Decoding Legalese Without Borders:

Multilingual Evaluation of Language Models on Long Legal Texts

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Joel Niklaus

from Müntschemier, BE

Supervisor of the doctoral thesis:
PD Dr. Matthias Stürmer

University of Bern

Decoding Legalese Without Borders:

Multilingual Evaluation of Language Models on Long Legal Texts

Inaugural dissertation
of the Faculty of Science,
University of Bern

presented by

Joel Niklaus

From Müntschemier, BE

Supervisor of the doctoral thesis :
PD Dr. Matthias Stürmer
University of Bern

Accepted by the Faculty of Science.

Bern, 24th of January 2024

The Dean
Prof. Dr. Marco Herwegh

ABSTRACT

Pretrained transformers have sparked an explosion of research in the field of Natural Language Processing (NLP). Scaling up language models based on the transformer architecture in terms of size, compute, and data led to impressive emergent capabilities that were considered unattainable in such a brief span, a mere three years ago, prior to the launch of GPT-3. These advances catapulted the previously niche field of legal NLP into the mainstream, at the latest, with GPT-4 passing the bar. Many products based on GPT-4 and other large language models are entering the market at an increasing pace, many of those targeting the legal field. This dissertation makes contributions in two key areas within Natural Language Processing (NLP) focused on legal text: resource curation and detailed model analysis. First, we curate an extensive set of multilingual legal datasets, train a variety of language models on these, and establish comprehensive benchmarks for evaluating Large Language Models (LLMs) in the legal domain. Second, we conduct a multidimensional analysis of model performance, focusing on metrics like explainability and calibration in the context of Legal Judgment Prediction. We introduce novel evaluation frameworks and find that while our trained models exhibit high performance and better calibration than human experts, they do not necessarily offer improved explainability. Furthermore, we investigate the feasibility of re-identification in anonymized legal texts, concluding that large-scale re-identification using LLMs is currently unfeasible. For future work, we propose exploring domain adaptation and instruction tuning to enhance language model performance on legal benchmarks, while also advocating for a detailed examination of dataset overlaps and model interpretability. Additionally, we emphasize the need for dataset extension to unexplored legal tasks and underrepresented jurisdictions, aiming for a more comprehensive coverage of the global legal landscape in NLP resources.

No matter what happens in life, be good to people.
Being good to people is a wonderful legacy to leave behind.

Taylor Swift

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to a number of individuals who have made significant contributions to my PhD journey.

I would like to express heartfelt thanks to my partner, Mara Häusler. Her unwavering love, support, and understanding have been the pillars of strength for me throughout this challenging journey. Her presence and encouragement have been indispensable to my successes and for this, I am deeply grateful.

I am immensely grateful to my friends and family, whose constant support and belief in me have been invaluable throughout my PhD journey. Their encouragement, understanding, and reassurance during challenging times provided the strength I needed to persevere. I deeply appreciate their unyielding faith in my abilities and their immense contributions to my personal and professional growth.

I am also deeply grateful to Michele Alberti for motivating me to pursue a PhD in the first place and for his invaluable advice at crucial stages of the journey.

I express deep appreciation for my supervisor, Matthias Stürmer, whose invaluable guidance and mentorship have been instrumental throughout my PhD journey. His belief in my abilities allowed me the freedom to explore and cultivate my academic potential, fostering significant professional development. His commitment to my growth extended to providing me with opportunities for teaching and public speaking, experiences that have been tremendously beneficial. His role in my journey has been transformative and I am profoundly grateful for his unwavering support.

I want to specially acknowledge Ilias Chalkidis. It was under his tutelage that I was first introduced to the intricate and compelling field of legal Natural Language Processing. His persistent mentorship, practical advice, and steadfast support have not only significantly shaped this dissertation, but also fundamentally enhanced my understanding of conducting research in legal NLP. I am immensely grateful for his guidance throughout my PhD journey.

I am thankful to Paolo Favaro and Daniel E. Ho for their insightful advice, which has been instrumental in shaping my research approach and thought processes.

I would like to express sincere thanks to Nikolaos Aletras for his insightful comments on the thesis and for serving as the external examiner at my PhD defense.

My heartfelt thanks go to the Swiss National Science Foundation (SNF) for funding this project. Their financial support has been crucial in allowing me to conduct this research and pursue my passion for NLP. Thanks to Andreas Lienhard, Daniel Kettiger, Thomas Myrach, Magda Chodup, Natalie Schwager, Tania Munz, and Leonie Grob for the collaboration.

I would also like to express my gratitude to my collaborators Veton Matoshi, Andrea Galassi, Pooja Rani, Gil Semo, Dor Bernsohn, Yaron Vazana, Ben Hagag, Gila Hayat, Daniele Giofré, Anastassia Shaitarova, Santosh T.Y.S.S, Matthias Grabmair, Robin Mamié, Daniel Brunner, Marcel Gygli, Magda Chodup, Thomas Lüthi, Daniel Kettiger, Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. The papers we worked on together were greatly enriched by their exceptional dedication and skill. Without them, this PhD would never have been possible. Their collaboration not only produced high-quality research, but also enhanced my skills and shaped my professional development in countless ways.

I extend my sincere gratitude to my students Vishvaksenan Rasiah, Ronja Stern, Tobias Brugger, Alex Nyffenegger, Ramona Christen, Nina Baumgartner, Marco Buchholz, and Alec Schürmann for their unwavering commitment and invaluable contributions to this research. Their diligent work, intellectual rigor, and collaborative spirit substantially elevated the quality of this project. Whether it was through meticulous data analysis, insightful discussions, or the persistent debugging of code, each student played a critical role in the advancement of this Ph.D. thesis in the field of Legal Natural Language Processing. Their collective effort not only enhanced the depth of this research but also greatly enriched my academic journey.

Finally, I would like to thank people from the University of Zurich Computational Linguistics Reading Group, Thomson Reuters Labs, the Stanford RegLab, the Stanford NLP Group, and (Google) X for insightful discussions around (legal) Natural Language Processing.

The fusion of machine learning and legal analysis embarks us on an unprecedented journey towards automating judicial reasoning, but it also summons a pressing inquiry into the transparency and ethics of algorithmic jurisprudence.

GPT-4

PUBLICATIONS

This dissertation summarizes the following publications. See the respective papers in Appendix A or online on https://niklaus.ai/files/theses/PhD_Thesis.pdf for the full content.

Bernsohn, Dor, Ben Hagag, Yaron Vazana, and Joel Niklaus (2023). *LegalLens: Leveraging Language Models for Legal Violation Identification in Unstructured Text*.

Brugger, Tobias, Matthias Stürmer, and Joel Niklaus (2023). "MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset." In: *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*. ICAIL '23. Braga, Portugal: Association for Computing Machinery, 42–51. ISBN: 9798400701979. DOI: 10.1145/3594536.3595132. URL: https://doi.org/10.1145/3594536.3595132.

Christen, Ramona, Anastassia Shaitarova, Matthias Stürmer, and Joel Niklaus (2023). *Resolving Legalese: A Multilingual Exploration of Negation Scope Resolution in Legal Documents*. arXiv: 2309.08695 [cs.CL].

Guha, Neel et al. (2023). *LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models*. arXiv: 2308.11462 [cs.CL].

Niklaus, Joel, Ilias Chalkidis, and Matthias Stürmer (Nov. 2021). "Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark." In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 19–35. URL: https://aclanthology.org/2021.nllp-1.3 (visited on 12/13/2021).

Niklaus, Joel, Magda Chodup, Thomas Lüthi, and Daniel Kettiger (Oct. 2023a). *Re-Identifizierung in Gerichtsurteilen mit Simap Daten*.

Niklaus, Joel and Daniele Giofre (July 2023). "Can we Pretrain a SotA Legal Language Model on a Budget From Scratch?" In: *Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*. Toronto, Canada (Hybrid): Association for Computational Linguistics, pp. 158–182. DOI: 10.18653/v1/2023.sustainlp-1.11. URL: https://aclanthology.org/2023.sustainlp-1.11.

Niklaus, Joel, Robin Mamié, Matthias Stürmer, Daniel Brunner, and Marcel Gygli (2023b). *Automatic Anonymization of Swiss Federal Supreme Court Rulings*. arXiv: 2310.04632 [cs.CL].

Niklaus, Joel, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis (2023c). *LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain*. arXiv: 2301.13126 [cs.CL].

Niklaus, Joel, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho (2023d). *MultiLegalPile: A 689GB Multilingual Legal Corpus*. arXiv: 2306.02069 [cs.CL].

Niklaus, Joel, Matthias Stürmer, and Ilias Chalkidis (Nov. 2022). "An Empirical Study on Cross-X Transfer for Legal Judgment Prediction." In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online only: Association for Computational Linguistics, pp. 32–46. URL: https://aclanthology.org/2022.aacl-main.3 (visited on 01/27/2023).

Nyffenegger, Alex, Matthias Stürmer, and Joel Niklaus (2023). *Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models*. arXiv: 2308.11103 [cs.CL].

Rasiah, Vishvaksenan, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E. Ho, and Joel Niklaus (2023). *SCALE: Scaling up the Complexity for Advanced Language Model Evaluation*. arXiv: 2306.09237 [cs.CL].

Semo, Gil, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus (Dec. 2022). "ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US." In: *Proceedings of the Natural Legal Language Processing Workshop 2022*. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, pp. 31–46. URL: https://aclanthology.org/2022.nllp-1.3 (visited on 04/17/2023).

T.Y.S.S, Santosh, Nina Baumgartner, Matthias Stürmer, Matthias Grabmair, and Joel Niklaus (2023). *Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset*.

CONTENTS

ACRONYMS

Life is really simple, but we insist on making it complicated —— Confucius

SFSC	Swiss Federal Supreme Court
ILDC	Indian Legal Documents Corpus
US	United States
EU	European Union
CoE	Council of Europe
AI	Artificial Intelligence
NLP	Natural Language Processing
ML	Machine Learning
SotA	state-of-the-art
NLU	Natural Language Understanding
TC	Text Classification
NER	Named Entity Recognition
QA	Question Answering
NLI	Natural Language Inference
SBD	Sentence Boundary Detection
NMT	Neural Machine Translation
LJP	Legal Judgment Prediction
CVG	Court View Generation
LDS	Leading Decision Summarization
GPT	Generative Pre-Trained Transformer
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bidirectional Long Short-Term Memory
GAN	Generative Adversarial Network
LM	Language Model
PLM	Pre-trained Language Model
LLM	Large Language Model
RTD	Replaced Token Detection
CLT	Cross-Lingual Transfer
CRF	Conditional Random Field

Part I

DISSERTATION

1 INTRODUCTION

Transformers (Vaswani et al., 2017) have revolutionized the field of Natural Language Processing (NLP) and Artificial Intelligence (AI) as a whole. With their advent, an unprecedented scale has been realized, both in terms of model size and data size (Brown et al., 2020). They have saturated many benchmarks and have achieved performances that supersede human capabilities (OpenAI, 2023). An exemplary case is Generative Pre-Trained Transformer (GPT)-4, a transformer-based model that even claims to pass the bar exam (a standardized test that evaluates a candidate's legal knowledge and skills, required for practicing law within a specific jurisdiction) (Katz et al., 2023).

The real-world impact of these developments is perhaps most evident in the burgeoning legal technology industry. Startups focusing on legal applications of AI have been thriving, with substantial funds being raised (e.g., Robin AI raising a 10.5M Series A[1] or Sequoia leading a 21M Series A for Harvey[2]). This trend is further cemented by recent acquisitions such as the case of Casetext being acquired by Thomson Reuters[3]. These advancements, however, do not mean we have fully realized the potential of AI in the legal field.

Despite progress, several challenges continue to persist, particularly in the legal field. A notable shortcoming is the lack of comprehensive datasets, especially when it comes to multilingual data. Furthermore, models tend to perform suboptimally when dealing with lower resourced languages, underlining the need for novel multilingual datasets and models. Specialized domains, such as the legal field, also pose unique challenges, often necessitating the development of models capable of handling more complex tasks. An additional challenge is the processing of long sequences, which remains problematic due to the quadratic nature of self-attention in transformer models (Tay et al., 2020a).

This thesis delves into these challenges and provides contributions towards improving the evaluation of models along three axes: long documents, multilinguality, domain specificity, and multi-tasking. Additionally, the publications in this thesis conduct rigorous analyses regarding Language Model (LM) performance, explainability, and calibration in Legal Judgment Prediction and re-identification risks of various Large Language Models (LLMs).

The explanation of the fundamental concepts of AI, Machine Learning (ML), and NLP is beyond the scope of this thesis. I refer the interested reader to excellent resources such as Goodfellow, Bengio, and

1 https://www.robinai.com/post/robin-raises-series-a
2 https://siliconangle.com/2023/04/27/legal-ai-focused-firm-harvey-raises-21m-led-sequoia/
3 https://www.reuters.com/markets/deals/thomson-reuters-acquire-legal-tech-provider-casetext-650-mln-2023-06-27/

Courville (2016) and LeCun, Bengio, and Hinton (2015), Speech and Language Processing (Jurafsky and Martin, 2009) and Natural Language Processing with Transformers for more in-depth information about the background.

2 GENERAL RELATED WORK

In this chapter, I review related work relevant to all the five papers. Specific related work is mentioned in the full texts of the papers.

2.1 LONG DOCUMENT PROCESSING

In recent years, substantial research efforts have been dedicated to tackling the issue of quadratic time and memory complexity inherent in the dense attention mechanism (Vaswani et al., 2017), which practically constrains the maximum sequence length to a severe extent (typically to 512 tokens) (Beltagy, Peters, and Cohan, 2020; Child et al., 2019; Kitaev, Kaiser, and Levskaya, 2020; Lee-Thorp et al., 2021; Roy et al., 2021; Tay et al., 2021, 2020b; Zaheer et al., 2020). This has led to the emergence of a new category of transformers, known as sparse or efficient transformers (Tay et al., 2020b). The fundamental concept behind efficient transformers is to decrease the computational cost associated with the dense attention matrix while preserving performance. This is typically achieved by introducing sparsity into the attention matrix in various ways, such as fixed patterns like local (windowed) attention (Beltagy, Peters, and Cohan, 2020; Child et al., 2019), global attention (Zaheer et al., 2020), learnable patterns like routing attention (Roy et al., 2021) and attention using locality-sensitive hashing (Kitaev, Kaiser, and Levskaya, 2020), or random patterns (Tay et al., 2021; Zaheer et al., 2020). Fourier transforms were proposed as an alternative to the attention layer by Lee-Thorp et al. (2021). A comprehensive overview of efficient transformers and their attention mechanisms is provided by Tay et al. (2020b).

While these methods listed above are very general and allow for a wide range of tasks, they often require dedicated pretraining, significantly adding to the costs. Though limited to classification tasks, but significantly reducing costs, Pappagari et al. (2019) make use of standard-length encoder models in a hierarchical setup.

Tay et al. (2020a) introduced a set of tasks, known as the "Long Range Arena", specifically designed to evaluate the capabilities of these models when dealing with longer inputs. Note that these tasks are largely artificial, aiming to assess the models independently of any pretraining.

In the publications of this thesis we evaluate a range of long document processing methods on legal datasets, finding good results for the Longformer architecture and hierarchical transformers.

Conneau et al. (2020) developed a multilingual language model trained on 2.5 TB of data crawled from the web in 100 languages. This model surpassed the performance of earlier multilingual models (Devlin et al., 2019; Lample and Conneau, 2019) in tasks such as classification, sequence labelling, and Question Answering (QA). Pfeiffer et al. (2020) introduced a framework for transferring knowledge across tasks and languages. By leveraging multilingually pretrained models like XLM-R (Conneau et al., 2020), they achieved impressive results even on languages not included in the pretraining corpus. Cross-Lingual Transfer (CLT) has become a vibrant research area, with the application of multilingual pretrained transformer-based models (Conneau et al., 2020; Devlin et al., 2019; Lample and Conneau, 2019; Xue et al., 2021a) showing excellent performance in Natural Language Understanding (NLU) benchmarks (Ruder et al., 2021). Adapter-based fine-tuning (Bhatia et al., 2023; Houlsby et al., 2019; Pfeiffer et al., 2021) has been proposed as a strategy to alleviate the misalignment of multilingual knowledge when CLT is applied, particularly in a zero-shot scenario where the target language is not seen during training.

However, the application of CLT in legal NLP remains relatively unexplored. Chalkidis, Fergadiotis, and Androutsopoulos (2021) experimented with standard fine-tuning and also explored the use of adapters (Houlsby et al., 2019) for zero-shot CLT on a legal topic classification dataset composed of European Union (EU) laws. They found that adapters provided the best balance between effectiveness and efficiency. Their work did not investigate the use of methods that incorporate translated versions of the original documents. Recently, Xenouleas et al. (2022) used an updated, unparalleled version of the dataset from Chalkidis, Fergadiotis, and Androutsopoulos to study Neural Machine Translation (NMT)-augmented CLT methods. Other multilingual legal NLP resources (Drawzeski et al., 2021; Galassi et al., 2020) have been recently made available, but none of them apply CLT in any form.

In this thesis, we extend the work on CLT for the multilingual legal domain and find that cross-jurisdiction transfer from translated Indian Legal Documents Corpus (ILDC) cases improve performance on Swiss decisions.

2.3 DOMAIN SPECIFICITY

While general-purpose LMs are typically trained on generic text corpora like Wikipedia and evaluated on widely used benchmarks such as GLUE (Wang et al., 2018), domain-specific models require specialized datasets for training and dedicated benchmarks for quality assessment. The following examples demonstrate the performance improvements achieved when using domain-specific datasets and benchmarks.

In the field of biomedical NLP (BioNLP), Lee et al. (2019) pioneered the development of a domain-specific LM based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), pretraining it on biomedical text corpora. The resulting BioBERT model outperformed BERT in biomedical NLP tasks, and these scores were later surpassed by Naseem et al. (2022) who conducted domain-specific pretraining of ALBERT (Lan et al., 2020) using biomedical text and the MIMIC-III (**M**edical **I**nformation **M**art for **I**ntensive **C**are) dataset (Johnson et al., 2016).

In the financial domain, Yang, Uy, and Huang (2020) pretrained FinBERT on financial data, outperforming generic BERT models in all financial datasets. This was later improved upon by Shah et al. (2022) with the introduction of FLANG-BERT and the FLUE (**F**inancial **L**anguage **U**nderstanding **E**valuation) benchmark. In May 2023, Bloomberg announced the BloombergGPT model for the financial domain (Wu et al., 2023), although no datasets, benchmarks, or weights have been publicly released.

Several other domain-specific LMs have been developed, such as SciBERT for scientific publications (Beltagy, Lo, and Cohan, 2019), ConfliBERT for monitoring political violence and conflicts (Hu et al., 2022), PoliBERTweet for analyzing political content on Twitter (Kawintiranon and Singh, 2022), SecureBERT for cybersecurity (Aghaei et al., 2023), and BlueBERT for the biomedical domain (Peng, Yan, and Lu, 2019), among others.

In the legal domain, several models have been pretrained on various legal corpora, such as LegalBERT (Chalkidis et al., 2020), CaseHoldBERT (Zheng et al., 2021), PoL-BERT (Henderson et al., 2022), and LegalReformer (Hua et al., 2022). Recently, Chalkidis et al. (2023) released LexFiles, a large English legal corpus, and trained two new legal English Pre-trained Language Models (PLMs), demonstrating improved performance in legal probing and classification tasks.

While there have been efforts to pretrain legal LMs in languages other than English, such as Italian (Licari and Comandè, 2022), Romanian (Masala et al., 2021), and Spanish (Gutiérrez-Fandiño et al., 2021), English remains the dominant language, highlighting the need for the compilation of multilingual legal corpora.

As part of this thesis, we pretrain a host of multilingual and monolingual models in 24 languages and make them publicly available to the community. Our models are the new SotA on LexGLUE and LEXTREME.

2.4 BENCHMARKS

Standardized benchmarks are important for measuring progress and for spurring new innovation. I briefly review the literature on benchmarks along four aspects: a) long documents, b) multilinguality, c) domain specificity and d) multitasking.

2.4.1 *Long Documents*

Most current LMs are capable of handling inputs in the range of a few thousand tokens. However, very long documents in the tens of thousands of tokens are still very hard for current models. Additionally, models seem to struggle with making use of all information, especially when located in the middle of long texts (Liu et al., 2023).

SCROLLS (**S**tandardized **C**ompa**R**ison **O**ver **L**ong **L**anguage **S**equences) (Shaham et al., 2022) is an ensemble of tasks encompassing summarization, QA, and Natural Language Inference (NLI). The distinct feature of this dataset is that the typical input examples comprise thousands of words in English. MuLD (**Mu**ltitask **L**ong **D**ocument) (Hudson and Moubayed, 2022) is a collection of six tasks: two instances of QA, style change detection, Text Classification (TC), summarization, and translation. Each input in the MuLD dataset has a minimum of 10K tokens, while some extend to nearly half a million tokens.

So far, there are only a limited number of benchmarks available for long documents and most are limited to English. We extend the literature with multiple challenging datasets for evaluating LMs on long documents such as Court View Generation (CVG) and Leading Decision Summarization (LDS).

2.4.2 *Multilinguality*

Currently, the majority of NLP research is still conducted in English, even though most people on earth speak another native language. Therefore, benchmarks measuring LM performance multilingually are very important for facilitating multilingual research.

The **C**ross-lingual **TR**ansfer **E**valuation of **M**ultilingual **E**ncoders (XTREME) benchmark (Hu et al., 2020) is formulated to scrutinize cross-lingual generalization capabilities. It incorporates six tasks dispersed across ten datasets, accommodating 40 languages. These datasets comprise both cross-lingual and professionally as well as automatically translated texts. Building upon XTREME, XTREME-UP (Ruder et al., 2023) places an emphasis on assessing multilingual models in a few-shot setting, catering to user-centric tasks. Notably, it encompasses 88 under-represented languages, including but not limited to Swahili, Burmese, and Telugu, which currently have scarce datasets.

The large majority of NLP benchmarks tackle the English, leaving many other languages under-represented. We release a host of datasets geared for multilingual evaluation in European languages.

2.4.3 *Domain Specificity*

While LM performance is already very high on generic texts like news, specialized domains still pose a host of harder problems, to a large part not solved to date.

The BLUE (**B**iomedical **L**anguage **U**nderstanding **E**valuation) benchmark (Peng, Yan, and Lu, 2019) houses five tasks spread over ten datasets, specifically geared towards biomedical and clinical texts. Similarly, CBLUE (Zhang et al., 2022), a benchmark for Chinese biomedical texts, offers eight NLU tasks such as Named Entity Recognition (NER), TC, QA, information extraction, diagnosis normalization, intent classification, and semantic similarity. LexGLUE (Chalkidis et al., 2022) encompasses six predictive tasks over five datasets comprising English documents sourced from the United States (US), the EU, and the Council of Europe (CoE). Finally, LBOX-OPEN (Hwang et al., 2022) consists of five legal tasks derived from South Korean legal documents.

While there exist domain specific benchmarks, the availability in the legal domain is low, especially multilingually. We fill this gap by releasing two challenging multilingual legal benchmarks focused on European languages.

2.4.4 *Multitasking*

A single model, capable of solving multiple tasks is much more useful than a model for just one single task. For that reason, many benchmarks evolved for measuring LM performance across a wide range of tasks.

GLUE (**G**eneral **L**anguage **U**nderstanding **E**valuation) (Wang et al., 2018), an early benchmark for sentence NLU tasks intended to evaluate general-purpose neural LMs, quickly became outmoded due to advanced models such as BERT (Devlin et al., 2019). Its successor, SuperGLUE (Wang et al., 2019), incorporated new tasks that pose a challenge for machines but are solvable by humans. MMLU (**M**assive **M**ultitask **L**anguage **U**nderstanding) emphasizes zero-shot and few-shot learning tasks, and includes approximately 16K multiple-choice questions segregated into 57 subtasks, spanning diverse subjects from humanities to hard sciences. The first Chinese language multitask benchmark, CLUE (Xu et al., 2020), comprises single sentence classification, sentence pair classification, and machine reading comprehension tasks. BIG-Bench (**B**eyond the **I**mitation **G**ame) (Srivastava et al., 2022) includes 204 language tasks developed by 450 authors from 132 institutions, covering a wide array of topics, such as linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, and software development. Lastly, HELM (**H**olistic **E**valuation of **L**anguage **M**odels) (Liang et al., 2022) is a comprehensive multi-metric benchmark covering seven metrics and seven targeted evaluations, involving 42 test scenarios with a large-scale evaluation of 30 LMs.

While there are many benchmarks available, tackling multiple tasks concurrently, few are available in the legal domain. We fill this gap by releasing datasets in a total of ten large task categories with multiple sub-categories within.

3 CONTRIBUTIONS

In this chapter, I briefly summarize the contributions of each paper and refer the interested reader to the full text at the end of the thesis in Appendix A. Table 1 shows an overview of the artifacts and resources released as part of the publications.

Table 1: Resources of the Publications

Title	Paper	Code GitHub	Datasets Hugging Face	Models Hugging Face
Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark	NLLP @ EMNLP 2021	Swiss-Judgment-Prediction	Swiss-Judgment-Prediction	-
An Empirical Study on **Cross-X Transfer** for Legal Judgment Prediction	AACL-IJCNLP 2022	Swiss-Judgment-Prediction	-	-
ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US	NLLP @ EMNLP 2022	Class-Action-Prediction	Class-Action-Prediction	-
MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset	ICAIL 2023	Multi-Legal-SBD	Multi-Legal-SBD	Multi-Legal-SBD
BudgetLongformer: Can we Pretrain a SotA Legal Language Model on a Budget From Scratch?	SustaiNLP @ ACL 2023	-	-	Budget-Longformer
LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models	NeurIPS Datasets and Benchmarks 2023	Legal-Bench	Legal-Bench	-
LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain	EMNLP Findings 2023	LEXTREME	LEXTREME	-
Automatic Anonymization of Swiss Federal Supreme Court Rulings	NLLP @ EMNLP 2023	-	-	Swiss Legal-LMs
MultiLegalPile: A 689GB Multilingual Legal Corpus	ArXiv Pre-Print, DMLR @ ICML 2023, NLLP @ EMNLP 2023, submitted to AAAI 2024	Legal-Datasets	Multi-Legal-Pile	Multilingual Legal-LMs
SCALE: Scaling up the Complexity for Advanced Language Model Evaluation	ArXiv Pre-Print, NLLP @ EMNLP 2023, submitted to ICLR 2024	SCALE	SCALE	Swiss Legal-LMs
Resolving Legalese: A Multilingual Exploration of Negation Scope Resolution in Legal Documents	ArXiv Pre-Print, NLLP @ EMNLP 2023, submitted to LREC-COLING 2024	Resolving-Legalese	Multi-Legal-Neg	Neg-XLM-RoBERTa
Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models	ArXiv Pre-Print, submitted to AAAI 2024	Anonymity-at-Risk	Anonymity-at-Risk	-
LegalLens: Leveraging Language Models for Legal Violation Identification in Unstructured Text	Submitted to EACL 2024	-	-	-
Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset	NLLP @ EMNLP 2023, submitted to LREC-COLING 2024	-	-	-
Re-Identification of Corporations in Swiss Court Decisions with Simap Data (in German)	Zenodo Pre-Print	Swiss-Court-Decision-ReIdentification	-	-

3.1 SWISS-JUDGMENT-PREDICTION: A MULTILINGUAL LEGAL JUDGMENT PREDICTION BENCHMARK

PROBLEM Court overloads often lead to significant delays in many jurisdictions, and predictive AI models can aid legal professionals in their work, thereby enhancing efficiency. Existing Legal Judgment Prediction datasets are mostly restricted to English, French, and Chinese.

OUR CONTRIBUTION We present a newly released multilingual corpus, composed of 85K cases from the Federal Supreme Court of Switzerland, spanning 2000-2020, in German, French, and Italian. We evaluate the corpus using BERT-based methods, including two variants capable of overcoming the 512 token limit of BERT, with Hierarchical BERT delivering the highest performance. We also examine how factors like the canton of origin, publication year, text length, and legal area influence performance, and we have publicly released both the dataset and the code for future research and reproducibility.

3.2 AN EMPIRICAL STUDY ON CROSS-X TRANSFER FOR LEGAL JUDGMENT PREDICTION

PROBLEM Cross-lingual transfer learning, while useful in many Natural Language Processing (NLP) tasks, remains understudied in the legal NLP realm, especially in Legal Judgment Prediction (LJP).

OUR CONTRIBUTION We investigate transfer learning techniques in LJP, utilizing the trilingual Swiss-Judgment-Prediction dataset. We observe that cross-lingual transfer enhances overall results across languages, particularly when adapter-based fine-tuning (updating only a small part of the weights) is employed. Performance is further boosted by augmenting the training dataset with machine-translated versions of the original documents, thus expanding the training corpus threefold. The research also highlights improved results when implementing cross-jurisdiction transfer, such as training the model across different legal areas and regions, and augmenting the dataset with Indian legal cases.

3.3 CLASSACTIONPREDICTION: A CHALLENGING BENCHMARK FOR LEGAL JUDGMENT PREDICTION OF CLASS ACTION CASES IN THE US

PROBLEM Legal NLP is an increasingly active field, with LJP becoming a prevalent research focus. However, most publicly available LJP datasets are derived from countries practicing civil law. Additionally, most LJP datasets study less realistic scenarios basing their input on the court decisions which are often written to support the final argument made.

our contribution We introduce a novel and challenging LJP dataset centered on US class action cases, the first of its kind in the common law system, focusing on complaint inputs rather than court-written fact summaries. An accompanying study reveals the task's complexity, with human experts achieving only 53% accuracy, while our Longformer model surpasses this human baseline with a 63% accuracy rate, despite only considering the first 2,048 tokens. Additionally, an error analysis demonstrates the model's superior calibration compared to human experts, and we publicly release the dataset and code for further exploration.

3.4 Can we pretrain a SOTA legal language model on a budget from scratch?

problem Pretrained transformer models, while achieving state-of-the-art results in many tasks, often struggle to process texts longer than 512 tokens, a limitation particularly restrictive in specialized fields like legal or scientific domains where texts can exceed 10,000 tokens. Efficient transformers like Longformer, BigBird, or FNet (all methods enabling efficient processing of longer sequences) have been proposed, but few are available for specialized domains, and their pretraining process is generally costly.

our contribution We use the Replaced Token Detection (RTD) task (using a training setup similar to Generative Adversarial Networks (GANs) (Goodfellow et al., 2014)) to make the pretraining of Longformer models on legal data more cost-effective. The resulting models are evaluated on challenging summarization tasks, demonstrating that they outperform their baselines on both the in-domain BillSum (a legal summarization task of US bills) and out-of-domain PubMed (a medical summarization task of biomedical scientific articles) tasks within their respective parameter range. The code and models used in the study have been made publicly available for further research.

3.5 MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset

problem Sentence Boundary Detection (SBD) a crucial component of NLP, is particularly challenging in the legal domain due to the complexity and variety of sentence structures.

our contribution To address this, we have curated a diverse multilingual legal dataset, comprising over 130K annotated sentences in six languages. Experimental results reveal that the performance of existing SBD models is unsatisfactory for multilingual legal data. We trained and evaluated monolingual and multilingual models based on Conditional Random Fields (CRFs), Bidirectional Long Short-Term

Memory (BiLSTM)-CRFs, and transformers, and achieved state-of-the-art performance, with our multilingual models surpassing all baselines in zero-shot settings on a Portuguese test set. We have made our dataset, models, and code publicly available to foster further research and development.

3.6 LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models

Problem As LLMs become increasingly integrated into the legal domain, questions arise about their capabilities in legal reasoning. The critical question is: What types of legal reasoning can LLMs effectively perform? Existing benchmarks fail to provide a comprehensive evaluation framework to examine these capabilities from the perspective of legal professionals. There is also a lack of interdisciplinary tools that bridge the gap between the computational and legal communities, making it challenging to establish a common vocabulary for describing and evaluating legal reasoning in LLMs.

Our contribution We introduce LegalBench, a benchmark comprising 162 tasks that cover six distinct types of legal reasoning. We assembled LegalBench through an interdisciplinary collaboration, predominantly led by legal professionals, ensuring the tasks are either practically useful or intellectually engaging for legal reasoning. We empirically evaluate 20 open-source and commercial LLMs on this benchmark, providing a comprehensive understanding of their legal reasoning capabilities. Additionally, we map popular legal frameworks for describing types of legal reasoning onto tasks in LegalBench, facilitating a shared vocabulary for both legal professionals and LLM developers. This work thus serves as a cornerstone for cross-disciplinary discussions and provides valuable insights into the types of research explorations that LegalBench enables.

3.7 LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain

Problem To measure progress in the fast-growing NLP field, well-curated and challenging benchmarks are crucial. Previous efforts have produced numerous benchmarks for general NLP models, typically based on news or Wikipedia. However, these may not fit specific domains such as law, with its unique lexicons and intricate sentence structures. Even though there is a rising need to build NLP systems for languages other than English, many benchmarks are available only in English and no multilingual benchmark exists in the legal NLP field.

Our contribution We survey the legal NLP literature and select 11 datasets covering 24 languages, creating LEXTREME. To fairly compare models, we propose two aggregate scores, i.e., dataset aggregate score and language aggregate score. Our results indicate that even the best baseline only achieves modest results, and also ChatGPT struggles with many tasks. This indicates that LEXTREME remains a challenging benchmark with ample room for improvement. To facilitate easy use for researchers and practitioners, we release LEXTREME on huggingface along with a public leaderboard and the necessary code to evaluate models.

3.8 Automatic Anonymization of Swiss Federal Supreme Court Rulings

Problem The Swiss Federal Supreme Court (SFSC) employs a system for anonymizing court decisions that merges traditional computational methods with human expertise. While the existing system performs adequately, there remains a need to improve its efficiency and accuracy to reduce manual labor and ensure the protection of all parties involved.

Our contribution We enrich the current anonymization software by introducing a large, annotated dataset comprising entities requiring anonymization. We then compare the performance of BERT-based models with models pre-trained on in-domain data. Our findings indicate that the use of in-domain data for pre-training enhances the F1-score by over 5% relative to the existing models. Further, we demonstrate that integrating ML techniques with existing anonymization methods, such as regular expressions, not only elevates the performance of automatic suggestions but also reduces the necessity for manual intervention.

3.9 MultiLegalPile: A 689GB Multilingual Legal Corpus

Problem The availability of large, high-quality datasets is instrumental for the training of LLMs, especially in specialized domains such as law. However, existing datasets in this domain are often limited to the English language and do not represent multiple jurisdictions. This poses a challenge for the development of universally applicable NLP solutions for legal tasks.

Our contribution To address this gap, we curate and release MultiLegalPile, a comprehensive 689GB corpus that covers 24 languages and 17 jurisdictions. The dataset encompasses diverse legal data sources and features varying licenses, allowing for the pretraining of NLP models under fair use conditions. Special subsets of the data, such as Eurlex Resources and Legal mC4, come with more permissive licenses. We pretrain two RoBERTa models and one Longformer model multilingually, along with 24 monolingual models tailored to each language-specific subset. We evaluate these models on the LEXTREME and LexGLUE benchmarks. Our multilingual models establish a new SotA on LEXTREME, and our English models surpass existing models on LexGLUE. We release the MultiLegalPile dataset, the pretrained models, and the corresponding code under the most open licenses available.

3.10 SCALE: Scaling up the Complexity for Advanced Language Model Evaluation

PROBLEM Despite the advancements in LLMs, several open challenges remain, particularly in the processing of long documents, leveraging domain-specific knowledge, multilingual understanding, and multitasking capabilities. Current NLP benchmarks are insufficiently rigorous to comprehensively evaluate LLMs along these dimensions. Most benchmarks are focused predominantly on English-language tasks and often do not include specialized, domain-specific, or multilingual aspects. This lack of comprehensive and domain-specific benchmarks is especially apparent in the realm of legal NLP.

OUR CONTRIBUTION We introduce a novel NLP benchmark designed to rigorously assess LLM capabilities across four key dimensions: long document processing (up to 50K tokens), domain-specific knowledge utilization (in legal texts), multilingual understanding (in five languages), and multitasking (including a wide range of legal NLP tasks). Our benchmark encompasses diverse legal NLP datasets from the inherently multilingual Swiss legal system. We establish strong baselines by evaluating several pre-trained multilingual LLMs on this benchmark. Notably, even the existing SotA models struggle with most tasks on our benchmark, highlighting the benchmark's rigor and challenging nature. We make all resources, including the benchmark suite, pre-trained models, and code, publicly available under a fully permissive open CC BY-SA license.

3.11 Resolving Legalese: A Multilingual Exploration of Negation Scope Resolution in Legal Documents

PROBLEM Resolving the scope of negation in a sentence is a complex task in NLP. The intricacy of legal texts and the absence of annotated negation corpora specifically tailored for the legal domain further exacerbate the challenges faced by SotA models when applied to multilingual legal data. Current models, particularly those not fine-tuned on legal data, exhibit underperformance in negation scope resolution tasks.

OUR CONTRIBUTION We release a new set of annotated court decisions in German, French, and Italian to bolster research on negation scope resolution in legal texts. Utilizing this corpus, we conduct experiments in both zero-shot and multilingual settings. Models fine-tuned exclusively on domains such as literary texts and medical data were found to underperform compared to prior cross-domain studies. In our zero-shot cross-lingual experiments, we achieve token-level F1-scores of up to 86.7%. Moreover, in our multilingual experiments, which employ models trained on all available negation data and evaluated on our new legal corpus, we report F1-scores reaching up to 91.1%. These results demonstrate significant performance gains in negation scope resolution within the context of multilingual legal texts. We publicly release the code, data and trained models for the wider community.

3.12 Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models

PROBLEM Anonymity in court rulings is a critical aspect of privacy protection in the European Union and Switzerland. The rise of LLMs has heightened concerns about the potential for large-scale re-identification of anonymized individuals in legal documents.

OUR CONTRIBUTION We explore the capability of LLMs to re-identify individuals in Swiss court rulings, given the significant implications for privacy and legal transparency. We construct a proof-of-concept experiment using actual legal data from the Federal Supreme Court of Switzerland to investigate the feasibility of re-identification by LLMs. To augment this, we also build an anonymized Wikipedia dataset as a more rigorous testing environment. We introduce a new task focused on the re-identification of individuals in texts and propose new performance metrics for evaluating this task. Our systematic analysis uncovers the factors influencing successful re-identification, such as model size, input length, and instruction tuning. Despite high re-identification rates in the Wikipedia dataset, LLMs face challenges with court decisions due to limitations like data sparsity and lack of adequate test datasets. Ultimately, our study suggests that re-identification through LLMs may not be currently feasible, although future possibilities cannot be ruled out. The insights gained aim to bolster confidence in the security of anonymized legal decisions, potentially encouraging more widespread publication by courts.

3.13 LegalLens: Leveraging Language Models for Legal Violation Identification in Unstructured Text

PROBLEM The proliferation of the internet has led to an immense volume of unstructured textual data, including news articles, reviews, and social media posts. Within this data, legal violations frequently go unnoticed, obscured by the sheer amount of information. These hidden violations have far-reaching implications, affecting individual rights, societal norms, and the principles of justice. Current methods, often specialized for particular domains, lack the versatility needed to identify the wide variety of legal violations across different contexts.

OUR CONTRIBUTION We address both the identification of legal violations in unstructured text and the association of these violations with potentially affected individuals. These tasks are particularly tailored for the context of class-action legal cases. We construct two specialized datasets using LLMs and validate them with annotations from domain experts. In our experimental design, we fine-tune models from the BERT family and employ both open-source and closed-source LLMs. Our results yield an F1-score of 62.69% for the task of identifying legal violations and 81.02% for associating these violations with potential victims. We release both the datasets and the code for our experiments.

3.14 TOWARDS EXPLAINABILITY AND FAIRNESS IN SWISS JUDGEMENT PREDICTION: BENCHMARKING ON A MULTILINGUAL DATASET

PROBLEM Assessing the explainability of LJP systems is imperative for building trustworthy and transparent models. This becomes particularly critical when these models rely on factors that may either lack legal relevance or contain sensitive attributes.

OUR CONTRIBUTION We curate a comprehensive dataset containing rationales, both supporting and opposing legal judgments, from domain experts for 108 cases in three languages: German, French, and Italian. Utilizing an occlusion-based explainability methodology, we assess the performance of state-of-the-art monolingual and multilingual BERT-based LJP models. We also evaluate models that incorporate data augmentation and cross-lingual transfer techniques. Our results reveal that an increase in prediction performance does not necessarily yield better explainability. To further scrutinize this, we introduce a novel evaluation framework, Lower Court Insertion, which quantifies the influence of lower court decisions on model predictions, exposing inherent biases.

3.15 RE-IDENTIFICATION OF CORPORATIONS IN SWISS COURT DECISIONS WITH SIMAP DATA (IN GERMAN)

PROBLEM Easy public access to court decisions is important for maintaining a healthy democracy and for the daily work of legal professionals. However, Swiss courts still do not publish a large portion of the decisions made. This is mainly due to the high manual effort required for anonymization. Another reason could be the fear of re-identification of anonymized individuals in the judgments.

OUR CONTRIBUTION We describe a possible attack vector: Public tenders from the Simap platform. We search the judgments for report numbers and project IDs and compare these with the Simap data. In this way, we can identify companies in 271 out of 340 (80%) examined anonymized Swiss judgments. The judgments involved disputes valued at up to 250 million CHF.

4 CONCLUSIONS AND FUTURE WORK

In this chapter, I summarize the conclusions and future work of all the publications presented in this thesis.

4.1 CONCLUSIONS

This thesis makes two main types of contributions to the field of NLP: The large-scale curation of resources and detailed analyses of LMs operating on legal text.

4.1.1 *Resource Curation*

In this work, we contribute to the field of multilingual legal NLP by curating diverse datasets, pretraining and finetuning LMs, and creating comprehensive benchmarks integrated into an existing evaluation framework. First, we curated a wide array of datasets for legal NLP across a wide range of tasks: From a) unlabeled large document collections for self-supervised pretraining, to b) lower level structural tasks such as Sentence Boundary Detection, Negation Scope Resolution, and Citation Extraction, to c) more complicated classification tasks such as Law Area Prediction, Judgment Prediction, Criticality Prediction, and Court Decision Re-Identification to d) Information Retrieval and finally e) complicated free-form generation problems such as Court View Generation and Leading Decision Summarization. Second, we a) pretrained a host of multilingual and monolingual LMs in 24 languages and b) finetuned models for all the above tasks. We openly release trained models for future work to build on. Finally, we created benchmarks, carefully selecting and aggregating suitable high-quality datasets for a more holistic and comprehensive evaluation of LLMs in the legal domain. We openly release the benchmarking code for easy evaluation of future models and also integrate them into the Holistic Evaluation of Language Models (HELM) framework for more widespread adoption. I hope that the created resources can kick-start future research projects in multilingual legal NLP.

4.1.2 *Analysis*

We evaluate LM performance, explainability, and calibration in Legal Judgment Prediction and re-identification risks of various LLMs. First, we analysed model performance across multiple dimensions such as date, legal area, language, and jurisdiction. We find that cross-lingual, cross-domain and even cross-jurisdiction transfer improve results on our Swiss Judgment Prediction dataset. Second, we assess the explainability of SotA LJP models using occlusion and find that increased prediction performance does not necessarily lead to better explainability. We also introduced a novel evaluation framework, Lower Court Insertion, which quantifies the influence of lower court decisions on model predictions, exposing inherent biases. Finally, we measure model calibration on a US LJP dataset and find that our trained model is better calibrated than human experts. Finally, we investigate the risk of re-identification of anonymized persons in court decisions. We manage to re-identify companies in 271 of 340 (80%) investigated court decisions by matching report numbers and project IDs with Simap data. More general re-identification using LLMs does not seem to be feasible currently at a large scale.

4.2 FUTURE WORK

We would like to pursue future work in three main categories: Pushing the SotA on legal benchmarks, further analysis of the resources, and further extension of the datasets.

4.2.1 *Pushing the SotA on Legal Benchmarks*

We would like to improve the performance of LMs on the benchmarks proposed in this thesis. First, Domain adaptation has proved advantageous in smaller encoder based LMs. We could test whether large-scale continued pretraining on a large legal corpus like the MultiLegalPile can also improve performance of large generative models. Second, large-scale instruction finetuning led to astonishing multi-task capabilities for large decoder based LMs (Chung et al., 2022; Longpre et al., 2023). We could curate a large scale legal instruction tuning dataset to test whether finetuning on such a dataset could yield further improvements. Finally, more sophisticated prompting techniques (Khattab et al., 2023; Wang et al., 2023; Yao et al., 2023) have shown drastic performance increases in reasoning tasks. We could investigate whether techniques like these can be applied successfully to legal reasoning tasks.

4.2.2 *Further Analysis of the Resources*

We would like to perform more in-depth analyses of the proposed datasets and models. First, it is unclear how large the overlap of the MultiLegalPile is with other large-scale pretraining datasets based on web scrapes such as mC4 (Xue et al., 2021b). Future work may calculate metrics for overlap of these corpora, giving more information about the value of additional pretraining on the MultiLegalPile. Second, explainability and interpretability of models is still an area of active research. Future work may analyse in more detail rationales given by the models for their decisions. Finally, we could only perform preliminary error analyses in the publications that introduce the resources. More comprehensive studies could provide valuable insights into current failure modes, and thus outlining avenues for further research to improve model performance.

4.2.3 *Further Extension of the Datasets*

We would like to further extend the proposed datasets, providing more coverage for the legal domain and thus facilitating research. First, while we already cover a wide range of tasks, many tasks done by lawyers in their daily work are not yet covered by structured NLP datasets. We would like to further expand our datasets to more tasks relevant for the legal profession, both in the public and private sector. Second, while coverage of legal tasks is high in the US and China, most of the rest of the world is poorly represented by legal NLP datasets. We would like to expand access to more languages and jurisdictions, enabling legal language technology to spread more widely. Finally, many datasets are created semi-automatically, due to lacking resources. Data annotated by legal professionals specifically for the purpose of training and evaluating AI models will be crucial in improving the (semi-)automated legal assistants of the future.

Aghaei, Ehsan, Xi Niu, Waseem Shadid, and Ehab Al-Shaer (2023). "SecureBERT: A Domain-Specific Language Model for Cybersecurity." In: *Security and Privacy in Communication Networks*. Ed. by Fengjun Li, Kaitai Liang, Zhiqiang Lin, and Sokratis K. Katsikas. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Place: Cham. Springer Nature Switzerland, pp. 39–56. DOI: 10.1007/978-3-031-25538-0_3.

Beltagy, Iz, Kyle Lo, and Arman Cohan (Sept. 2019). "SciBERT: A Pretrained Language Model for Scientific Text." In: *arXiv:1903.10676 [cs]*. arXiv: 1903.10676. URL: http://arxiv.org/abs/1903.10676 (visited on 11/18/2021).

Beltagy, Iz, Matthew E. Peters, and Arman Cohan (Dec. 2020). "Longformer: The Long-Document Transformer." In: *arXiv:2004.05150 [cs]*. arXiv: 2004.05150. URL: http://arxiv.org/abs/2004.05150 (visited on 03/05/2021).

Bhatia, Kush, Avanika Narayan, Christopher De Sa, and Christopher Ré (June 2023). *TART: A plug-and-play Transformer module for task-agnostic reasoning*. en. arXiv:2306.07536 [cs]. URL: http://arxiv.org/abs/2306.07536 (visited on 06/15/2023).

Brown, Tom et al. (2020). "Language Models are Few-Shot Learners." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chalkidis, Ilias, Manos Fergadiotis, and Ion Androutsopoulos (Nov. 2021). "MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 6974–6996. DOI: 10.18653/v1/2021.emnlp-main.559. URL: https://aclanthology.org/2021.emnlp-main.559.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos (2020). "LEGAL-BERT: The Muppets straight out of Law School." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904.

Chalkidis, Ilias, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard (May 2023). *LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development*. en. arXiv:2305.07507 [cs]. URL: http://arxiv.org/abs/2305.07507 (visited on 05/16/2023).

Chalkidis, Ilias, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras (2022). “LexGLUE: A Benchmark Dataset for Legal Language Understanding in English.” In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4310–4330.

Child, Rewon, Scott Gray, Alec Radford, and Ilya Sutskever (Apr. 2019). “Generating Long Sequences with Sparse Transformers.” In: *arXiv:1904.10509 [cs, stat]*. arXiv: 1904.10509. URL: http://arxiv.org/abs/1904.10509 (visited on 10/04/2021).

Chung, Hyung Won et al. (Oct. 2022). *Scaling Instruction-Finetuned Language Models*. arXiv:2210.11416 [cs]. DOI: 10.48550/arXiv.2210.11416. URL: http://arxiv.org/abs/2210.11416 (visited on 10/24/2022).

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (Apr. 2020). “Unsupervised Cross-lingual Representation Learning at Scale.” In: *arXiv:1911.02116 [cs]*. arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116 (visited on 10/05/2021).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423.

Drawzeski, Kasper, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni (Nov. 2021). “A Corpus for Multilingual Analysis of Online Terms of Service.” In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1–8. DOI: 10.18653/v1/2021.nllp-1.1. URL: https://aclanthology.org/2021.nllp-1.1.

Galassi, Andrea, Kasper Drazewski, Marco Lippi, and Paolo Torroni (Dec. 2020). “Cross-lingual Annotation Projection in Legal Texts.” In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 915–926. DOI: 10.18653/v1/2020.coling-main.79. URL: https://aclanthology.org/2020.coling-main.79.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Goodfellow, Ian J., Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (June 2014). *Generative Adversarial Networks*. arXiv:1406.2661 [cs,

stat]. DOI: 10.48550/arXiv.1406.2661. URL: http://arxiv.org/abs/1406.2661 (visited on 07/21/2022).
Gutiérrez-Fandiño, Asier, Jordi Armengol-Estapé, Aitor Gonzalez-Agirre, and Marta Villegas (Oct. 2021). *Spanish Legalese Language Model and Corpora*. arXiv:2110.12201 [cs]. DOI: 10.48550/arXiv.2110.12201. URL: http://arxiv.org/abs/2110.12201 (visited on 10/26/2022).
Henderson, Peter, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho (July 2022). *Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset*. en. arXiv:2207.00220 [cs]. URL: http://arxiv.org/abs/2207.00220 (visited on 07/19/2022).
Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly (June 2019). "Parameter-Efficient Transfer Learning for NLP." In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 2790–2799. URL: https://proceedings.mlr.press/v97/houlsby19a.html.
Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson (Sept. 2020). *XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization*. arXiv:2003.11080 [cs]. DOI: 10.48550/arXiv.2003.11080. URL: http://arxiv.org/abs/2003.11080 (visited on 04/28/2023).
Hu, Yibo, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio (July 2022). "ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence." In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Place: Seattle, United States. Association for Computational Linguistics, pp. 5469–5482. DOI: 10.18653/v1/2022.naacl-main.400. URL: https://aclanthology.org/2022.naacl-main.400 (visited on 05/31/2023).
Hua, Wenyue, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber (Dec. 2022). *LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension*. arXiv:2212.08204 [cs]. DOI: 10.48550/arXiv.2212.08204. URL: http://arxiv.org/abs/2212.08204 (visited on 03/29/2023).
Hudson, G Thomas and Noura Al Moubayed (2022). "MuLD: The Multitask Long Document Benchmark." In: *arXiv preprint arXiv:2202.07362*.
Hwang, Wonseok, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo (Oct. 2022). *A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction*. en. arXiv:2206.05224 [cs]. URL: http://arxiv.org/abs/2206.05224 (visited on 04/28/2023).
Johnson, Alistair E W, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark (May 2016). "MIMIC-

III, a freely accessible critical care database." In: *Scientific data* 3, p. 160035. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.35. URL: https://europepmc.org/articles/PMC4878278.

Jurafsky, Dan and James H. Martin (2009). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall. ISBN: 9780131873216 0131873210. URL: http://www.amazon.com/Speech-Language-Processing-2nd-Edition/dp/0131873210/ref=pd_bxgy_b_img_y.

Katz, Daniel Martin, Michael James Bommarito, Shang Gao, and Pablo Arredondo (Mar. 2023). *GPT-4 Passes the Bar Exam*. en. SSRN Scholarly Paper. Rochester, NY. DOI: 10.2139/ssrn.4389233. URL: https://papers.ssrn.com/abstract=4389233 (visited on 03/27/2023).

Kawintiranon, Kornraphop and Lisa Singh (June 2022). "PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Place: Marseille, France. European Language Resources Association, pp. 7360–7367. URL: https://aclanthology.org/2022.lrec-1.801 (visited on 05/31/2023).

Khattab, Omar et al. (Oct. 2023). *DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines*. arXiv:2310.03714 [cs]. DOI: 10.48550/arXiv.2310.03714. URL: http://arxiv.org/abs/2310.03714 (visited on 10/28/2023).

Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya (Feb. 2020). "Reformer: The Efficient Transformer." In: *arXiv:2001.04451 [cs, stat]*. arXiv: 2001.04451. URL: http://arxiv.org/abs/2001.04451 (visited on 11/17/2020).

Lample, Guillaume and Alexis Conneau (Jan. 2019). *Cross-lingual Language Model Pretraining*. arXiv:1901.07291 [cs]. DOI: 10.48550/arXiv.1901.07291. URL: http://arxiv.org/abs/1901.07291 (visited on 07/04/2023).

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (Feb. 2020). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations." In: *arXiv:1909.11942 [cs]*. arXiv: 1909.11942. URL: http://arxiv.org/abs/1909.11942 (visited on 10/29/2020).

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (May 2015). "Deep learning." en. In: *Nature* 521.7553. Number: 7553 Publisher: Nature Publishing Group, pp. 436–444. ISSN: 1476-4687. DOI: 10.1038/nature14539. URL: https://www.nature.com/articles/nature14539 (visited on 10/28/2023).

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang (Sept. 2019). "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." In: *Bioinformatics*. arXiv: 1901.08746, btz682. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btz682. URL: http://arxiv.org/abs/1901.08746 (visited on 11/18/2021).

Lee-Thorp, James, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon (Sept. 2021). "FNet: Mixing Tokens with Fourier Transforms." In: *arXiv:2105.03824 [cs]*. arXiv: 2105.03824. URL: http://arxiv.org/abs/2105.03824 (visited on 10/04/2021).

Liang, Percy et al. (Nov. 2022). *Holistic Evaluation of Language Models*. arXiv:2211.09110 [cs]. DOI: 10.48550/arXiv.2211.09110. URL: http://arxiv.org/abs/2211.09110 (visited on 11/19/2022).

Licari, Daniele and Giovanni Comandè (2022). "ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law." en. In.

Liu, Nelson F., Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang (July 2023). *Lost in the Middle: How Language Models Use Long Contexts*. arXiv:2307.03172 [cs]. DOI: 10.48550/arXiv.2307.03172. URL: http://arxiv.org/abs/2307.03172 (visited on 10/29/2023).

Longpre, Shayne et al. (Feb. 2023). *The Flan Collection: Designing Data and Methods for Effective Instruction Tuning*. arXiv:2301.13688 [cs]. DOI: 10.48550/arXiv.2301.13688. URL: http://arxiv.org/abs/2301.13688 (visited on 10/28/2023).

Masala, Mihai, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu (Nov. 2021). "jurBERT: A Romanian BERT Model for Legal Judgement Prediction." In: *Proceedings of the Natural Legal Language Processing Workshop 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 86–94. DOI: 10.18653/v1/2021.nllp-1.8. URL: https://aclanthology.org/2021.nllp-1.8.

Naseem, Usman, Adam G Dunn, Matloob Khushi, and Jinman Kim (2022). "Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT." In: *BMC bioinformatics* 23.1. Publisher: BioMed Central, pp. 1–15.

OpenAI (Mar. 2023). *GPT-4 Technical Report*. arXiv:2303.08774 [cs]. DOI: 10.48550/arXiv.2303.08774. URL: http://arxiv.org/abs/2303.08774 (visited on 05/25/2023).

Pappagari, Raghavendra, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak (Oct. 2019). "Hierarchical Transformers for Long Document Classification." In: *arXiv:1910.10781 [cs, stat]*. arXiv: 1910.10781. URL: http://arxiv.org/abs/1910.10781 (visited on 06/08/2021).

Peng, Yifan, Shankai Yan, and Zhiyong Lu (2019). "Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets." In: *BioNLP@ACL*. Association for Computational Linguistics, pp. 58–65. DOI: 10.18653/v1/W19-5006.

Pfeiffer, Jonas, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych (Apr. 2021). "AdapterFusion: Non-Destructive Task Composition for Transfer Learning." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 487–503. DOI: 10.18653/v1/2021.eacl-main.39. URL: https://aclanthology.org/2021.eacl-main.39.

Pfeiffer, Jonas, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder (Nov. 2020). "MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7654–7673. DOI: 10.18653/v1/2020.emnlp-main.617. URL: https://aclanthology.org/2020.emnlp-main.617.

Roy, Aurko, Mohammad Saffar, Ashish Vaswani, and David Grangier (2021). "Efficient Content-Based Sparse Attention with Routing Transformers." In: *Transactions of the Association for Computational Linguistics* 9. Place: Cambridge, MA Publisher: MIT Press, pp. 53–68. DOI: 10.1162/tacl_a_00353. URL: https://aclanthology.org/2021.tacl-1.4 (visited on 07/21/2022).

Ruder, Sebastian et al. (Nov. 2021). "XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10215–10245. DOI: 10.18653/v1/2021.emnlp-main.802. URL: https://aclanthology.org/2021.emnlp-main.802.

Ruder, Sebastian et al. (2023). *XTREME-UP: A User-Centric Scarce-Data Benchmark for Under-Represented Languages*. _eprint: 2305.11938.

Shah, Raj, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang (Dec. 2022). "When FLUE Meets FLANG: Benchmarks and Large Pretrained Language Model for Financial Domain." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2322–2335. URL: https://aclanthology.org/2022.emnlp-main.148.

Shaham, Uri et al. (Oct. 2022). *SCROLLS: Standardized CompaRison Over Long Language Sequences*. en. arXiv:2201.03533 [cs, stat]. URL: http://arxiv.org/abs/2201.03533 (visited on 12/24/2022).

Srivastava, Aarohi et al. (June 2022). *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models*. arXiv:2206.04615 [cs, stat]. DOI: 10.48550/arXiv.2206.04615. URL: http://arxiv.org/abs/2206.04615 (visited on 04/28/2023).

Tay, Yi, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng (May 2021). "Synthesizer: Rethinking Self-Attention in Transformer Models." In: *arXiv:2005.00743 [cs]*. arXiv: 2005.00743. URL: http://arxiv.org/abs/2005.00743 (visited on 10/04/2021).

Tay, Yi, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler (Nov. 2020a). "Long Range Arena: A Benchmark for Efficient Transformers." In: *arXiv:2011.04006 [cs]*. arXiv: 2011.04006. URL: http://arxiv.org/abs/2011.04006 (visited on 10/04/2021).

Tay, Yi, Mostafa Dehghani, Dara Bahri, and Donald Metzler (Sept. 2020b). "Efficient Transformers: A Survey." In: *arXiv:2009.06732*

[cs]. arXiv: 2009.06732. URL: http://arxiv.org/abs/2009.06732 (visited on 10/07/2021).
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (Dec. 2017). "Attention Is All You Need." In: *arXiv:1706.03762 [cs]*. arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 11/17/2020).
Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." en. In: p. 30.
Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (Nov. 2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://aclanthology.org/W18-5446 (visited on 08/19/2021).
Wang, Xuezhi, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou (Mar. 2023). *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. arXiv:2203.11171 [cs]. DOI: 10.48550/arXiv.2203.11171. URL: http://arxiv.org/abs/2203.11171 (visited on 10/28/2023).
Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann (May 2023). *BloombergGPT: A Large Language Model for Finance*. arXiv:2303.17564 [cs, q-fin]. DOI: 10.48550/arXiv.2303.17564. URL: http://arxiv.org/abs/2303.17564 (visited on 05/25/2023).
Xenouleas, Stratos, Alexia Tsoukara, Giannis Panagiotakis, Ilias Chalkidis, and Ion Androutsopoulos (Sept. 2022). "Realistic Zero-Shot Cross-Lingual Transfer in Legal Topic Classification." In: *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*. SETN '22. New York, NY, USA: Association for Computing Machinery, pp. 1–8. ISBN: 978-1-4503-9597-7. DOI: 10.1145/3549737.3549760. URL: https://doi.org/10.1145/3549737.3549760 (visited on 07/04/2023).
Xu, Liang et al. (Dec. 2020). "CLUE: A Chinese Language Understanding Evaluation Benchmark." In: *COLING*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4762–4772. DOI: 10.18653/v1/2020.coling-main.419. URL: https://aclanthology.org/2020.coling-main.419.
Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (June 2021a). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Lin-

guistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (Mar. 2021b). "mT5: A massively multilingual pre-trained text-to-text transformer." en. In: *arXiv:2010.11934 [cs]*. arXiv: 2010.11934. URL: http://arxiv.org/abs/2010.11934 (visited on 11/12/2021).

Yang, Yi, Mark Christopher Siy Uy, and Allen Huang (2020). "Finbert: A pretrained language model for financial communications." In: *arXiv preprint arXiv:2006.08097*.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan (May 2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*. arXiv:2305.10601 [cs]. DOI: 10.48550/arXiv.2305.10601. URL: http://arxiv.org/abs/2305.10601 (visited on 10/28/2023).

Zaheer, Manzil et al. (July 2020). "Big Bird: Transformers for Longer Sequences." In: *arXiv:2007.14062 [cs, stat]*. arXiv: 2007.14062. URL: http://arxiv.org/abs/2007.14062 (visited on 11/02/2020).

Zhang, Ningyu et al. (May 2022). "CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark." In: *ACL*. Dublin, Ireland: Association for Computational Linguistics, pp. 7888–7915. DOI: 10.18653/v1/2022.acl-long.544. URL: https://aclanthology.org/2022.acl-long.544.

Zheng, Lucia, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho (July 2021). "When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset." In: *arXiv:2104.08671 [cs]*. arXiv: 2104.08671 version: 3. URL: http://arxiv.org/abs/2104.08671 (visited on 08/27/2021).

Part II

PUBLICATIONS

A PUBLICATIONS

This section lists the full content of the publications.

Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark

Joel Niklaus
Research Center for Digital Sustainability
Institute of Computer Science
University of Bern
Bern, Switzerland
joel.niklaus@inf.unibe.ch

Ilias Chalkidis
Coastal NLP Group
Department of Computer Science
University of Copenhagen
Copenhagen, Denmark
ilias.chalkidis@di.ku.dk

Matthias Stürmer
Research Center for Digital Sustainability
Institute of Computer Science
University of Bern
Bern, Switzerland
matthias.stuermer@inf.unibe.ch

Abstract

In many jurisdictions, the excessive workload of courts leads to high delays. Suitable predictive AI models can assist legal professionals in their work, and thus enhance and speed up the process. So far, Legal Judgment Prediction (LJP) datasets have been released in English, French, and Chinese. We publicly release a multilingual (German, French, and Italian), diachronic (2000-2020) corpus of 85K cases from the Federal Supreme Court of Switzerland (FSCS). We evaluate state-of-the-art BERT-based methods including two variants of BERT that overcome the BERT input (text) length limitation (up to 512 tokens). Hierarchical BERT has the best performance (approx. 68-70% Macro-F1-Score in German and French). Furthermore, we study how several factors (canton of origin, year of publication, text length, legal area) affect performance. We release both the benchmark dataset and our code to accelerate future research and ensure reproducibility.

1 Introduction

Frequently, legal information is available in textual form (e.g. court decisions, laws, legal articles or commentaries, contracts). With the abundance of legal texts comes the possibility of applying Natural Language Processing (NLP) techniques to tackle challenging tasks (Chalkidis and Kampas, 2018; Zhong et al., 2020; Chalkidis et al., 2021b). In this work, we study the task of Legal Judgment Prediction (LJP) where the goal is to predict the outcome (verdict) of a decision given its facts (Aletras et al., 2016; Şulea et al., 2017; Luo et al., 2017; Zhong et al., 2018; Hu et al., 2018; Chalkidis et al., 2019). Many relevant applications and tasks, such as court opinion generation (Ye et al., 2018) and analysis (Wang et al., 2012) have been also studied, while there is also work aiming to interpret (explain) the decisions of particular courts (Ye et al., 2018; Chalkidis et al., 2021a).

Models developed for LJP and relevant supportive tasks may assist both lawyers, e.g., help them prepare their arguments by identifying their strengths and weaknesses, and judges and clerks, e.g., review or prioritize cases, thus speeding up judicial processes and improving their quality. Especially in areas with many pending cases such as Indian[1] and Brazilian[2] jurisdictions or US immigration cases[3] the deployment of such models may drastically shorten the backlog. Such models can also help legal scholars to study case law (Katz, 2012) and help sociologists and research ethicists to expose irresponsible use of AI in the justice system (Angwin et al., 2016; Dressel and Farid, 2018). So far, LJP datasets have been released for English (Katz et al., 2017; Medvedeva et al., 2018; Chalkidis et al., 2019), French (Şulea et al., 2017) and Chinese (Xiao et al., 2018; Long et al., 2019).

[1] https://tinyurl.com/mjy2uf9a
[2] https://tinyurl.com/2uttucmn
[3] https://tinyurl.com/4ybhhff8

Natural Legal Language Processing Workshop 2021, pages 19–35
November 10, 2021.

We introduce a new multilingual, diachronic LJP dataset of FSCS cases, which spans 21 years (from 2000 to 2020) containing over 85K (50K German, 31K French and 4K Italian) cases. To the best of our knowledge, it is the only publicly available multilingual LJP dataset to date. Additionally, it is annotated with publication years, legal areas and cantons of origin; thus it can be used also as test-bed for fairness and robustness in the critical application of NLP to law (Wang et al., 2021).

Rogers (2021) argues that the NLP community is investing many more resources in the development of models rather than data. As a result, there are not enough challenging, high-quality and well curated benchmarks available. Rogers assumes that the main reason for this imbalance is that the ”data work“ is considered less prestigious and top conferences are more likely to reject resource (dataset) papers. With our work (and the associated code and data) we hope to make a valuable contribution to the legal NLP field, where there are not many ready-to-use benchmarks available.

Contributions

The contributions of this paper are threefold:

- We publicly release a large, high quality, curated, multilingual, diachronic dataset of 85K Swiss Federal Supreme Court (FSCS) cases annotated with the respective binarized judgment outcome (*approval/dismissal*), posing a challenging text classification task. We also provide additional metadata, i.e., the publication year, the legal area and the canton of origin per case, to promote robustness and fairness studies on the critical area of legal NLP (Wang et al., 2021).

- We provide experimental results with strong baselines representing the current state-of-the-art in NLP. Since the average length of the facts (850 tokens in the French part) is longer than the 512 tokens limit by BERT (Devlin et al., 2019), special methods are needed to cope with that. We show results comparing standard BERT models (up to 512 tokens) with two variants (hierarchical and prolonged BERT) that use up to 2048 tokens.

- We analyze the results of the German dataset in terms of diachronicity (publication year), legal area and input (text) length and the French dataset by canton of origin. We find that performance deteriorates as cases are getting more complex (longer facts), while also performance varies across legal areas. There is no sign of performance fluctuation across years.

2 Related Work

European Court of Human Rights (ECtHR)

Aletras et al. (2016) introduced a dataset of 584 ECtHR cases concerning the violation or not of three articles of the European Convention of Human Rights (ECHR). They used a Support Vector Machine (SVM) (Cortes and Vapnik, 1995) with Bag-of-Words (BoW) (n-grams) and topical features on a simplified binarized LJP. In contrast to our work, they evaluated with random 10-fold cross-validation instead of the more realistic temporal split based on the date (Søgaard et al., 2021). Medvedeva et al. (2018) extended the ECtHR dataset to include 9 instead of 3 Articles resulting in a total of approx. 11.5K cases. They also experimented with an SVM operating on n-grams on the LJP task. Chalkidis et al. (2019) experimented on a similarly sized dataset using neural methods. On the binary LJP task, they improve the state-of-the-art using a hierarchical version of BERT. Additionally, they experimented with a multi-label LJP task predicting for each of the 66 ECHR Articles whether it is violated or not.

Supreme Court of the United States (SCOTUS)

Katz et al. (2017) experimented on LJP with 28K cases from the SCOTUS spanning almost two centuries. They trained a Random Forest (Breiman, 2001) classifier using extensive feature engineering with many non textual features. Kaufman et al. (2019) improved results using an ADABoost (Freund and Schapire, 1997) classifier, while also incorporating more textual information (i.e., statements made by the court judges during oral arguments).

French Supreme Court (Court of Cassation)

Şulea et al. (2017) studied the LJP task on a dataset of approx. 127K French Supreme Court cases. They experimented on a 6-class and a 8-class setting using an SVM with BoW features. They reported very high scores, which they claim are justified by the high predictability of the French Supreme Court. Although they used as input the entire case description and not only the facts, thus there is a strong possibility of label information leak. They also used 10-fold stratified cross-validation selecting the test part at random.

German Courts

Urchs et al. (2021) present a corpus of over 32K German court decisions from 131 Bavarian courts. The corpus is annotated with rich metadata including, among others, facts and judgment outcome needed for the LJP task. They present sample experiments predicting the type of the decision (judgment, resolution or other) and detecting conclusion, definition and subsumption in a subset of 200 randomly chosen and manually annotated decisions. They used traditional Machine Learning (ML) methods such as Logistic Regression (LR) on unigrams (BoW features) and SVM on Term Frequency - Inverse Document Frequency (TF-IDF) features.

Supreme People's Court of China (SPC)

Luo et al. (2017) experimented with the Hierarchical Attention Network (Yang et al., 2016) on Chinese criminal cases. They trained a model jointly on charge prediction, a form of LJP, and the relevant criminal law article extraction task using the relevant articles as support for the charge prediction. Xiao et al. (2018) introduced a large-scale LJP dataset of more than 2.6M Chinese criminal cases from the SPC. Their dataset is annotated with extensive metadata such as applicable law articles, charges, and prison terms. Zhong et al. (2018) viewed the dependencies between the different subtasks of LJP as a Directed Acyclic Graph (DAG) and apply a topological multitask learning framework. They work on three different datasets each containing Chinese criminal cases. Long et al. (2019) studied the LJP task on 100K Chinese divorce proceedings considering three types of information as input: applicable law articles, fact description, and plaintiffs' pleas. Li et al. (2019) use a multichannel attentive neural network on four datasets containing Chinese criminal cases. They considered all three subtasks of the Chinese LJP datasets: charges, law articles and prison term. Yang et al. (2019) apply a recurrent attention network on three Chinese LJP datasets.

3 Data Description

3.1 Dataset Construction

The decisions were downloaded from the platform entscheidsuche.ch and have been pre-processed by the means of HTML parsers and Regular Expressions (RegExps). The dataset contains more than 85K decisions from the FSCS written in three languages (50K German, 31K French, 4K Italian) from the years 2000 to 2020.[4] The FSCS is the last level of appeal in Switzerland and hears only the most controversial cases which could not have been sufficiently well solved by (up to two) lower courts. In their decisions, they often focus only on small parts of previous decision, where they discuss possible wrong reasoning by the lower court. This makes these cases particularly challenging.

In order to fight the reproducibility crisis (Britz, 2020), we release the Swiss-Judgment-Prediction dataset on Zenodo[5] and on Hugging Face[6], while also open-sourcing the complete code used for constructing the dataset[7] as well as for running the experiments[8] on GitHub.

3.2 Structure of Court Decisions

A typical Swiss court decision is made up of the following four main sections: *rubrum*, *facts*, *considerations* and *rulings*.[9] The *rubrum* (introduction) contains the date and chamber, mentions the involved judge(s) and parties and finally states the topic of the decision. The *facts* describe what happened in the case and form the basis for the considerations of the court. The higher the level of appeal, the more general and summarized the facts. The *considerations* reflect the formal legal reasoning which form the basis for the final ruling. Here the court cites laws and other influential rulings. The *rulings*, constituting the final section, are an enumeration of the binding decisions made by the court. This section is normally rather short and summarizes the considerations.

3.2.1 Use of Facts instead of Considerations

We deliberately did not consider the considerations as input to the model, unlike Aletras et al. (2016) for the following reasons.The facts are the section which is most similar to a general description of the case, which may be more widely available, while

[4]The dataset is not parallel, all cases are unique and decision are written only in a single language.

[5]https://zenodo.org/record/5529712

[6]https://huggingface.co/datasets/swiss_judgment_prediction

[7]https://github.com/JoelNiklaus/SwissCourtRulingCorpus

[8]https://github.com/JoelNiklaus/SwissJudgementPrediction

[9]See examples in Figures 5 and 6 of Appendix B

Split	de			fr			it		
	approval	dismissal	total	approval	dismissal	total	approval	dismissal	total
train	8369 (24%)	27003 (76%)	35452	5197 (25%)	15982 (75%)	21179	625 (20%)	2447 (80%)	3072
val	959 (20%)	3746 (80%)	4705	649 (21%)	2446 (79%)	3095	65 (16%)	343 (84%)	408
test	1915 (20%)	7810 (80%)	9725	1264 (19%)	5556 (81%)	6820	152 (19%)	660 (81%)	812
all	11243 (23%)	38639 (77%)	49882	7110 (23%)	23984 (77%)	31094	842 (20%)	3450 (80%)	4292

Table 1: The number of cases per label (*approval*, *dismissal*) in each language subset.

being less biased.[10] Additionally, the facts do not change that much from one to the next level of appeal (apart from being more concise and summarized in the higher levels of appeal). According to estimations from several court clerks we consulted, the facts take approximately 10% of the time for drafting a decision while the considerations take 85% and the outcome 5% (45%, 50% and 5% in penal law respectively). So, most of the work being done by the judges and clerks results in the legal considerations. Therefore, we would expect the model to perform better if it had access to the considerations. But on the other hand, the value of the model would be far smaller, since most of the work is already done, once the considerations are written. Thus, to create a more realistic and challenging scenario, we consider only the facts as input for the predictive models.

3.3 The Binarized LJP Task - Verdict Labeling Simplification

The cases have been originally labeled with 6 labels: *approval*, *partial approval*, *dismissal*, *partial dismissal*, *inadmissible* and *write off*. The first four are judged on the basis of the facts (merits) and the last two for formal reasons. A case is considered *inadmissible*, if there are formal deficiencies with the appeal or if the court is not responsible to rule the case. A court rules *write off* if the case has become redundant so there is no reason for the proceeding anymore. This can be for several reasons, such as an out-of-court settlement or procedural association (two proceedings are unified). *Approval* and *partial approval* mean that the request is deemed valid or partially valid respectively. *Dismissal* and *partial dismissal* mean that the request is denied or partially denied respectively. A *partial* decision is usually ruled in parallel with a decision of the opposite kind or with *inadmissible*.

In practice, court decisions may have multiple requests (questions), where each can be judged individually. Since the structure of the outcomes in the decisions is non-standard, parsing them automatically is very challenging. Therefore, we decided to focus on the main request only and discard all side (secondary) requests. Even the main request sometimes contains multiple judgments referring to different parts of the main request, with some more important than others (it is very hard to automatically detect their criticality). So, to simplify the task and make it more concise, we transform the document labeling from a list of partial judgments into a single judgment, as follows:

1. We excluded all cases that have been ruled with both an approval and a dismissal in the main request, since that could be rather confusing.
2. We excluded cases ruled with *write off* outcomes since these cases are rejected for formal reasons that are not written (described) in the facts. Therefore, a model has no chance of inferring it correctly. We also excluded cases with *inadmissible* outcomes for similar reasons.
3. Since *partial* approvals/dismissals are very hard to distinguish from *full* approvals/dismissals respectively, we converted all the partial ones to full ones. Thus, the final labeling includes two possible outcomes, approvals and dismissals (i.e., the court "leans" positive or negative to the request).

By implementing these simplifications, we made the dataset more feasible (solvable) and semantically coherent targeting the core ruling process (see Section 5). Table 2 shows the numbers of decisions after each processing step. Note that we reduced the dataset with these preprocessing steps significantly (from over 141K to close to 85K decisions) to achieve higher quality. We also made the task structurally simpler by converting it from a multi-label to a binary classification task.[11]

The dataset is highly imbalanced containing more than $\frac{3}{4}$ dismissed cases (see Table 1 for de-

[10] Note however, that the facts are drafted together with the considerations and are often formulated in a way to support the reasoning in the considerations.

[11] Although, we look forward to recover at least part of the complexity in the future, if we have the appropriate resources to manually extract per-request judgments, introducing a new multi-task (multi-question) LJP dataset.

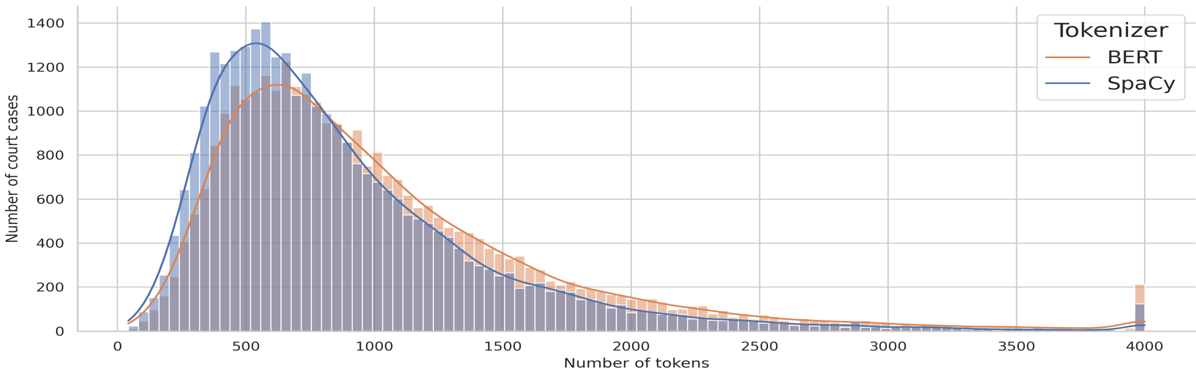


Figure 1: The distribution of the document (the facts of a case) length for French decisions. The blue histogram shows the document (case) length distribution in regular words (using the spacy tokenizer (Honnibal et al., 2020)). It is useful for a human estimation of the length and for methods building upon word embeddings (Mikolov et al., 2013; Pennington et al., 2014). The orange histogram shows the distribution in sub-word units (generated by the SentencePiece tokenizer (Kudo and Richardson, 2018) used in BERT). It is useful e.g. for estimating the maximum sequence length of a BERT-like model. Decisions with length over 4000 tokens have been grouped in the last bin.

tails). The label skewness makes the classification task quite hard and beating dummy baselines, e.g., predicting always the majority class, on micro-averaged measures (e.g., Micro-F1) is challenging. In our opinion, macro-averaged measures (e.g., Macro-F1) are more suitable in this setting, since they consider both outcomes (classes); they can also better discriminate better methods. In other words, they favor models that can actually learn the task (discriminate the two classes) and they do not always predict the majority class, i.e., *dismissal*, regardless of the facts.

Language	Total	2000-2020	Rulings	Judgments	Binarized
de	96337	95449	95273	84083	49882
fr	52278	51748	49132	49083	31094
it	8784	8643	8457	8441	4292
all	157399	155840	152862	141607	85268

Table 2: *Rulings* is the number of cases where rulings could be extracted. *Judgments* is the number of cases where we could extract any judgment types described in Section 3.3. *Binarized* is the number of cases considered in the final dataset after removing decisions containing labels other than *approval* or *dismissal*.

3.4 Case Distribution

This Section presents statistics about the distribution of cases according to different metadata like input (text) length, legal area and origin cantons.

3.4.1 The Curse of Long Documents

Figure 1 shows the distribution of the document (facts of the case) length of French cases.[12] We see that there are very few decisions with more than 2K tokens in German (very similar for Italian). The French decisions are more evenly distributed, including a large portion of decisions with more than 4K tokens. For all languages, there is a considerable portion of decisions (50%+) containing more than 512 sub-word units (BERTs maximum sequence length) posing a fundamental challenge for standard BERT models.

3.4.2 Legal Areas

Table 3 presents the distribution of legal areas across languages. The legal areas are derived from the chambers where the decisions were heard. The website of the FSCS[13] describes in detail what kinds of cases the different chambers hear.

Legal Area	de	fr	it
public law	12182 (24%)	8514 (27%)	1583 (37%)
penal law	10942 (22%)	8039 (26%)	692 (16%)
social law	10742 (22%)	4048 (13%)	673 (16%)
civil law	8208 (16%)	7348 (24%)	763 (18%)
insurance law	7625 (15%)	2950 (9%)	561 (13%)
other	183 (0.4%)	195 (0.6%)	20 (0.5%)

Table 3: The distribution of legal areas in each language subset.

3.4.3 Origin Cantons

To study robustness and fairness in terms of geographical (regional) groups, we extracted the canton of origin from the decisions. As we observe in Table 4, most of the cantons (e.g., Zürich, Ticino) are monolingual and the distribution of the cases across cantons is very skewed with 1-2 cantons per language covering a large portion of the total cases.

[12]See Figures 7 and 8 in Appendix C for the German and Italian cases, respectively.

[13]https://tinyurl.com/52a4x8yz (in German)

Canton of Origin	de	fr	it
Zürich (ZH)	**12749 (25%)**	-	-
Berne (BE)	4705 (9%)	469 (2%)	-
Lucerne (LU)	3124 (6%)	-	-
Uri (UR)	248 (0.5%)	-	-
Schwyz (SZ)	1408 (3%)	-	-
Obwalden (OW)	190 (0.4%)	-	-
Nidwalden (NW)	364 (0.7%)	-	-
Glarus (GL)	363 (0.7%)	-	-
Zug (ZG)	1321 (3%)	-	-
Fribourg (FR)	487 (1%)	1826 (6%)	-
Soleure (SO)	2022 (4%)	-	-
Basel-City (BS)	1651 (3%)	-	-
Basel-Country (BL)	1578 (3%)	-	-
Schaffhausen (SH)	591 (1%)	-	-
Appenzell Outer-Rhodes (AR)	73 (0.2%)	-	-
Appenzell Inner-Rhodes (AI)	103 (0.2%)	-	-
St. Gall (SG)	3188 (6%)	-	-
Grisons (GR)	1300 (3%)	-	85 (2%)
Argovia (AG)	5494 (11%)	-	-
Thurgovia (TG)	2066 (4%)	-	-
Ticino (TI)	-	-	**3302 (77%)**
Vaud (VD)	-	**8926 (29%)**	-
Valais (VS)	502 (1%)	2095 (7%)	-
Neuchâtel (NE)	-	1732 (6%)	-
Genève (GE)	-	**9320 (30%)**	-
Jura (JU)	-	630 (2%)	-
Swiss Confederation (CH)	1854 (4%)	348 (1%)	83 (2%)
uncategorized	4488 (9%)	5742 (18%)	818 (19%)

Table 4: The distribution of cantons of origin in each language subset. No entry means that this language is not spoken in that canton. The cantons are ordered in the official order determined by the Swiss Confederation (mostly based on the date of entry into the confederation). High-resource cantons (> 20% of decisions per language) are marked in bold. Low-resource cantons (< 5% of decisions per language) are underlined.

4 Methods

4.1 Baselines

We first experiment with three baselines. The first one is a *majority* baseline that selects the majority (*dismissal*) class always across cases. The *stratified* baseline predicts labels randomly, respecting the training distribution. The last baseline is a *linear* classifier relying on TF-IDF features for the 35K most frequent n-grams in the training set.

4.2 BERT-based methods

BERT (Devlin et al., 2019) and its variants (Yang et al., 2020; Liu et al., 2019; Lan et al., 2020), inter alia, dominate NLP as state-of-the-art in many tasks (Wang et al., 2018, 2019). Hence, we examine an arsenal of BERT-based methods.

Standard BERT We experimented with monolingual BERT models for German (Chan et al., 2019), French (Martin et al., 2020) and Italian (Parisi et al., 2020) and also the multilingual BERT of (Devlin et al., 2019). Since the facts are often longer than 512 tokens (see Section 3 for details), there is a need to adapt the models to long textual input.

Long BERT is an extension of the standard BERT models, where we extend the maximum sequence length by introducing additional positional embeddings. In our case, the additional positional encodings have been initialized by replicating the original pre-trained 512 ones 4 times (2048 in total). While Long BERT can process the full text in the majority of the cases, its extension leads to longer processing time and higher memory requirements.

Hierarchical BERT, similar to the one presented in Chalkidis et al. (2019), uses a shared standard BERT encoder processing segments up to 512 tokens to encode each segment independently. To aggregate all (in our case 4) segment encodings, we pass them through an additional Bidirectional Long Short-Term Memory (BiLSTM) encoder and concatenate the final LSTM output states to form a single document representation for classification.

5 Experiments

In this Section, we describe the conducted experiments alongside the presentation of the results and an analysis of the results of the German dataset in terms of diachronicity (judgment year), legal area, input (text) length and canton of origin.

5.1 Experimental SetUp

During training, we over-sample the cases representing the minority class (*approval*).[14] Across BERT-based methods, we use Early Stopping on development data, an initial learning rate of 3e-5 and batch size 64 across experiments. The standard BERT models have been trained and evaluated with maximum sequence length 512 and the two variants of BERT with maximum sequence length 2048. The 2048 input length has been chosen based on a balance between memory and compute restrictions and the statistics of the length of facts (see Section 3.4.1), where we see that the vast majority of cases contains less than 2K tokens. Additionally, this gives us the possibility to investigate differences by input (text) length (see Section 5.3.2). We report both micro- and macro-averaged F1-score on the test set. Micro-F1 is averaged across samples whereas Macro-F1 is averaged across samples inside each class and then across the classes. Therefore, a test example in

[14] In preliminary experiments, we find that this sampling methodology outperforms both the standard Empirical Risk Minimization (ERM) and the class-wise weighting of the loss penalty, i.e., considering each class loss 50-50.

Model	de		fr		it	
	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
baselines						
Majority	**80.3**	44.5	**81.5**	44.9	**81.3**	44.8
Stratified	66.7 ± 0.3	50.0 ± 0.4	66.3 ± 0.2	50.0 ± 0.4	69.9 ± 1.8	48.8 ± 2.4
Linear (BoW)	65.4 ± 0.2	52.6 ± 0.1	71.2 ± 0.1	56.6 ± 0.2	67.4 ± 0.5	53.9 ± 0.6
standard (up to 512 tokens)						
Native BERT	74.0 ± 4.0	63.7 ± 1.7	74.7 ± 1.8	58.6 ± 0.9	76.1 ± 3.7	55.2 ± 3.7
Multilingual BERT	68.4 ± 5.1	58.2 ± 4.8	71.3 ± 4.3	55.0 ± 0.8	77.6 ± 2.4	53.0 ± 1.1
long (up to 2048 tokens)						
Native BERT	76.5 ± 3.7	67.9 ± 1.8	77.2 ± 3.4	68.0 ± 1.8	77.1 ± 3.9	$\mathbf{59.8 \pm 4.6}$
Multilingual BERT	75.9 ± 1.6	66.5 ± 0.8	73.3 ± 1.9	64.3 ± 1.5	76.0 ± 2.6	58.4 ± 3.5
hierarchical (two-tier $4\times$ 512 tokens)						
Native BERT	77.1 ± 3.7	$\mathbf{68.5 \pm 1.6}$	80.2 ± 2.0	$\mathbf{70.2 \pm 1.1}$	75.8 ± 3.5	57.1 ± 6.1
Multilingual BERT	76.8 ± 3.2	57.1 ± 0.8	76.3 ± 4.1	67.2 ± 2.9	72.4 ± 16.6	55.5 ± 9.5

Table 5: All the models have been trained and evaluated in the same language. With *Native BERT* we mean the BERT model pre-trained in the respective language. The best scores for each language are in bold. Given the high class imbalance, BERT-based methods under-perform in Micro-F1 compared to the *Majority* baseline, while being substantially better in Macro-F1.

a minority class has a higher weight in Macro-F1 than an example from the majority class. In classification problems with imbalanced class distributions (such as the one we examine), Macro-F1 is more realistic than Micro-F1 given that we are equally interested in both classes. Each experiment has been run with 5 different random seeds. We report the average score and standard deviation across experiments. The experiments have been performed on a single GeForce RTX 3090 GPU with mixed precision and gradient accumulation. We used the Hugging Face Transformers library (Wolf et al., 2020) and the BERT models available from `https://huggingface.co/models`.

5.2 Main Results

Table 5 shows the results across methods for all language subsets. We observe that the native BERT models outperform their multi-lingual counterpart; while not being domain-specific, these models can still better model the case facts. Given the high class imbalance, all BERT-based methods under-perform in Micro-F1, being biased towards *dismissal* performance compared to the naive Majority baseline, while doing substantially better in Macro-F1. Hierarchical and Long BERT-based methods consistently out-perform the linear classifiers across languages (+10% in Macro-F1), while standard BERT is comparable or better than linear models, although it considers only up to 512 tokens. While performance of BERT-based methods is quite comparable between the German and French subsets with 35K and 21K training samples respectively, it is far worse in the Italian subset, where there are only 3K training samples. In two out of three languages (German and French with 20K+ training samples) hierarchical BERT has borderline better performance compared to long BERT (+1.6-2.2% in Macro-F1), but in both cases the difference is very close to the error margin (standard deviation). We would like to remark that the results of Hierarchical BERT could possibly be improved considering a finer segmentation of the text into sentences or paragraphs.[15] We leave the investigation for alternative segmentation schemes for future work.

5.3 Discussion - Bivariate Analysis

In this section, we analyze the results in relation to specific attributes (publication year, input (text) length, legal area and canton of origin) in order to evaluate the model robustness and identify how specific aspects affect the model performance.

[15] Currently, we segment the text into chunks of 512 tokens to avoid excessive padding that will further increase the needed number of segments and will lead to even higher time and memory demands.

Legal Area			standard		long		hierarchical	
Legal Area	# cases	approval rate	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
public law	2587	20.6%	66.6 ± 6.2	53.1 ± 1.8	64.6 ± 6.7	53.8 ± 2.1	64.8 ± 8.1	53.7 ± 3.0
penal law	2900	21.0%	83.6 ± 1.8	74.8 ± 1.5	87.6 ± 1.6	81.1 ± 2.3	88.4 ± 1.0	82.6 ± 2.5
social law	661	19.3%	71.1 ± 4.3	65.2 ± 2.6	74.8 ± 4.0	69.1 ± 2.8	75.4 ± 3.9	69.4 ± 2.5
civil law	1574	16.5%	73.6 ± 4.8	55.5 ± 1.0	79.0 ± 3.4	65.1 ± 2.4	78.9 ± 3.8	65.9 ± 2.8

Table 6: We used the German native BERT model pre-trained and evaluated on the German data. In the German test set there are no insurance law cases and only 3 cases with other legal areas. The area where models perform best is in bold and the area where they perform worst is underlined.

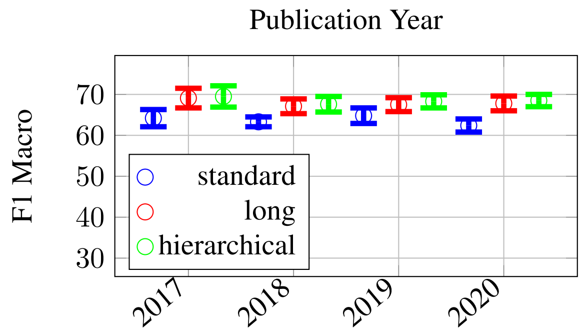


Figure 2: This table compares the different BERT types on cases from different years. We used the native German BERT model.

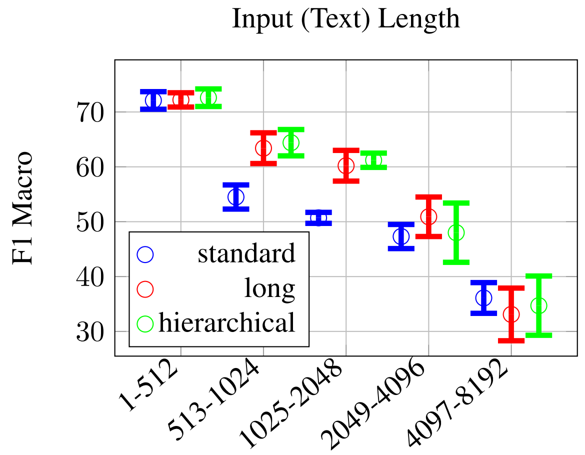


Figure 3: This table compares the different long BERT types on different input (text) lengths. We used the native German BERT model.

5.3.1 Diachronicity

In Figure 2, we present the results grouped by years in the test set (2017-2020). We cannot identify a notable fluctuation in performance across years as there is a very small decrease in performance (approx. -2% in Macro-F1); most probably because the testing time-frame is really short (4 years). Comparing the performance between the validation (2015-2016) and the test (2017-2020) set (approx. 70% vs. 68.5%), again we do not observe an exceptional fluctuation time-wise.

5.3.2 Input (Text) Length

In Figure 3, we observe that model performance deteriorates as input (text) length increases, i.e., there is an absolute negative correlation between performance and input (text) length. The two variants of BERT improve results, especially in cases with 512 to 2048 tokens. Since the two variants of BERT have a maximum length of 2048 they perform similar to the standard BERT type in cases longer than 2048 tokens.

5.3.3 Legal Area

In Table 6, we observe that the models do not equally perform across legal areas. All models seem to be much more accurate in penal law cases, while the performance is much worse (approx. 30%) in public law cases. According to the experts, the jurisprudence in penal law is more united and aligned in Switzerland and outlier judgments are rarer making the task more predictable. Additionally, in the case of not enough evidence the principle of *"in dubio pro reo"* (reasonable doubt) is applied. [16] Another possible reason for the higher performance in penal law could be the increased work performed by the legal clerks in drafting the facts of the case (see Section 3.2.1), thus including more useful information relevant to the task.

5.3.4 Canton of Origin

In Figure 4, we observe a performance disparity across cantons, although this is neither correlated with the number of cases per canton, nor with the dismissal/approval rate per canton. Thus, the disparity is either purely coincidental and has to do with the difficulty of particular cases in some cantons or there are other factors (e.g., societal, economics) worth considering in future work.

[16] The principle of *"in dubio pro reo"*, i.e., "When in doubt, in favor of the defendant.", is only applicable in penal law cases.

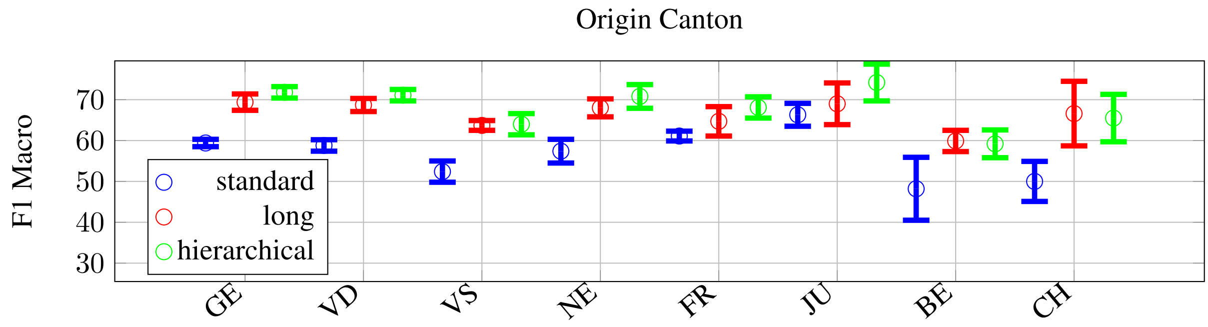


Figure 4: This table compares the different long BERT types on different origin cantons. We used the native French BERT model. The cantons are sorted by the number of cases in the training set descending.

6 Conclusions & Future Work

We introduced a new multilingual, diachronic dataset of 85K Swiss Federal Supreme Court (FSCS) cases, including cases in German, French, and Italian. We presented results considering three alternative BERT-based methods, including methods that can process up to 2048 tokens and thus can read the entirety of the facts in most cases. We found that these methods outperform the standard BERT models and have the best results in Macro-F1, while the naive majority classifier has the best overall results in Micro-F1 due to the high class imbalance of the dataset (more than $\frac{3}{4}$ of the cases are dismissed). Further on, we presented a bivariate analysis between performance and multiple factors (diachronicity, input (text) length, legal area, and canton of origin). The analysis showed that performance deteriorates as input (text) length increases, while the results in cases from different legal areas or cantons vary raising questions on models' robustness under different attributes.

In future work, we would like to investigate the application of cross-lingual transfer learning techniques, for example the use of Adapters (Houlsby et al., 2019; Pfeiffer et al., 2020). In this case, we could possibly improve the poor performance in the Italian subset, where approx. 3K cases exists, by training a multilingual model across all languages, thus exploiting all available resources, ignoring the traditional language barrier. In the same direction, we could also exploit and transfer knowledge from other annotated datasets that aim at the LJP task (e.g., ECtHR and SCOTUS).

More in depth analysis on robustness is also an interesting future avenue. In this direction, we would like to explore distributional robust optimization (DRO) techniques (Koh et al., 2021; Wang et al., 2021) that aim to mitigate disparities across groups of interest, i.e., labels, cantons and/or legal areas could be both considered in this framework.

Another interesting direction is a deeper analysis with models handling long textual input (Beltagy et al., 2020; Zaheer et al., 2020) using alternative attention schemes (window-based, dilated, etc.). Furthermore, none of the examined pre-trained models is legal-oriented, thus pre-training and evaluating such specialized models is also needed, similarly to the English Legal-BERT of Chalkidis et al. (2020).

Ethics Statement

The scope of this work is not to produce a robot lawyer, but rather to study LJP in order to broaden the discussion and help practitioners to build assisting technology for legal professionals. We believe that this is an important application field, where research should be conducted (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize law, while also highlight (inform the audience on) the various multi-aspect shortcomings seeking a responsible and ethical (fair) deployment of technology. In this direction, we provide a well-documented public resource for three languages (German, French, and Italian) that are underrepresented in legal NLP literature. We also provide annotations for several attributes (year of publication, legal area, canton/region) and provide a bivariate analysis discussing the shortcomings to further promote new studies in terms of fairness and robustness (Wang et al., 2021), a critical part of NLP application in law. All decisions (original material) are publicly available on the entscheidsuche.ch platform and the names of the parties have been redacted (See Figures 5 and 6) by the court according to its official guidelines[17].

[17] https://tinyurl.com/mtu23szy (In German)

Acknowledgements

This work has been supported by the Swiss National Research Programme "Digital Transformation" (NRP-77)[18] grant number 187477. This work is also partly funded by the Innovation Fund Denmark (IFD)[19] under File No. 0175-00011A. We would like to thank: Daniel Kettiger, Magda Chodup, and Thomas Lüthi for their legal advice, Adrian Jörg for help in coding, and Entscheidsuche.ch for providing the data.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93. Publisher: PeerJ Inc.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*. ArXiv: 2004.05150.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Denny Britz. 2020. AI Research, Replicability and Incentives.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021a. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, online.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021b. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English.

Ilias Chalkidis and Dimitrios Kampas. 2018. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27:171–198.

Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. 2019. deepset - Open Sourcing German BERT.

C. Cortes and V. Vapnik. 1995. Support vector networks. *Machine Learning*, 20:273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(10).

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, Long Beach, CA, USA.

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-Shot Charge Prediction with Discriminative Legal Attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Daniel Martin Katz. 2012. Quantitative legal prediction-or-how I learned to stop worrying and start preparing for the data-driven future of the legal services industry. *Emory Law Journal*, 62:909.

Daniel Martin Katz, Michael J. Bommarito Ii, and Josh Blackman. 2017. A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE*, 12(4):e0174698. Publisher: Public Library of Science.

[18] `https://www.nfp77.ch/en/`

[19] `https://innovationsfonden.dk/en`

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. 2021. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. *arXiv:1808.06226 [cs]*. ArXiv: 1808.06226.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*. ArXiv: 1909.11942.

Shang Li, Hongli Zhang, Lin Ye, Xiaoding Guo, and Binxing Fang. 2019. MANN: A Multichannel Attentive Neural Network for Legal Judgment Prediction. *IEEE Access*, 7:151144–151155. Conference Name: IEEE Access.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic Judgment Prediction via Legal Reading Comprehension. In *Chinese Computational Linguistics*, Lecture Notes in Computer Science, pages 558–572, Cham. Springer International Publishing.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to Predict Charges for Criminal Cases with Legal Basis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2727–2736, Copenhagen, Denmark. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte De La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a Tasty French Language Model. pages 7203–7219.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the European Court of Human Rights: Looking into the crystal ball. In *Proceedings of the Conference on Empirical Legal Studies*, page 24.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*. ArXiv: 1301.3781.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. UmBERTo: an Italian Language Model trained with Whole Word Masking. Original-date: 2020-01-10T09:55:31Z.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online.

Anna Rogers. 2021. Changing the World by Changing the Data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2182–2194, Online. Association for Computational Linguistics.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. 2021. We need to talk about random splits. In *Proceedings of the 2021 Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Online.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2021. Design and Implementation of German Legal Decision Corpora:. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, pages 515–521, Online Streaming, — Select a Country —. SCITEPRESS - Science and Technology Publications.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. page 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

William Yang Wang, Elijah Mayfield, Suresh Naidu, and Jeremiah Dittmar. 2012. Historical Analysis of Legal Opinions with a Sparse Mixed-Effects Latent Variable Model. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 740–749, Jeju Island, Korea. Association for Computational Linguistics.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *Science China - Information Sciences*, abs/2103.13868.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv:1807.02478 [cs]*. ArXiv: 1807.02478.

Ze Yang, Pengfei Wang, Lei Zhang, Linjun Shou, and Wenwen Xu. 2019. A Recurrent Attention Network for Judgment Prediction. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, Lecture Notes in Computer Science, pages 253–266, Cham. Springer International Publishing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*. ArXiv: 1906.08237.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal Judgment Prediction via Topological Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

A Training Effort

Type	BERT	RoBERTa
standard	3.377E+11	3.398E+11
long	1.365E+12	1.374E+12
hierarchical	1.476E+12	1.477E+12

Table 7: This table shows the total floating point operations per epoch per training example used for training each type. Each model has been trained for 2 to 4 epochs (variable because of early stopping). This table can be used to choose a suitable model with limited resources. Additionally, it can be used to measure the environmental impact.

Table 7 shows the training effort required for finetuning each type. Training one of the types capable of handling long input results in 4 to 5 times more training operations compared to the standard model. This seems justifiable since the gain from the longer models in terms of F1 score is considerable. Also, the entire cost of finetuning is relatively small.

B Examples

In this appendix we show some examples of court decisions with their respective labels. Figure 5 shows an example of a dismissed decision and Figure 6 an example of an approved decision. Both decisions are relatively short, but still contain all sections (rubrum, facts, considerations and judgments). They are both very recent, dating from 2019 and 2017 respectively.

C Input Length Distribution

In this appendix we show the input length distributions for the German (Figure 7) and Italian (Figure 8) datasets. We observe that the average Italian decision is longer than the average German decision. Additionally, there is also a higher density in moderately long decisions (over 1000 tokens) and there are many more decisions over 4000 tokens. Apart from the availability of more training data in the German dataset, the shorter decisions may also be an important factor in the better performance we see in most models trained on the German dataset in comparison to the Italian case and to some extent the French case (see Table 5).

D Tables to Plots

In this appendix, we show tables belonging to plots in the main paper to show the exact numbers. Table 8 shows the results regarding the different input lengths. Table 9 shows the results regarding different years in the test set. Table 10 shows the model performance across different cantons.

E Training with Class Weights

In this appendix we show the results of training the models with class weights instead of oversampling. Table 11 shows the training results. We notice, that for many configurations (especially with XLM-R), the model only learns the majority classifier. This leads to a very low Macro-F1 score. We also experimented with undersampling as an alternative to oversampling, but saw similar results to the training with class weights.

F Classifier Confidence

In this appendix, we discuss the reliability of the confidence scores of the classifier output alongside the predictions. The confidence scores are computed by taking the softmax on the classifier outputs, so that we get a probability (confidence) score of a given class between 0 and 100. The hierarchical and long BERT types show an increase in both the confidence in the correct predictions and the incorrect predictions compared to the standard BERT type (with the increase in the correct predictions being more pronounced). This finding holds across all three languages.

Model	standard		long		hierarchical	
	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
1-512 (5479 decisions)	81.1 ± 2.7	72.1 ± 1.6	80.8 ± 2.5	72.2 ± 1.3	39.3 ± 37.2	25.1 ± 17.4
513-1024 (3364 decisions)	65.3 ± 6.2	65.3 ± 6.2	71.8 ± 5.4	63.4 ± 2.8	43.3 ± 30.8	30.5 ± 13.2
1025-2048 (788 decisions)	63.8 ± 4.9	50.7 ± 1.0	69.1 ± 5.4	60.2 ± 2.8	54.9 ± 26.7	37.2 ± 15.3
2049-4096 (82 decisions)	64.9 ± 6.7	47.3 ± 2.2	65.1 ± 9.2	50.9 ± 3.6	60.2 ± 13.3	48.0 ± 5.4
4097-8192 (12 decisions)	56.7 ± 7.0	36.1 ± 2.8	50.0 ± 10.2	33.1 ± 4.8	50.0 ± 11.8	34.7 ± 5.4

Table 8: Results on the German data grouped by text length. Performance deteriorates as text length is increased.

Model	standard		long		hierarchical	
	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
2017	73.9 ± 4.2	64.2 ± 2.1	77.1 ± 3.9	69.1 ± 2.4	77.4 ± 3.9	69.5 ± 2.6
2018	74.2 ± 3.8	63.3 ± 1.2	76.6 ± 3.7	67.1 ± 1.8	76.7 ± 4.0	67.6 ± 1.9
2019	74.5 ± 4.0	64.8 ± 1.9	76.0 ± 3.7	67.5 ± 1.7	76.9 ± 3.8	68.3 ± 1.6
2020	73.5 ± 4.2	62.4 ± 1.6	76.6 ± 3.4	67.8 ± 1.8	77.4 ± 3.1	68.5 ± 1.5

Table 9: We used the German native BERT model pretrained and evaluated on the German data.

Canton			standard		long		hierarchical	
Canton	# cases	approval rate	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
Berne (BE)	332	9.5%	79.4 ± 4.6	48.2 ± 7.7	78.7 ± 4.7	59.9 ± 2.6	78.5 ± 2.7	59.2 ± 3.4
Fribourg (FR)	1121	14.7%	76.7 ± 3.1	61.1 ± 1.2	75.8 ± 5.2	64.7 ± 3.6	79.5 ± 3.4	68.1 ± 2.6
Vaud (VD)	5684	17.0%	76.0 ± 1.8	58.8 ± 1.4	78.9 ± 3.0	68.7 ± 1.6	82.5 ± 1.7	71.1 ± 1.4
Valais (VS)	1399	20.6%	75.1 ± 1.0	52.4 ± 2.6	75.0 ± 2.6	63.7 ± 1.2	76.1 ± 3.3	64.0 ± 2.6
Neuchâtel (NE)	1226	14.9%	76.2 ± 3.6	57.4 ± 2.9	79.0 ± 3.9	68.0 ± 2.2	82.3 ± 2.7	70.8 ± 2.9
Genève (GE)	6017	21.8%	72.0 ± 3.1	59.4 ± 0.9	76.0 ± 3.3	69.4 ± 2.0	79.4 ± 2.3	71.8 ± 1.7
Jura (JU)	425	15.7%	80.1 ± 3.2	66.3 ± 2.8	78.9 ± 5.8	69.0 ± 5.1	83.8 ± 4.3	74.2 ± 4.5
Swiss Confederation (CH)	227	26.7%	70.0 ± 2.7	50.0 ± 4.9	72.0 ± 8.7	66.6 ± 7.9	73.3 ± 4.4	65.5 ± 5.8

Table 10: We used the French native BERT model pretrained and evaluated on the French data. The number of cases is counted on the training set per canton. The approval rate is calculated on the test set.

Model	de		fr		it	
	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑	Micro-F1↑	Macro-F1↑
baselines						
Most Frequent	80.3	44.5	81.5	44.9	81.3	44.8
Stratified	66.7 ± 0.3	50 ± 0.4	66.3 ± 0.2	50 ± 0.4	69.9 ± 1.8	48.8 ± 2.4
Uniform	50 ± 0.3	44.8 ± 0.4	50 ± 0.6	44.5 ± 0.5	49.7 ± 2.4	44 ± 2.3
standard						
Native BERT	71.1 ± 3.3	62.6 ± 1.6	72.8 ± 5.5	58.2 ± 1.2	67 ± 13.1	49.4 ± 5.1
XLM-RoBERTa	77.8 ± 6.3	47.3 ± 6.3	76.1 ± 7.4	48.4 ± 4.9	80.4 ± 1.9	44.7 ± 0.4
long						
Native BERT	81.9 ± 1.2	**69.5 ± 0.9**	**81.8 ± 1.5**	69.4 ± 1.7	80.2 ± 1.4	46.1 ± 2.2
XLM-RoBERTa	**81.5 ± 0.7**	59.4 ± 9.6	81.5 ± 0.5	51.3 ± 8.8	81.3	44.8
hierarchical						
Native BERT	78.6 ± 2.1	69.2 ± 0.6	79.3 ± 0.8	**70 ± 0.7**	80.6 ± 1.1	**50.5 ± 6.5**
XLM-RoBERTa	80.3	44.5	80.3 ± 1.8	49.6 ± 9.8	81.3	44.8

Table 11: All the models have been trained and evaluated in the same language. With *Native BERT* we mean the BERT model pretrained in the respective language. The *Most Frequent* baseline just selects the majority class always. The *Stratified* baseline predicts randomly, respecting the training distribution. The best scores for each language are in bold. To combat label imbalance, we weighted the minority class samples more in the loss function.

Model	de		fr		it	
	Correct↑	Incorrect↓	Correct↑	Incorrect↓	Correct↑	Incorrect↓
standard	75.8 ± 13.6	64.7 ± 10.6	71.9 ± 12.2	64.4 ± 9.8	77.6 ± 12.2	68.3 ± 11.3
long	78.9 ± 12.2	65.8 ± 10.9	78.3 ± 11.6	67.8 ± 11.0	81.2 ± 11.2	68.4 ± 10.5
hierarchical	86.6 ± 15.9	69.3 ± 13.6	85.9 ± 15.2	70.8 ± 13.9	88.7 ± 14.7	71.4 ± 13.4

Table 12: This table shows the average confidence scores (0-100) of the different types of multilingual BERT models on the test set for correct and incorrect predictions respectively. Both the mean and standard deviation are averaged over 5 random seeds. The model has been finetuned on the entire dataset (all languages) and evaluated on the respective language.

Bundesgericht
Tribunal fédéral
Tribunale federale
Tribunal federal

Rubrum

5F_5/2019

Urteil vom 28. Mai 2019

II. zivilrechtliche Abteilung

Besetzung
Bundesrichter Herrmann, Präsident,
Bundesrichter von Werdt, Bovey,
Gerichtsschreiber Möckli.

Verfahrensbeteiligte
A.________,
Gesuchsteller.

Gegenstand
Gesuch um Revision des bundesgerichtlichen Urteils 5A_89/2018 vom 5. Februar 2018.

facts

Sachverhalt:
Gegen die KESB Pfäffikon, diverse Sozialdienste, die Postfinance, Sozialversicherungen, eine Einwohnergemeinde, verschiedene Versicherungsgesellschaften, Banken und weitere Gesellschaften sowie mehrere Scientology Kirchen und andere religiöse Vereinigungen erhebt A.________ eine als subsidiäre Verfassungsbeschwerde betitelte Eingabe, mit welcher er die Revision des bundesgerichtlichen Urteils 5A_89/2018 vom 5. Februar 2018 verlangt. Ferner verlangt er die Verurteilung der Gegenparteien wegen Persönlichkeitsverletzung, die Aufhebung eines Erbenscheines, die Überweisung von IV-Renten, die Erstellung von Abschlussrechnungen und vieles mehr.

considerations

Erwägungen:

1.
Der Gesuchsteller zählt zwar verschiedene Revisionsgründe auf. Indes begründet er mit keinem Wort, inwiefern ein Revisionsgrund vorliegen soll. Ebenso wenig äussert er sich zur Einhaltung der Fristen (Art. 124 Abs. 1 BGG). Somit ist auf das Revisionsgesuch nicht einzutreten (Art. 42 Abs. 2 BGG).

2.
Soweit die Eingabe auch einen Beschwerdecharakter haben sollte, wäre darauf ebenfalls nicht einzutreten. Es ist zwar die Rede von "Urteilen des Obergerichts des Kantons Zürich" und solche könnten grundsätzlich Anfechtungsobjekt sein (Art. 75 Abs. 1 BGG). Jedoch werden diese nicht näher bezeichnet und es liegt auch keine Beschwerdebegründung im Sinn von Art. 42 Abs. 2 BGG vor, wenn festgehalten wird, die Beschwerde schütze die körperliche und geistige Unversehrtheit vor ausländischen Vereinen und Folgedelikten der ausserordentlichen Gerichte, um entsprechend den Beweisen, welche sich jeden Tag mit Völkermord beweisen, die Rechtsgleichheit und Unversehrtheit entsprechend der Mehrheit zu beweisen.

3.
Die Urteile des Bundesgerichtes erfolgen grundsätzlich im schriftlichen Verfahren (zu den Ausnahmen vgl. Art. 57 und 58 BGG), weshalb der Antrag auf eine öffentliche Verhandlung abzuweisen ist.

4.
Die Gerichtskosten sind dem Gesuchsteller aufzuerlegen (Art. 66 Abs. 1 BGG).

rulings

Demnach erkennt das Bundesgericht:

1.
Der Antrag auf öffentliche Verhandlung wird abgewiesen.

2.
Auf das Revisionsgesuch wird nicht eingetreten.

3.
Soweit eine Beschwerde erhoben werden sollte, wird auf diese nicht eingetreten.

4.
Die Gerichtskosten von Fr. 1'000.-- werden dem Gesuchsteller auferlegt.

5.
Dieses Urteil wird dem Gesuchsteller schriftlich mitgeteilt.

Lausanne, 28. Mai 2019

Im Namen der II. zivilrechtlichen Abteilung
des Schweizerischen Bundesgerichts

Der Präsident: Herrmann

Der Gerichtsschreiber: Möckli

Figure 5: This is an example of a dismissed decision: https://tinyurl.com/n44hathc

Bundesgericht
Tribunal fédéral
Tribunale federale
Tribunal federal

Rubrum

9C_502/2017

Urteil vom 21. September 2017

II. sozialrechtliche Abteilung

Besetzung
Bundesrichterin Pfiffner, Präsidentin,
Bundesrichterin Glanzmann, Bundesrichter Parrino,
Gerichtsschreiberin Oswald.

Verfahrensbeteiligte
A.________,
vertreten durch Rechtsanwalt Jan Herrmann,
Beschwerdeführerin,

gegen

IV-Stelle Basel-Stadt,
Lange Gasse 7, 4052 Basel,
Beschwerdegegnerin.

Gegenstand
Invalidenversicherung
(vorinstanzliches Verfahren; Prozessvoraussetzung),

Beschwerde gegen den Entscheid des Sozialversicherungsgerichts des Kantons Basel-Stadt vom 10. Juni 2017 (IV.2016.186).

facts

Nach Einsicht
in die Beschwerde vom 18. Juli 2017 (Poststempel) gegen den Entscheid des Sozialversicherungsgerichts des Kantons Basel-Stadt vom 10. Juni 2017, mit welchem auf das Gesuch vom 21. November 2016 um Revision des Entscheids vom 11. Oktober 2016 nicht eingetreten wurde,

considerations

in Erwägung,
dass das kantonale Gericht erkannte, das einschlägige Prozessrecht (§ 18 Abs. 1 lit. a des Gesetzes über das Sozialversicherungsgericht des Kantons Basel-Stadt und über das Schiedsgericht in Sozialversicherungssachen vom 9. Mai 2001 [Sozialversicherungsgerichtsgesetz, SVGG/BS SGS 154.200]) sehe die Revision unter anderem bei Entdeckung neuer erheblicher Tatsachen oder Beweismittel vor (Art. 61 lit. i ATSG),
dass es erwog, beim Revisionsgesuch handle es sich um ein ausserordentliches Rechtsmittel, das sich gegen einen rechtskräftigen Beschwerdeentscheid richte, ein solcher jedoch vorliegend nicht

gegeben sei, da die Gesuchstellerin gegen den Entscheid des Sozialversicherungsgerichts Basel-Stadt vom 11. Oktober 2016 Beschwerde beim Bundesgericht erhoben habe,
dass diese Beschwerde vom 21. November 2016 beim Bundesgericht noch hängig ist (9C_782/2016),
dass eine Vorinstanz des Bundesgerichts auf ein Revisionsgesuch nicht einzig mit der Begründung nicht eintreten darf, gegen den zu revidierenden Entscheid sei Beschwerde beim Bundesgericht erhoben worden (**BGE 138 II 386** E. 6 S. 389 ff.; Urteil 8C_921/2014 vom 12. Mai 2015 E. 2.3),
dass die Beschwerde damit offensichtlich begründet und deshalb im Verfahren nach Art. 109 Abs. 2 lit. b BGG mit summarischer Begründung (Art. 109 Abs. 3 Satz 1 BGG) gutzuheissen ist,
dass umständehalber auf die Erhebung von Gerichtskosten zu verzichten ist (Art. 66 Abs. 1 Satz 2 BGG),

Demnach erkennt das Bundesgericht:

1.
Die Beschwerde wird gutgeheissen. Der Entscheid des Sozialversicherungsgerichts des Kantons Basel-Stadt vom 10. Juni 2017 wird aufgehoben. Die Sache wird an die Vorinstanz zurückgewiesen, damit sie über die übrigen Eintretensvoraussetzungen bezüglich des Gesuchs vom 21. November 2016 entscheide und dieses gegebenenfalls materiell behandle.

2.
Es werden keine Gerichtskosten erhoben.

3.
Die Beschwerdegegnerin hat die Beschwerdeführerin für das bundesgerichtliche Verfahren mit Fr. 2'000.- zu entschädigen.

4.
Dieses Urteil wird den Parteien, dem Sozialversicherungsgericht des Kantons Basel-Stadt und dem Bundesamt für Sozialversicherungen schriftlich mitgeteilt.

Luzern, 21. September 2017

Im Namen der II. sozialrechtlichen Abteilung
des Schweizerischen Bundesgerichts

Die Präsidentin: Pfiffner

Die Gerichtsschreiberin: Oswald

Figure 6: This is an example of an approved decision: https://tinyurl.com/mjxfjn65

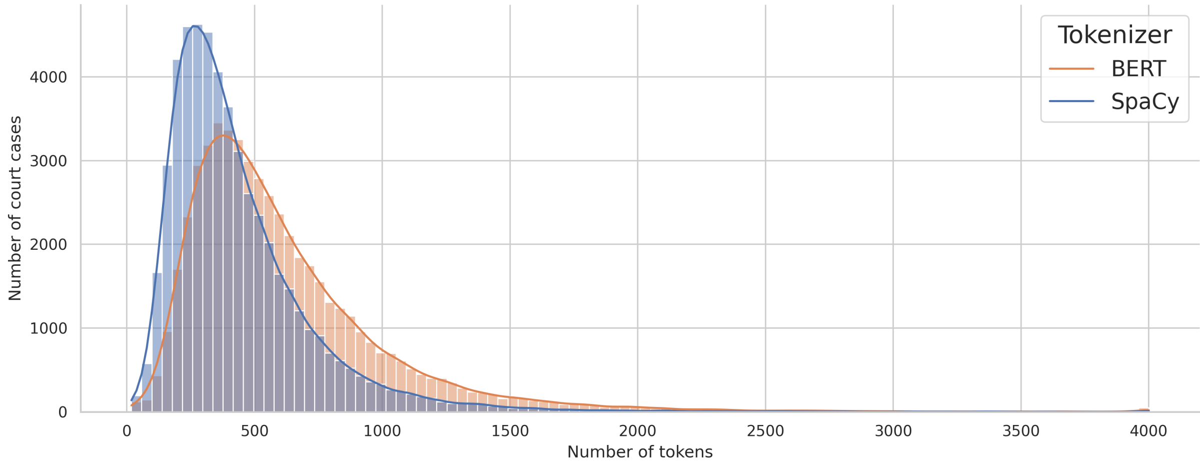


Figure 7: This histogram shows the distribution of the input length for German decisions. The blue histogram is generated from tokens generated by the spacy tokenizer (regular words). The orange histogram is generated from tokens generated by the SentencePiece tokenizer used in BERT (subword units). Decisions with length over 4000 tokens are grouped in the last bin (before 4000).

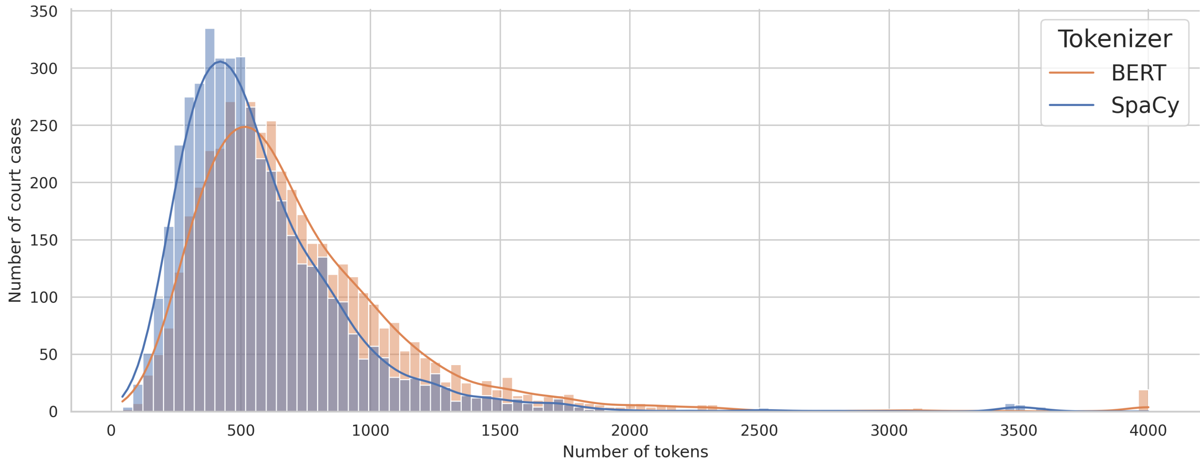


Figure 8: This histogram shows the distribution of the input length for Italian decisions. The blue histogram is generated from tokens generated by the spacy tokenizer (regular words). The orange histogram is generated from tokens generated by the SentencePiece tokenizer used in BERT (subword units). Decisions with length over 4000 tokens are grouped in the last bin (before 4000).

An Empirical Study on Cross-X Transfer for Legal Judgment Prediction

Joel Niklaus † * **Matthias Stürmer** † **Ilias Chalkidis** ‡ ◇ *
† Institute of Computer Science, University of Bern, Switzerland
‡ Department of Computer Science, University of Copenhagen, Denmark
◇ Cognitiv+, Athens, Greece

Abstract

Cross-lingual transfer learning has proven useful in a variety of Natural Language Processing (NLP) tasks, but it is understudied in the context of legal NLP, and not at all in Legal Judgment Prediction (LJP). We explore transfer learning techniques on LJP using the trilingual Swiss-Judgment-Prediction dataset, including cases written in three languages. We find that cross-lingual transfer improves the overall results across languages, especially when we use adapter-based fine-tuning. Finally, we further improve the model's performance by augmenting the training dataset with machine-translated versions of the original documents, using a 3× larger training corpus. Further on, we perform an analysis exploring the effect of cross-domain and cross-regional transfer, i.e., train a model across domains (legal areas), or regions. We find that in both settings (legal areas, origin regions), models trained across all groups perform overall better, while they also have improved results in the worst-case scenarios. Finally, we report improved results when we ambitiously apply cross-jurisdiction transfer, where we further augment our dataset with Indian legal cases.

1 Introduction

Rapid development in Cross-Lingual Transfer (CLT) has been achieved by pre-training transformer-based models in large multilingual corpora (Conneau et al., 2020; Xue et al., 2021), where these models have state-of-the-art results in multilingual NLU benchmarks (Ruder et al., 2021). Moreover, adapter-based fine-tuning (Houlsby et al., 2019; Pfeiffer et al., 2020) has been proposed to minimize the misalignment of multilingual knowledge (alignment) when CLT is applied, especially in a zero-shot fashion, where the target language is unseen during training. CLT is severely understudied in legal NLP applications except for Chalkidis et al. (2021) who experimented with several methods for CLT on MultiEURLEX, a newly introduced multilingual legal topic classification dataset, including EU laws.

* Equal contribution.

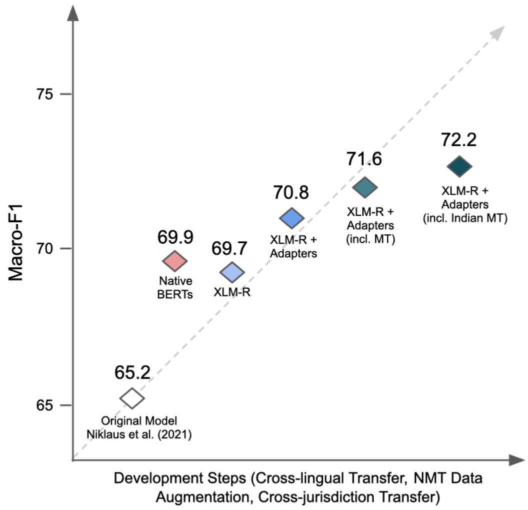


Figure 1: Incremental performance improvement through several development steps.

To the best of our knowledge, CLT has not been applied to the Legal Judgment Prediction (LJP) task (Aletras et al., 2016; Xiao et al., 2018; Chalkidis et al., 2019; Malik et al., 2021), where the goal is to predict the verdict (court decision) given the facts of a legal case. In this setting, positive impact of cross-lingual transfer is not as conceptually straight-forward as in other general applications (NLU), since there are known complications for sharing legal definitions and interpreting law across languages (Gotti, 2014; McAuliffe, 2014; Robertson, 2016; Ramos, 2021).

Following the work of Niklaus et al. (2021), we experiment with their newly released trilingual Swiss-Judgment-Prediction (SJP) dataset, containing cases from the Federal Supreme Court of Switzerland (FSCS), written in three official Swiss languages (German, French, Italian). The dataset covers four legal areas (public, penal, civil, and social law) and lower courts located in eight regions of Switzerland (Zurich, Ticino, etc.), which poses

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 32–46
November 20–23, 2022.

interesting new challenges on model robustness / fairness and the effect of cross-domain and cross-regional knowledge sharing. In their experiments, Niklaus et al. (2021) find that the performance in cases written in Italian is much lower compared to the rest, while also performance varies a lot across regions and legal areas.

Main Research Questions

We pose and examine four main research questions:
RQ1: *Is cross-lingual transfer beneficial across all or some of the languages?*
RQ2: *Do models benefit or not from cross-regional and cross-domain transfer?*
RQ3: *Can we leverage data from another jurisdiction to improve performance?*
RQ4: *How does representational bias (wrt. language, origin region, legal area) affect model's performance?*

Contributions

The contributions of this paper are fourfold:

- We explore, for the first time, the application of cross-lingual transfer learning in the challenging LJP task in several settings (Section 3.3). We find that a pre-trained language model fine-tuned multilingually, outperforms its monolingual counterparts, especially when we use adapter-based fine-tuning and augment the training data with machine-translated versions of the original documents (3× larger training corpus) with larger gains in a low-resource setting (Italian).
- We perform cross-domain and cross-regional analyses (Section 3.4) exploring the effects of cross-domain and cross-regional transfer, i.e., train a model across domains, i.e., legal areas (e.g., civil, penal law), or regions (e.g., Zurich, Ticino). We find that in both settings (legal areas, regions), models trained across all groups perform overall better and more robustly; while always improving performance in the worst-case (region or legal area) scenario.
- We also report improved results when we apply cross-jurisdiction transfer (Section 3.5) , where we further augment our dataset with Indian legal cases originally written in English.
- We release the augmented dataset (incl. 100K machine-translated documents) and our code for replicability and future experimentation.[1]

[1]https://huggingface.co/datasets/swiss_judgment_prediction

The cumulative performance improvement amounts to 7% overall and 16+% in the low-resource Italian subset, compared to the best reported scores in Niklaus et al. (2021), while using cross-lingual and cross-jurisdiction transfer we improve for 2.3% overall and 4.6% for Italian over our strongest baseline (NativeBERTs).

2 Dataset and Task description

2.1 Swiss Legal Judgment Prediction Dataset

We investigate the LJP task on the Swiss-Judgment-Prediction (SJP) dataset (Niklaus et al., 2021). The dataset contains 85K cases from the Federal Supreme Court of Switzerland (FSCS) from the years 2000 to 2020 written in German, French, and Italian. The court hears appeals focusing on small parts of the previous (lower court) decision, where they consider possible wrong reasoning by the lower court. The dataset provides labels for a simplified binary (*approval*, *dismissal*) classification task. Given the facts of the case, the goal is to predict if the plaintiff's request is valid or partially valid (i.e., the court *approved* the complaint).

Since the dataset contains rich metadata, such as legal areas and origin regions, we can conduct experiments on the robustness of the models (see Section 3.4). The dataset is not equally distributed; in fact, there is a notable representation disparity where Italian have far fewer documents (4K), compared to German (50K) and French (31K). Representation disparity is also vibrant with respect to legal areas and regions. We refer readers to the work of Niklaus et al. for detailed dataset statistics.

2.2 Indian Legal Judgment Prediction Dataset

The Indian Legal Documents Corpus (ILDC) dataset (Malik et al., 2021) comprises 30K cases from the Indian Supreme Court in English. The court hears appeals that usually include multiple petitions and rules a decision (*accepted* vs. *rejected*) per petition. Similarly to Niklaus et al. (2021), Malik et al. released a simplified version of the dataset with binarized labels. In effect, the two datasets (SJP, ILDC) target the very same task (partial or full approval of plaintiff's claims), nonetheless in two different jurisdictions (Swiss Federation and India). Our main goal, when we use ILDC as a complement of SJP, is to assess the possibility of cross-jurisdiction transfer from Indian to Swiss cases (see Section 3.5), an experimental scenario that has not been explored so far in the literature.

2.3 NMT-based Data Augmentation

In some of our experiments, we perform data augmentation using machine-translated versions of the original documents, i.e., translate a document originally written in a single language to the other two (e.g., from German to French and Italian). We performed the translations using the EasyNMT[2] framework utilizing the *many-to-many* Neural Machine Translation (NMT) model of Fan et al. (2020).[3] A preliminary manual check of some translated samples showed sufficient translation quality to proceed forward. We release the machine-translated additional dataset for future consideration on cross-lingual experiments or quality assessment.

To the best of our knowledge, machine translation for data augmentation has not been studied in legal Natural Language Processing (NLP) applications, while it is generally a straight-forward, though under-studied idea. As we show in the experiments (see Section 3.3), the translations are effective, leading to an average improvement of 1.6% macro-F1 for standard fine-tuning and 0.8% for adapter-based one (see Table 1). For the low-resource Italian subset, the improvement even amounts to 3.2% and 1.6%, respectively.

3 Experiments

3.1 Hierarchical BERT

Since the examined dataset (SJP) contains many documents with more than 512 tokens (90% of the documents are up to 2048), we use Hierarchical BERT models (Chalkidis et al., 2019; Niklaus et al., 2021; Dai et al., 2022) to encode up to 2048 tokens per document (4×512 blocks).

We split the text into consecutive blocks of 512 tokens and feed the first 4 blocks to a shared standard BERT encoder. Then, we aggregate the block-wise `CLS` tokens by passing them through another 2-layer transformer encoder, followed by max-pooling and a final classification layer.

We re-use and expand the implementation released by Niklaus et al. (2021),[4] which is based on the Hugging Face library (Wolf et al., 2020). Notably, we first improve the masking of the blocks. Specifically, when the document has less than the maximum number (4) of blocks, we pad with extra sequences of `PAD` tokens, without the use of special tokens (`CLS`, `SEP`), as was previously performed. This minor technical improvement seems to affect the model's performance at large (group A1 Prior SotA vs. NativeBERTs — Table 1).

We experiment with monolingually pre-trained BERT models (aka NativeBERTs) and the multilingually pre-trained XLM-R of Conneau et al. (2020). Specifically, for monolingual experiments (Native BERTs), we use German-BERT (Chan et al., 2019) for German, CamemBERT (Martin et al., 2020) for French, and UmBERTo (Parisi et al., 2020) for Italian, similar to Niklaus et al. (2021).

In our multilingual experiments, we also assess the effectiveness of adapter-based fine-tuning (Houlsby et al., 2019; Pfeiffer et al., 2020), in comparison to standard full fine-tuning. In this setting, adapter layers are placed after all feed-forward layers of XLM-R and are trained together with the parameters of the layer-normalization layers. The rest of the model parameters remain untouched.

3.2 Experimental Set Up

We follow Niklaus et al. (2021) and report macro-averaged F1 score to account for the high class-imbalance in the dataset (approx. 20/80 approval/dismissal ratio). We repeat each experiment with 3 different random seeds and report the average score and standard deviation across runs (seeds). We perform grid-search for the learning rate and report test results, selecting the hyper-parameters with the best development scores.[5]

3.3 Cross-lingual Transfer

We first examine *cross-lingual transfer*, where the goal is to share (transfer) knowledge across languages, and we compare models in three main settings: (a) *Monolingual* (see Section 3.3.1): fine-tuned per language, using either the documents originally written in the language, or an augmented training set including the machine-translated versions of all other documents (originally written in another language), (b) *Cross-lingual* (see Section 3.3.2): fine-tuned across languages with or without the additional translated versions, and (c) *Zero-shot cross-lingual* (see Section 3.3.3): fine-tuned across a subset of the languages excluding the target language at a time. We present the results in Table 1.

[2] `https://github.com/UKPLab/EasyNMT`

[3] The *one-to-one* OPUS-MT (Tiedemann and Thottingal, 2020) models did not have any model available from French to Italian (fr2it) at the time of the experiments.

[4] `https://github.com/JoelNiklaus/SwissJudgementPrediction`

[5] Additional details on model configuration, training, and hyper-parameter tuning can be found in Appendix A.

Model	#D	#M	German ↑	French ↑	Italian ↑	All ↑	(Diff. ↓)
A1. *Monolingual: Fine-tune on the* **tgt training set** (**src** = **tgt**) — Baselines							
Prior SotA (Niklaus et al.)	3-35K	N	68.5 ± 1.6	70.2 ± 1.1	57.1 ± 0.4	65.2 ± 0.8	(13.1)
NativeBERTs	3-35K	N	69.6 ± 0.4	72.0 ± 0.5	68.2 ± 1.3	69.9 ± 1.6	(3.8)
XLM-R	3-35K	N	68.2 ± 0.3	69.9 ± 1.6	65.9 ± 1.2	68.0 ± 2.0	(4.0)
A2. *Monolingual: Fine-tune on the* **tgt training set incl. machine-translations** (**src** = **tgt**)							
NativeBERTs	60K	N	70.0 ± 0.7	71.0 ± 1.3	71.9 ± 2.5	71.0 ± 0.8	(0.9)
XLM-R	60K	N	68.8 ± 1.4	70.7 ± 2.1	71.9 ± 2.6	70.4 ± 1.3	(1.1)
B1. *Cross-lingual: Fine-tune on* **all training sets** (**src** ⊂ **tgt**)							
XLM-R	60K	1	68.9 ± 0.3	71.1 ± 0.3	68.9 ± 1.4	69.7 ± 1.0	(2.2)
XLM-R + Adapters	60K	1	69.9 ± 0.6	71.8 ± 0.7	70.7 ± 1.8	70.8 ± 0.8	(0.9)
B2. *Cross-lingual: Fine-tune on* **all training sets incl. machine-translations** (**src** ⊂ **tgt**)							
XLM-R	180K	1	70.2 ± 0.5	71.5 ± 1.1	72.1 ± 1.2	71.3 ± 0.7	(1.9)
XLM-R + Adapters	180K	1	**70.3** ± 0.9	**72.1** ± 0.8	**72.3** ± 2.1	**71.6** ± 0.8	(2.0)
C. *Zero-shot Cross-lingual: Fine-tune on* **all training sets excl. tgt language** (**src** ≠ **tgt**)							
XLM-R	25-57K	1	58.4 ± 1.2	58.7 ± 0.8	68.1 ± 0.2	61.7 ± 4.5	(9.7)
XLM-R + Adapters	25-57K	1	62.5 ± 0.6	58.8 ± 1.5	67.5 ± 2.2	62.8 ± 3.7	(8.7)

Table 1: Test results for all training set-ups (monolingual w/ or w/o translations, multilingual w/ or w/o translations, and zero-shot) w.r.t source (src) and target (tgt) language. Best overall results are in **bold**, and best per setting (group) are underlined. #D is the number of training documents used. #M is the number of models trained/used. The mean and standard deviation are computed across random seeds and across languages for the last column. Diff. shows the difference between the best and the worst performing language. ***The adapter-based multilingually fine-tuned XLM-R model including machine-translated versions (3× larger corpus) has the best overall results***.

3.3.1 Mono-Lingual Training

We observe that the baseline of *monolingually* pre-trained and fine-tuned models (NativeBERTs) have the best results compared to the *multilingually* pre-trained but *monolingually* fine-tuned XLM-R (group A1 – Table 1). Representational bias across languages (Section 2.1) seems to be a key part of performance disparity, considering the performance of the least represented language (Italian) compared to the rest (3K vs. 21-35K training documents). However, this is not generally applicable, i.e., French have better performance compared to German, despite having approx. 30% less training documents.

Translating the full training set provides a 3× larger training set (approx. 180K in total) that "equally" represents all three languages.[6] Augmenting the original training sets with translated versions of the documents (group A2 – Table 1), originally written in another language, improves performance in almost all (5/6) cases (languages per model). Interestingly, the performance improvement in Italian, which has the least documents (less than 1/10 compared to German), is the largest across languages with 3.7% for NativeBERT (68.2 to 71.9) and 6% for XLM-R (65.9 to 71.9) making Italian the best performing language after augmentation. Data augmentation seems more beneficial for XLM-R, which does not equally represent the three examined languages.[7]

3.3.2 Cross-Lingual Training

We now turn to the *cross-lingual transfer* setting, where we train XLM-R across all languages in parallel. We observe that cross-lingual transfer (group B1 – Table 1) improves performance (+4.5% p.p.) across languages compared to the same model (XLM-R) fine-tuned in a monolingual setting (group A1 – Table 1). This finding suggests that cross-lingual transfer (and the inherited benefit of using larger multilingual corpora) has a signifi-

[6] Representational equality with respect to number of training documents per language, but possibly not considering text quality, since we use NMT to achieve that goal.

[7] Refer to Conneau et al. (2020) for resources per language used to pre-train XLM-R (50% less tokens for Italian).

Origin Region	#D	#L	ZH	ES	CS	NWS	EM	RL	TI	FED	All
Region-specific fine-tuning with MT data augmentation											
Zürich (ZH)	26.4K	de	65.5	65.6	63.7	68.2	62.0	57.9	63.2	54.8	62.6
Eastern Switzerland (ES)	17.1K	de	62.9	66.9	62.8	65.2	62.2	60.2	57.8	55.1	61.6
Central Switzerland (CS)	14.4K	de	62.5	65.5	63.2	65.1	60.7	57.8	60.5	55.9	61.4
Northwestern Switzerland (NWS)	17.1K	de	66.0	68.6	65.2	67.9	61.6	57.0	57.1	55.5	62.4
Espace Mittelland (EM)	24.9K	de,fr	64.1	66.6	63.3	66.7	64.0	66.8	63.2	58.4	64.1
Région Lémanique (RL)	40.2K	fr,de	61.0	64.7	60.2	63.7	63.4	69.8	67.6	54.3	63.1
Ticino (TI)	6.9K	it	55.0	56.3	53.2	54.5	56.0	54.7	66.0	53.1	56.1
Federation (FED)	3.9K	de,fr,it	57.5	59.6	56.8	58.9	55.0	56.5	53.5	54.9	56.6
Cross-regional fine-tuning w/o MT data augmentation											
XLM-R	60K	de,fr,it	68.5	71.3	67.7	71.2	69.0	71.4	67.4	64.6	68.9
XLM-R + Adapters	60K	de,fr,it	**69.2**	**73.9**	67.9	72.6	69.0	**72.1**	70.1	64.2	69.9
Cross-regional fine-tuning with MT data augmentation											
NativeBERTs	180K	de,fr,it	69.0	72.1	68.6	72.0	69.9	71.9	68.8	64.8	69.6
XLM-R	180K	de,fr,it	**69.2**	72.9	68.3	**73.3**	69.9	71.7	70.4	**65.0**	70.1
XLM-R + Adapters	180K	de,fr,it	**69.2**	73.3	**69.9**	73.0	**70.3**	**72.1**	**70.9**	63.8	**70.3**

Table 2: Test results for models trained per region or across all regions. Best overall results are in **bold**, and in-domain are underlined. #D is the total number of training examples. #L are the languages covered. ***Cross-regional transfer is beneficial for all regions and has the best overall results. The shared multilingual model trained across all languages and regions slightly outperforms the baseline (NativeBERTs).***

cant impact, despite the legal complication of sharing legal definitions across languages. Augmenting the original training sets with the documents translated across all languages, further improves performance (group B2 – Table 1).

3.3.3 Zero-Shot Cross-Lingual Training

We also present results in a *zero-shot cross-lingual* setting (group C – Table 1), where XLM-R is trained in two languages and evaluated in the third one (unseen in fine-tuning). We observe that German has the worst performance (approx. 10% drop), which can be justified as German is a *Germanic* language, while both French and Italian are *Romance* and share a larger part of the vocabulary.

Contrarily, in case of Italian, the low-resource language in our experiments, the model strongly benefits from zero-shot cross-lingual transfer, leading to 2.2% p.p. improvement, compared to the monolingually trained XLM-R. In other words, training XLM-R with much more (approx 20×) out-of-language (57K in German and French) data is better compared to training on the limited (3K) in-language (Italian) documents (68.1 vs. 65.9).

3.3.4 Fine-tuning with Adapters

Across all cross-lingual settings (groups B-C – Table 1), the use of Adapters improves substantially the overall performance. The multilingual adapter-based XLM-R in group B1 (Table 1) has comparable performance to the NativeBERTs models of group A2, where the training dataset has been artificially augmented with machine translations. In a similar setting (group B2 – Table 1), the multilingual adapter-based XLM-R in group B2 has the best overall results, combining the benefits of both cross-lingual transfer and data augmentation.

With respect to *cross-lingual performance parity*, the adapter-based XLM-R model has also the highest performance parity (least diff. in the last column of Table 1), while augmenting the dataset with NMT translations leads to both the worst-case (language) performance and best performance for the least represented language (Italian).

In conclusion, cross-lingual transfer with an augmented dataset comprised of the original and machine-translated versions of all documents, has the best overall performance with a vibrant improvement (3% compared to our strong baselines – second part of Group A1 in Table 1) in Italian, the least represented language.

3.4 Cross-Domain/Regional Transfer Analysis

Further on, we examine the benefits of transfer learning (knowledge sharing) in other dimensions. Hence, we analyze model performance with respect to origin regions and legal areas (domains of law).

Legal Area	#D	Public Law	Civil Law	Penal Law	Social Law	All
Domain-specific fine-tuning with MT data augmentation						
Public Law	45.6K	<u>56.4</u> ± 2.2	52.2 ± 2.0	59.7 ± 4.9	60.1 ± 5.8	57.1 ± 3.2
Civil Law	34.5K	44.4 ± 7.9	<u>64.2</u> ± 0.6	45.5 ± 13.1	43.6 ± 5.2	49.4 ± 8.6
Penal Law	35.4K	40.8 ± 10.1	55.8 ± 2.9	<u>**84.5**</u> ± 1.3	61.1 ± 7.5	60.6 ± 15.7
Social Law	29.1K	52.6 ± 4.2	56.6 ± 2.0	69.0 ± 5.5	<u>70.2</u> ± 2.0	62.1 ± 7.6
Cross-domain fine-tuning w/o MT data augmentation						
XLM-R	60K	57.4 ± 2.0	66.1 ± 3.1	81.4 ± 1.4	70.8 ± 2.0	68.9 ± 8.7
XLM-R + Adapters	60K	58.4 ± 2.5	66.1 ± 2.4	83.1 ± 1.2	71.1 ± 1.4	69.7 ± 9.0
Cross-domain fine-tuning with MT data augmentation						
NativeBERTs	180K	58.1 ± 3.0	64.5 ± 3.7	83.0 ± 1.3	71.1 ± 4.3	69.2 ± 9.2
XLM-R	180K	58.0 ± 3.0	**67.2** ± 1.6	84.4 ± 0.2	70.2 ± 1.3	**70.0** ± 9.5
XLM-R + Adapters	180K	**58.6** ± 2.7	66.8 ± 2.8	83.1 ± 1.3	**71.3** ± 2.4	69.9 ± 8.8

Table 3: Test results for models (XLM-R with MT unless otherwise specified) **fine-tuned** per legal area (domain) or across all legal areas (domains). Best overall results are in **bold**, and in-domain are <u>underlined</u>. The mean and standard deviations are computed across languages per legal area and across legal areas for the right-most column. #D is the total number of training examples. ***Cross-domain transfer is beneficial for 3 out of 4 legal areas and has the best overall results.*** The shared multilingual model trained across all languages and legal areas outperforms the baseline (monolingual BERT models).

3.4.1 Origin Regions

In Table 2 we present the results for *cross-regional* transfer. In the top section of the table, we present results with region-specific multilingual (XLM-R) models evaluated across regions (in-region on the diagonal, zero-shot otherwise). We observe that the cross-regional models (two lower groups of Table 2) always outperform the region-specific models. Moreover, cross-lingual transfer is beneficial across cases, while adapter-based fine-tuning further improves results in 5 out of 8 cases (regions). Data augmentation is also beneficial in most cases.

In the top part of Table 2, in 60% of the cases (regions: ZH, ES, CS, NWS, TI), a “zero-shot” model, i.e., trained in the cases of another region, slightly outperforms the in-region model. In other words, in almost every case (target region), there is another *monolingual* region-specific model that outperforms the in-region one.

We consider two main factors that may explain these results: (a) the region-wise *representational bias* considering the number of cases per region, and (b) the cross-regional *topical similarity* of the training and test subsets across different regions. To approximate the cross-regional topical similarity, we consider the distributional similarity (or dissimilarity) w.r.t. legal areas (Table 6 in Appendix C). None of these factors can fully explain the results. Although in 3 out of 5 cases, the best performing (out-of-region) model has been trained on more data compared to the in-region one. There are also other confounding factors (e.g., language), i.e., models trained on the cases of either Espace Mittelland (EM) or Région Lémanique (RL), both bilingual with 8-10K cases, have the best results across all single-region models, hence a further exploration of the overall dynamics is needed.

3.4.2 Legal Areas

In Table 3 we present the results for *cross-domain* transfer between legal areas (domains of law). The results on the diagonal (<u>underlined</u>) are in-domain, i.e., fine-tuned and evaluated in the same legal area. We observe that for each domain, the models trained on in-domain data have the best results in the respective domain compared to the rest.

Interesting to note is that the best results (**bold**) are achieved in the cross-domain setting in 3 out of 4 legal areas. Such an outcome is not anticipated based on the current trends in law industry, where legal experts (judges, lawyers) over-specialize and excel in specific legal areas, e.g., criminal defense lawyers. Penal law poses the only exception where the domain-specific model is on par with the cross-domain model. Again, the results per area do not correlate with the volume of training data (*cross-*

Model	Training Dataset	#D	German ↑	French ↑	Italian ↑	All	(Diff. ↓)
			Cross-lingual fine-tuning w/ or w/o MT data augmentation				
XLM-R	Original	60K	68.9 ± 0.3	71.1 ± 0.3	68.9 ± 1.4	69.7 ± 1.0	(2.2)
XLM-R + Adapters	Original	60K	<u>69.9</u> ± 0.6	<u>71.8</u> ± 0.7	<u>70.7</u> ± 1.8	<u>70.8</u> ± 0.8	(0.9)
XLM-R	+ MT Swiss	180K	70.2 ± 0.5	71.5 ± 1.1	<u>72.1</u> ± 1.2	71.3 ± 0.7	(1.9)
XLM-R + Adapters	+ MT Swiss	180K	<u>70.3</u> ± 0.8	<u>72.1</u> ± 0.8	<u>72.1</u> ± 1.2	<u>71.5</u> ± 0.9	(1.8)
			Cross-jurisdiction fine-tuning w/ MT data augmentation				
XLM-R	+ MT {Swiss, Indian}	276K	70.5 ± 0.4	71.8 ± 0.3	<u>**73.5**</u> ± 1.4	72.0 ± 0.9	(3.0)
XLM-R + Adapters	+ MT {Swiss, Indian}	276K	<u>**71.0**</u> ± 0.4	<u>**73.0**</u> ± 0.6	72.6 ± 1.1	**72.2** ± 1.2	(2.0)
			Cross-jurisdiction zero-shot fine-tuning w/ MT data augmentation				
XLM-R	MT Indian	96K	50.4 ± 1.5	47.9 ± 1.0	49.5 ± 1.3	49.3 ± 1.0	(2.5)
XLM-R + Adapters	MT Indian	96K	<u>51.6</u> ± 2.9	<u>49.7</u> ± 1.4	<u>50.1</u> ± 1.4	<u>50.5</u> ± 1.0	(1.9)

Table 4: Test results for cross-jurisdiction transfer. We present results in four settings: *standard* (Original) *augmented* (+ MT Swiss), *further augmented incl. cross-jurisdiction* (+ MT Swiss + MT Indian) and *zero-shot* (MT Indian). Best results are in **bold**. Diff. shows the difference between the best performing language and the worst performing language (max - min). ***Further augmenting with translated Indian cases is overall beneficial.***

domain representational bias), and suggest that other qualitative characteristics (e.g., the idiosyncrasies of criminal law) affect the task complexity.

Similarly to the cross-regional experiments, the shared multilingual model (XLM-R) trained across all languages and legal areas with an augmented dataset outperforms the NativeBERTs models trained in a similar setting, giving another indication that the performance gains from cross-lingual transfer and data augmentation via machine translation are robust across domains as well.

3.5 Cross-Jurisdiction Transfer

We, finally, "ambitiously" stretch the limits of transfer learning in LJP and we apply *cross-jurisdiction* transfer, i.e., use of cases from different legal systems, another form of cross-domain transfer. For this purpose, we further augment the SJP dataset of FSCS cases, with cases from the Supreme Court of India (SCI), published by Malik et al. (2021).[8] We consider and translate all (approx. 30K) Indian cases ruled up to the last year (2014) of our training dataset, originally written in English, to all target languages (German, French, and Italian).[9]

In Table 4, we present the results for two cross-jurisdiction settings: *zero-shot* (Only MT Indian), where we train XLM-R on the machine-translated version of Indian cases, and *further augmented* (Original + MT Swiss + MT Indian), where we further augment the (already augmented) training set of Swiss cases with the translated Indian ones. While zero-shot transfer clearly fails; interestingly, we observe improvement for all languages in the further augmented setting. This opens a fascinating new direction for LJP research.

Similar to our results in Section 3.3 with respect to cross-lingual performance parity, the standard adapter-based XLM-R model has also the highest performance parity (least diff. on Table 4), while the same model trained on the fully augmented dataset leads to the worst-case (language; German) performance and best performance for the least represented language (Italian).

The cumulative improvement from all applied enhancements adds up to 7% macro-F1 compared to the XLM-R baseline and 16% to the best method by Niklaus et al. (2021) in the low-resource Italian subset, while using cross-lingual and cross-jurisdiction transfer we improve for 2.3% overall and 4.6% for Italian over our strongest baseline (NativeBERTs).

Since our experiments present several incremental improvements, we assess the stability of the performance improvements with statistical significance testing by comparing the most crucial settings in Appendix B.

4 Related Work

Legal Judgment Prediction (LJP) is the task, where given the facts of a legal case, a system

[8] Although the SCI rules under the Indian jurisdiction (law), while the FSCS under the Swiss one, we hypothesize that the fundamentals of law in two modern legal systems are quite common and thus transferring knowledge could potentially have a positive effect. We discuss this matter in Section 5.

[9] We do not use the original documents written in English, as English is not one of our target languages.

has to predict the correct outcome (legal judgement). Many prior works experimented with some forms of LJP, however, the precise formulation of the LJP task is non-standard as the jurisdictions and legal frameworks vary. Aletras et al. (2016); Medvedeva et al. (2018); Chalkidis et al. (2019) predict the plausible violation of European Convention of Human Rights (ECHR) articles of the European Court of Human Rights (ECtHR). Xiao et al. (2018, 2021) study Chinese criminal cases where the goal is to predict the ruled duration of prison sentences and/or the relevant law articles.

Another setup is followed by Şulea et al. (2017); Malik et al. (2021); Niklaus et al. (2021), which use cases from Supreme Courts (French, Indian, Swiss, respectively), hearing appeals from lower courts relevant to several fields of law (legal areas). Across tasks (datasets), the goal is to predict the binary verdict of the court (approval or dismissal of the examined appeal) given a textual description of the case. None of these works have explored neither cross-lingual nor cross-jurisdiction transfer, while the effects of cross-domain and cross-regional transfer are also not studied.

Cross-Lingual Transfer (CLT) is a flourishing topic with the application of pre-trained transformer-based models trained in a multilingual setting (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020; Xue et al., 2021) excelling in NLU benchmarks (Ruder et al., 2021). Adapter-based fine-tuning (Houlsby et al., 2019; Pfeiffer et al., 2021) has been proposed as an anti-measure to mitigate misalignment of multilingual knowledge when CLT is applied, especially in a zero-shot fashion, where the target language is unseen during training (or even pre-training).

Meanwhile, CLT is understudied in legal NLP applications. Chalkidis et al. (2021) experiment with standard fine-tuning, while they also examined the use of adapters (Houlsby et al., 2019) for zero-shot CLT on a legal topic classification dataset comprising European Union (EU) laws. They found adapters to achieve the best tradeoff between effectiveness and efficiency. Their work did not examine the use of methods incorporating translated versions of the original documents in any form, i.e., translate train documents or test ones. Recently, Xenouleas et al. (2022) used an updated, unparalleled version of Chalkidis et al. dataset to study NMT -augmented CLT methods. Other multilingual legal NLP resources (Galassi et al., 2020; Drawzeski et al., 2021) have been recently released, although CLT is not applied in any form.

5 Motivation and Challenges for Cross-Jurisdiction Transfer

Legal systems vary from country to country. Although they develop in different ways, legal systems also have some similarities based on historically accepted justice ideals, i.e., the rule of law and human rights. Switzerland has a civil law legal system (Walther, 2001), i.e., statutes (legislation) is the primary source of law, at the crossroads between Germanic and French legal traditions.

Contrary, India has a hybrid legal system with a mixture of civil, common law, i.e., judicial decisions have precedential value, and customary, i.e., Islamic ethics, or religious law (Bhan and Rohatgi, 2021). The legal and judicial system derives largely from the British common law system, coming as a consequence of the British colonial era (1858-1947) (Singh and Kumar, 2019).

Based on the aforementioned, cross-jurisdiction transfer is challenging since the data (judgments) abide to different law standards. Although the Supreme Court of India (SCI) rules under the Indian jurisdiction (law), while the Federal Supreme Court of Switzerland (FSCS) under the Swiss one, we hypothesize that the fundamentals of law in two modern legal systems are quite common and thus transferring knowledge could potentially have a positive effect, and thus it is an experiment worth considering, while we acknowledge that from a legal perspective equating legal systems is deeply problematic, since the legislation, the case law, and legal practice are different.

Our empirical work and experimental results shows that cross-jurisdiction transfer in this specific setting (combination of Swiss and Indian decisions) has a positive impact in performance, but we cannot provide any profound hypothesis neither we are able to derive any conclusions on the importance of this finding on legal literature and practice. We leave these questions in the hands of those who can responsibly bear the burden, the legal scholars.

6 Conclusions and Future Work

6.1 Answers to the Research Questions

Following the experimental results (Section 3), we answer the original predefined research questions:

RQ1: *Is cross-lingual transfer beneficial across all or some of the languages?* In Section 3.3, we find that vanilla CLT is beneficial in a low-resource setting (Italian), with comparable results in the rest of the languages. Moreover, CLT leveraging NMT -based data augmentation is beneficial across all languages. Overall, our experiments lead to a single multi-lingual cross-lingually "fairer" model.

RQ2: *Do models benefit or not from cross-regional and cross-domain transfer?* In Section 3.4, we find that models benefit from cross-regional transfer across all cases, since they are exposed to (trained in) many more documents (cases). We believe cross-regional diversity is not a significant aspect, compared to the importance of the increased data volume and language diversity. Cross-domain transfer is beneficial in three out of four cases (legal areas), with comparable results on penal (criminal) law, where the application of law seems to be more straight-forward / standardized (higher performing legal area). Cross-regional and cross-domain transfer lead to more robust models.

RQ3: *Can we leverage data from another jurisdiction to improve performance?* In Section 3.5, we find that cross-jurisdiction transfer in our specific setup, i.e., very similar LJP tasks, is beneficial. Again, we believe that this is mostly a matter of additional unique data (cases), rather than a matter of jurisdictional similarity. Cross-jurisdiction transfer leads to a better performing model.

RQ4: *How does representational bias (wrt. language, origin region, legal area) affect model's performance?* We observe that representational bias – in non-extreme cases (e.g., w.r.t. language) – does not always explain performance disparities across languages, regions, or domains, and other characteristics also need to be considered.

6.2 Conclusions - Summary

We examined the application of Cross-Lingual Transfer (CLT) in Legal Judgment Prediction (LJP) for the very first time, finding a multilingually trained model to be superior when augmenting the dataset with NMT. Adapter-based fine-tuning leads to even better results. We also examined the effects of cross-domain (legal areas) and cross-regional transfer, which is overall beneficial in both settings, leading to more robust models. Cross-jurisdiction transfer by augmenting the training set with machine-translated Indian cases further improves performance.

6.3 Future Work

In future work, we would like to explore the use of a legal-oriented multilingual pre-trained model by either continued pre-training of XLM-R, or pre-training from scratch in multilingual legal corpora. Legal NLP literature (Chalkidis et al., 2022; Zheng et al., 2021) suggests that domain-specific language models positively affect performance.

In another interesting direction, we will consider other data augmentation techniques (Feng et al., 2021; Ma, 2019) that rely on textual alternations (e.g., paraphrasing, etc.). We would also like to further investigate cross-jurisdictional transfer, either exploiting data for similar LJP tasks, or via multi-task learning on multiple LJP datasets with dissimilar task specifications.

7 Ethics Statement

The scope of this work is to study LJP to broaden the discussion and help practitioners to build assisting technology for legal professionals and laypersons. We believe that this is an important application field, where research should be conducted (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize law, while also highlight (inform the audience on) the various multi-aspect shortcomings seeking a responsible and ethical (fair) deployment of legal-oriented technologies.

In this direction, we study how we could better exploit all the available resources (from various languages, domains, regions, or even different jurisdictions). This combination leads to models that improve overall performance – more robust models –, while having improved performance in the worst-case scenarios across many important demographic or legal dimensions (low-resource language, worst performing legal area and region).

Nonetheless, irresponsible use (deployment) of such technology is a plausible risk, as in any other application (e.g., online content moderation) and domain (e.g., medical). We believe that similar technologies should only be deployed to assist human experts (e.g., legal scholars in research, or legal professionals in forecasting or assessing legal case complexity) with notices on their limitations.

The main examined dataset, Swiss-Judgment-Prediction (SJP), released by Niklaus et al. (2021), comprises publicly available cases from the FSCS, where cases are pre-anonymized, i.e., names and other sensitive information are redacted. The same applies for the second one, Indian Legal Documents Corpus (ILDC) of Malik et al. (2021).

Acknowledgements

This work has been supported by the Swiss National Research Program "Digital Transformation" (NRP-77)[10] grant number 187477. This work is also partly funded by the Innovation Fund Denmark (IFD)[11] under File No. 0175-00011A. This research has been also co-financed by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (T2EΔK-03849).

We would like to thank Thomas Lüthi for his legal advice, Mara Häusler for great discussions regarding the evaluation process of the models, and Phillip Rust and Desmond Elliott for providing valuable feedback on the original draft of the manuscript.

References

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93. Publisher: PeerJ Inc.

Ashish Bhan and Mohit Rohatgi. 2021. Legal systems in India: Overview. *Thomsons Reuters - Practical Law*.

Rich Caruana, Steve Lawrence, and C. Giles. 2001. Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping. In *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. 2019. deepset - Open Sourcing German BERT.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*. ArXiv: 1911.02116.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A corpus for multilingual analysis of online terms of service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 1–8, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence, Italy. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. *CoRR*, abs/2010.11125.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Andrea Galassi, Kasper Drazewski, Marco Lippi, and Paolo Torroni. 2020. Cross-lingual annotation projection in legal texts. In *Proceedings of the 28th International Conference on Computational Linguistics*,

[10] https://www.nfp77.ch/en/

[11] https://innovationsfonden.dk/en

pages 915–926, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maurizio Gotti. 2014. Linguistic Features of Legal Texts: Translation Issues. *Statute Law Review*, 37(2):144–155.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Karen McAuliffe. 2014. Translating Ambiguity. *Journal of Comparative Law*, 9(2).

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the European Court of Human Rights: Looking into the crystal ball. In *Proceedings of the Conference on Empirical Legal Studies*, page 24.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2020. On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. UmBERTo: an Italian Language Model trained with Whole Word Masking. Original-date: 2020-01-10T09:55:31Z.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Fernando Prieto Ramos. 2021. Translating legal terminology and phraseology: between inter-systemic incongruity and multilingual harmonization. *Perspectives*, 29(2):175–183.

C.D. Robertson. 2016. *Multilingual Law: A Framework for Analysis and Understanding*. Law, language and communication. Routledge.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mahendra Pal Singh and Niraj Kumar. 2019. 1Tracing the History of the Legal System in India. In *The Indian Legal System: An Enquiry*. Oxford University Press.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Dennis Ulmer. 2021. deep-significance: Easy and Better Significance Testing for Deep Neural Networks. Https://github.com/Kaleidophon/deep-significance.

Fridolin M.R. Walther. 2001. The swiss legal system a guide for foreign researchers. *International Journal of Legal Information*, 29(1):1–24.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Stratos Xenouleas, Alexia Tsoukara, Giannis Panagiotakis, Ilias Chalkidis, and Ion Androutsopoulos. 2022. Realistic zero-shot cross-lingual transfer in legal topic classification. In *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, SETN '22, New York, NY, USA. Association for Computing Machinery.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *CoRR*, abs/2105.03887.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. 2018. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv:1807.02478 [cs]*. ArXiv: 1807.02478.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

A Hyperparameter Tuning

We experimented with learning rates in {1e-5, 2e-5, 3e-5, 4e-5, 5e-5} as suggested by Devlin et al. (2019). However, like reported by Mosbach et al. (2020), we also found RoBERTa-based models to exhibit large training instability with learning rate 3e-5, although this learning rate worked well for BERT-based models. 1e-5 worked well enough for all models. To avoid either over- or under-fitting, we use Early Stopping (Caruana et al., 2001) on development data. To combat the high class imbalance, we use oversampling, following (Niklaus et al., 2021).

We opted to use the standard Adapters of Houlsby et al. (2019), as the language Adapters introduced by Pfeiffer et al. (2020) are more resource-intensive and require further pre-training per language. We tuned the adapter reduction factor in {2×, 4×, 8×, 16×} and got the best results with 2× and 4×; we chose 4× for the final experiments to favor less additional parameters. We tuned the learning rate in {1e-5, 5e-5, 1e-4, 5e-4, 1e-3} and achieved the best results with 5e-5.

We additionally applied label smoothing (Szegedy et al., 2015) on cross-entropy loss. We achieved the best results with a label smoothing factor of 0.1 after tuning with {0, 0.1, 0.2, 0.3}.

Model Type	M1	M2	M3	M4
M1: NativeBERTs	1.0	1.0	1.0	1.0
M2: NativeBERTs + MT CH	0.0	1.0	1.0	1.0
M3: XLM-R + MT CH	0.0	0.0	1.0	1.0
M4: XLM-R + MT CH + IN	0.0	0.0	0.0	1.0

Table 5: Almost stochastic dominance ($\epsilon_{\min} < 0.5$) with ASO. *+ MT CH* stands for augmentation with machine translation inside the Swiss dataset and *+ MT CH+IN* is the code for augmentation with machine-translations with the Swiss **and** Indian dataset.

B Statistical Significance Testing

Since our experiments present several incremental improvements, we assessed the stability of the performance improvements with statistical significance testing by comparing the most crucial settings. Using Almost Stochastic Order (ASO) (Dror et al., 2019) with a confidence level $\alpha = 0.05$, we find the score distributions of the core models (NativeBERTs, w/ and w/o MT Swiss, XLM-R w/ and w/o MT Indian and/or Swiss) stochastically dominant ($\epsilon_{\min} = 0$) over each other in order. We compared all pairs of models based on three random seeds each using ASO with a confidence level of $\alpha = 0.05$ (before adjusting for all pair-wise comparisons using the Bonferroni correction). Almost stochastic dominance ($\epsilon_{\min} < 0.5$) is indicated in Table 5 in Appendix A. We use the deep-significance Python library of Ulmer (2021).

C Distances Between Legal Area Distributions per Origin Regions

	ZH	ES	CS	NWS	EM	RL	TI	FED
ZH	<u>.02</u>	.02	.03	.02	.01	.02	.05	.12
ES	.03	<u>.03</u>	.04	.03	.02	.01	.06	.11
CS	.02	.01	<u>.01</u>	.02	.01	.04	.06	.13
NWS	.05	.04	.06	<u>.04</u>	.04	.03	.04	.09
EM	.03	.03	.04	.02	<u>.03</u>	.03	.04	.10
RL	.06	.05	.07	.05	.05	<u>.05</u>	.04	.07
TI	.07	.07	.08	.05	.07	.08	<u>.02</u>	.06
FED	.10	.10	.12	.09	.10	.10	.06	<u>.02</u>

Table 6: Wasserstein distances between the legal area distributions of the training and the test set per origin region across languages. The training sets are in the columns and the test sets in the rows.

In Table 6 we show the Wasserstein distances between the legal area distributions of the training and the test sets per origin region across languages. Unfortunately, this analysis does not explain why the NWS model (zero-shot) outperforms the ZH model (in-domain) on the ZH test set, as found in Table 2.

D Additional Results

In Tables 7, 8, 9 and 10 we present detailed results for all experiments. All tables include both the average score across repetitions, as reported in the original tables in the main article, but also the standard deviations across repetitions.

E Responsible NLP Research

We include information on limitations, licensing of resources, and computing foot-print, as suggested by the newly introduced Responsible NLP Research checklist.

E.1 Limitations

In this appendix, we discuss core limitations that we identify in our work and should be considered in future work.

Data size fluctuations We did not control for the sizes of the training datasets, which is why we reported them in the Tables 2, 3 and 4. This mimics a more realistic setting, where the training set size differs based on data availability. Although we discussed representational bias in RQ4, we cannot completely rule out different performance based on simply more training data.

Mismatch in in/out of region model performance As described in Section 3.4.1, certain zero-shot evaluations outperform in-domain evaluations. Although we try to find an explanation for this in Section 3.4, and Appendix C, it remains an open question since there are many confounding factors.

Re-use of Indian cases Although we have empirical results confirming the statistically significant positive effect of training with additional translated Indian cases, we do not have a profound legal justification or even a hypothesis for this finding at the moment.

E.2 Licensing

The SJP dataset (Niklaus et al., 2021) we mainly use in this work is available under a CC-BY-4 license. The second dataset, ILDC (Malik et al., 2021), comprising Indian cases is available upon request. The authors kindly provided their dataset. All used software and libraries (EasyNMT, Hugging Face Transformers, deep-significance, and several other typical scientific Python libraries) are publicly available and free to use, while we always cite the original work and creators. The artifacts (i.e., the translations and the code) we created, target academic research and are available under a CC-BY-4 license.

E.3 Computing Infrastructure

We used an NVIDIA GeForce RTX 3090 GPU with 24 GB memory for our experiments. In total, the experiments took approx. 80 GPU days, excluding the translations. The translations took approx. 7 GPU days per language from Indian to German, French, and Italian. The translation within the Swiss corpus took approx. 4 GPU days in total.

Legal Area	#D	Public Law	Civil Law	Penal Law	Social Law	All
Public Law	45.6K	56.4 ± 2.2	52.2 ± 2.0	59.7 ± 4.9	60.1 ± 5.8	57.1 ± 3.2
Civil Law	34.5K	44.4 ± 7.9	64.2 ± 0.6	45.5 ± 13.1	43.6 ± 5.2	49.4 ± 8.6
Penal Law	35.4K	40.8 ± 10.1	55.8 ± 2.9	**84.5** ± 1.3	61.1 ± 7.5	60.6 ± 15.7
Social Law	29.1K	52.6 ± 4.2	56.6 ± 2.0	69.0 ± 5.5	70.2 ± 2.0	62.1 ± 7.6
All	60K	58.0 ± 3.0	**67.2** ± 1.6	84.4 ± 0.2	70.2 ± 1.3	**70.0** ± 9.5
All (w/o MT)	60K	57.4 ± 2.0	66.1 ± 3.1	81.4 ± 1.4	70.8 ± 2.0	68.9 ± 8.7
All (Native)	60K	**58.1 ± 3.0**	64.5 ± 3.7	83.0 ± 1.3	**71.1 ± 4.3**	69.2 ± 9.2

Table 7: Test results for models (XLM-R with MT unless otherwise specified) **fine-tuned** per legal area (domain) or across all legal areas (domains). Best overall results are in **bold**, and in-domain are underlined. ***Cross-domain transfer is beneficial for 3 out of 4 legal areas and has the best overall results.*** The shared multilingual model trained across all languages and legal areas outperforms the baseline (monolingual BERT models). The mean and standard deviations are computed across languages per legal area and across legal areas for the right-most column. #D is the number of training examples per legal area.

Legal Area	#D	Public Law	Civil Law	Penal Law	Social Law	All
Public Law	45.6K	57.2 ± 1.8	53.8 ± 2.1	58.9 ± 5.2	61.7 ± 4.1	57.9 ± 2.9
Civil Law	34.5K	41.4 ± 6.6	57.6 ± 1.1	42.8 ± 9.1	43.0 ± 4.1	46.2 ± 6.6
Penal Law	35.4K	37.4 ± 12.8	56.4 ± 2.0	**86.3** ± 0.1	61.6 ± 6.7	60.4 ± 17.4
Social Law	29.1K	51.4 ± 5.8	54.8 ± 2.8	73.9 ± 1.9	70.3 ± 2.2	62.6 ± 9.7
All	60K	**58.6** ± 2.7	**66.8** ± 2.8	83.1 ± 1.3	**71.3** ± 2.4	**69.9** ± 8.8
All (w/o MT)	60K	58.4 ± 2.5	66.1 ± 2.4	83.1 ± 1.2	71.1 ± 1.4	69.7 ± 9.0

Table 8: Test results for models (XLM-R with MT unless otherwise specified) **adapted** per legal area (domain) or across all legal areas (domains). Best overall results are in **bold**, and in-domain are underlined. The mean and standard deviations are computed across languages per legal area and across legal areas for the right-most column. #D is the number of training examples per legal area.

Region	#D	#L	ZH	ES	CS	NWS	EM	RL	TI	FED	All
ZH	26.4K	de	<u>65.5 ± 0.0</u>	65.6 ± 0.0	63.7 ± 0.0	68.2 ± 0.0	62.0 ± 2.9	57.9 ± 6.7	63.2 ± 0.0	54.8 ± 5.1	62.6 ± 4.1
ES	17.1K	de	62.9 ± 0.0	<u>66.9 ± 0.0</u>	62.8 ± 0.0	65.2 ± 0.0	62.2 ± 1.1	60.2 ± 5.3	57.8 ± 0.0	55.1 ± 6.3	61.6 ± 3.6
CS	14.4K	de	62.5 ± 0.0	65.5 ± 0.0	<u>63.2 ± 0.0</u>	65.1 ± 0.0	60.7 ± 1.6	57.8 ± 3.7	60.5 ± 0.0	55.9 ± 0.5	61.4 ± 3.1
NWS	17.1K	de	66.0 ± 0.0	68.6 ± 0.0	65.2 ± 0.0	<u>67.9 ± 0.0</u>	61.6 ± 1.7	57.0 ± 4.9	57.1 ± 0.0	55.5 ± 5.7	62.4 ± 4.9
EM	24.9K	de,fr	64.1 ± 0.0	66.6 ± 0.0	63.3 ± 0.0	66.7 ± 0.0	<u>64.0 ± 0.7</u>	66.8 ± 2.9	63.2 ± 0.0	58.4 ± 0.3	64.1 ± 2.6
RL	40.2K	fr,de	61.0 ± 0.0	64.7 ± 0.0	60.2 ± 0.0	63.7 ± 0.0	63.4 ± 3.3	<u>69.8 ± 2.7</u>	67.6 ± 0.0	54.3 ± 7.2	63.1 ± 4.4
TI	6.9K	it	55.0 ± 0.0	56.3 ± 0.0	53.2 ± 0.0	54.5 ± 0.0	56.0 ± 0.4	54.7 ± 0.9	<u>66.0 ± 0.0</u>	53.1 ± 6.4	56.1 ± 3.9
FED	3.9K	de,fr,it	57.5 ± 0.0	59.6 ± 0.0	56.8 ± 0.0	58.9 ± 0.0	55.0 ± 1.0	56.5 ± 1.1	53.5 ± 0.0	<u>54.9 ± 2.9</u>	56.6 ± 1.9
All	60K	de,fr,it	**69.2** ± 0.0	**72.9** ± 0.0	68.3 ± 0.0	**73.3** ± 0.0	**69.9** ± 1.6	71.7 ± 2.8	**70.4** ± 0.0	**65.0** ± 3.9	**70.1** ± 2.5
All (w/o MT)	60K	de,fr,it	68.5 ± 0.0	71.3 ± 0.0	67.7 ± 0.0	71.2 ± 0.0	69.0 ± 1.5	71.4 ± 0.3	67.4 ± 0.0	64.6 ± 5.2	68.9 ± 2.2
All (Native)	60K	de,fr,it	69.0 ± 0.0	72.1 ± 0.0	**68.6 ± 0.0**	72.0 ± 0.0	**69.9 ± 1.6**	**71.9 ± 0.7**	68.8 ± 0.0	64.8 ± 7.0	69.6 ± 2.3

Table 9: Test results for models (XLM-R with MT unless otherwise specified) **fine-tuned** per region (domain) or across all regions (domains). Best overall results are in **bold**, and in-domain are <u>underlined</u>. The mean and standard deviations are computed across languages per origin region and across origin regions for the right-most column. The regions where only one language is spoken thus show std 0. #D is the number of training examples per origin region. #L are the languages covered.

Region	#D	#L	ZH	ES	CS	NWS	EM	RL	TI	FED	All
ZH	26.4K	de	65.4 ± 0.0	68.7 ± 0.0	63.9 ± 0.0	68.2 ± 0.0	63.6 ± 3.5	61.0 ± 2.8	66.4 ± 0.0	56.3 ± 1.8	64.2 ± 3.8
ES	17.1K	de	64.2 ± 0.0	69.4 ± 0.0	63.9 ± 0.0	66.0 ± 0.0	61.7 ± 2.3	59.4 ± 4.6	61.2 ± 0.0	56.5 ± 6.1	62.8 ± 3.7
CS	14.4K	de	63.1 ± 0.0	66.5 ± 0.0	64.1 ± 0.0	65.0 ± 0.0	61.0 ± 2.6	57.5 ± 2.1	62.2 ± 0.0	56.7 ± 2.5	62.0 ± 3.2
NWS	17.1K	de	65.8 ± 0.0	69.0 ± 0.0	63.8 ± 0.0	67.4 ± 0.0	59.9 ± 3.3	58.6 ± 1.1	58.9 ± 0.0	54.2 ± 2.7	62.2 ± 4.8
EM	24.9K	de,fr	63.9 ± 0.0	67.5 ± 0.0	64.4 ± 0.0	66.8 ± 0.0	64.7 ± 0.5	69.1 ± 1.7	66.4 ± 0.0	59.5 ± 1.0	65.3 ± 2.7
RL	40.2K	fr,de	62.3 ± 0.0	66.2 ± 0.0	62.0 ± 0.0	64.7 ± 0.0	65.2 ± 4.2	70.8 ± 6.8	65.5 ± 0.0	56.9 ± 6.0	64.2 ± 3.7
TI	6.9K	it	56.4 ± 0.0	62.1 ± 0.0	53.7 ± 0.0	56.3 ± 0.0	55.1 ± 0.2	57.4 ± 1.1	68.3 ± 0.0	50.5 ± 2.3	57.5 ± 5.1
FED	3.9K	de,fr,it	52.7 ± 0.0	52.7 ± 0.0	51.3 ± 0.0	53.1 ± 0.0	52.8 ± 0.7	52.0 ± 2.3	52.8 ± 0.0	50.0 ± 4.0	52.2 ± 1.0
All	60K	de,fr,it	**69.2** ± 0.0	73.3 ± 0.0	**69.9** ± 0.0	**73.0** ± 0.0	**70.3** ± 1.9	**72.1** ± 0.7	**70.9** ± 0.0	63.8 ± 6.1	**70.3** ± 2.8
All (w/o MT)	60K	de,fr,it	**69.2** ± 0.0	**73.9** ± 0.0	67.9 ± 0.0	72.6 ± 0.0	69.0 ± 2.1	**72.1** ± 0.3	70.1 ± 0.0	**64.2** ± 4.6	69.9 ± 2.9

Table 10: Test results for models (XLM-R with MT unless otherwise specified) **adapted** per region (domain) or across all regions (domains). Best overall results are in **bold**, and in-domain are <u>underlined</u>. The mean and standard deviations are computed across languages per origin region and across origin regions for the right-most column. The regions where only one language is spoken thus show std 0. #D is the number of training examples per origin region. #L are the languages covered.

ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US

Gil Semo* **Dor Bernsohn** **Ben Hagag** **Gila Hayat**
Darrow AI Ltd.
30 Ha'arbaa Street, Tel Aviv, Israel
firstname.lastname@darrow.ai

Joel Niklaus*†
Niklaus.ai
Schwarztorstrasse 108, Bern, Switzerland
joel@niklaus.ai

Abstract

The research field of Legal Natural Language Processing (NLP) has been very active recently, with Legal Judgment Prediction (LJP) becoming one of the most extensively studied tasks. To date, most publicly released LJP datasets originate from countries with civil law. In this work, we release, for the first time, a challenging LJP dataset focused on class action cases in the US. It is the first dataset in the common law system that focuses on the harder and more realistic task involving the complaints as input instead of the often used facts summary written by the court. Additionally, we study the difficulty of the task by collecting expert human predictions, showing that even human experts can only reach 53% accuracy on this dataset. Our Longformer model clearly outperforms the human baseline (63%), despite only considering the first 2,048 tokens. Furthermore, we perform a detailed error analysis and find that the Longformer model is significantly better calibrated than the human experts. Finally, we publicly release the dataset and the code used for the experiments.

1 Introduction

Recently, the literature in Legal Natural Language Processing (NLP) has grown at a fast pace, firmly establishing it as an important specialized domain in the broader NLP ecosystem. As part of this strong growth and as a first step establishing Legal NLP in the field, many legal datasets have been released in the fields of Legal Judgment Prediction (LJP) (Niklaus et al., 2021a; Chalkidis et al., 2019), Law Area Prediction (Glaser and Matthes, 2020), Legal Information Retrieval (Wrzalik and Krechel, 2021), Argument Mining (Urchs et al., 2022), Topic Classification (Chalkidis et al., 2021a), Named Entity Recognition (Luz de Araujo et al., 2018; Angelidis et al., 2018; Leitner et al., 2019), Natural Language Inference (Koreeda and Manning, 2021), Question Answering (Zheng et al., 2021; Hendrycks et al., 2021), and Summarization (Shen et al., 2022; Kornilova and Eidelman, 2019).

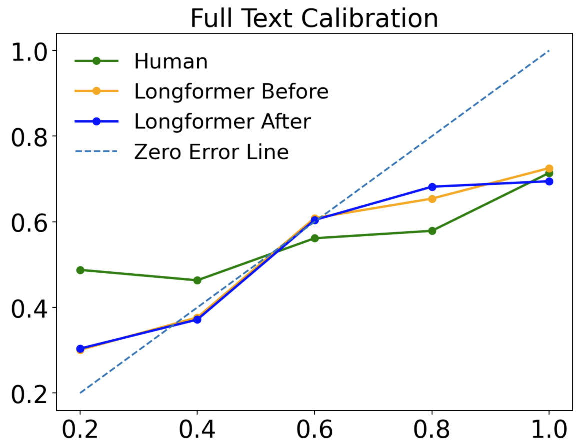


Figure 1: Calibration plot on the Full Text dataset. The human experts rated the confidence of their predictions on a score from 1 to 5. The confidence scores of the Longformer models were binned into 5 buckets.

In particular, the field of LJP has been very active, with many datasets released recently. Cui et al. (2022) surveyed the field and divided the datasets into five subtasks. In this work, we release a dataset belonging to the category of the Plea Judgment Prediction (PJP) task. Most other PJP datasets use the facts summary, written by the court (clerks or judges) as input (Cui et al., 2022). The facts are written in such a way as to support the final decision (Niklaus et al., 2021a) and require extensive work by highly qualified legal experts (Ma et al., 2021). In contrast, in this work we consider the plaintiff's pleas (AKA complaints) as input, making the task more realistic for use in real-world applications.

Most LJP datasets released so far are from countries with civil law. Our dataset originates from the United States, the largest country employing the common law legal system. To the best of our knowledge, we are the first to release a dataset specifically targeting class action lawsuits.

* Equal Contribution
† Corresponding Author

Proceedings of the Natural Legal Language Processing Workshop 2022, pages 31 - 46
December 8, 2022

Motivation

The 16th United Nations Sustainable Development Goal (UNSDG) is to "Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels". Class actions are a private enforcement instrument that enables courts to organize the mass adjudication of meritorious claims by underrepresented individuals and communities. Without class actions, many victims of illegal action would never get their day in court. Making case outcomes and facts accessible is crucial to strengthen the effective use of class actions and private enforcement to drive UNSDG 16. With the power of early LJP, plaintiffs will have the ability to bring only meritorious cases to court, and defendants are more likely to resolve them faster.

Main Research Questions

In this work, we pose and examine three main research questions:
RQ1: *To what extent is it possible to determine the outcome of US class action cases using only the textual part of the complaints (without metadata)?*
RQ2: *To what extent can we use Temperature Scaling (TS) to better calibrate our models?*
RQ3: *To what extent can expert human lawyers solve the proposed task?*

Contributions

The contributions of this paper are four-fold:

- We curate a new specialized dataset of 10.8K class action complaints in the US from 2012 to 2022 annotated with the binary outcome: win or lose (plaintiff side). In contrast to most other LJP datasets it is (a) from a country with the common law system (where there are less datasets available), (b) it is specialized to class actions (important types of complaints ensuring justice for numerous often under-represented individuals), and (c) it uses the plaintiff's pleas as input instead of the facts, making the task more realistic. To the best of our knowledge, our work is the first dataset with plaintiff's pleas in the common law system and in the English language.
- We conduct a detailed analysis of the studied models using Integrated Gradients (IG) and model calibration using TS (Guo et al., 2017a).
- We perform an experiment with human experts on a randomly selected subset of the dataset, showing that our Longformer model both outperforms the human experts in terms of accuracy and calibration.
- We publicly release a sample of 3,000 cases from the annotated dataset[1] together with the human expert labels[2] and the code for the experiments[3].

2 Legal Background

2.1 Class Action Lawsuits

Class actions are a unique procedural instrument that allows one person to sue a company, not only on behalf of himself, but for everyone that has been injured by the same wrongdoing. In contrast to traditional lawsuits, in a class action lawsuit a plaintiff sues the defendant(s) on behalf of a class of absent parties. Class action lawsuits typically involve a minimum of 40 claimants. Rather than filing individual lawsuits for each damaged person, class actions allow the plaintiffs to unite and sue through a single proceeding. Thus, class actions are usually large and important cases and contain more complexity due to the high number of represented plaintiffs. These characteristics make class action a legal enforcement mechanism, along with police and regulators. Class actions both deter companies from harming people in the first place, and give compensation to the large number of victims hurt by the violation, giving consumers power over large corporations.

2.2 Definitions

Civil Law vs. Common Law: In both civil law and common law systems, courts rule based on laws and precedents (previous case law, mostly from the Supreme Court). However, in common law countries (mainly present in the UK and its former Colonies), case law dominates, whereas in civil law countries (most other countries) laws are more important. Note, that the differences are often not clear-cut, and courts usually use a combination of both laws and precedent for their rulings.
Complaint: A complaint is a written pleading to initiate a lawsuit. It includes the plaintiff's cause of action, the court's jurisdiction, and the plaintiff's demand for judicial relief. It is necessary for

[1] https://huggingface.co/datasets/darrow-ai/USClassActions
[2] https://huggingface.co/datasets/darrow-ai/USClassActionOutcomes_ExpertsAnnotations
[3] https://github.com/darrow-labs/ClassActionPrediction

the complaint to state all of the plaintiff's claims against the defendant, as well as what remedy the plaintiff seeks. A complaint must state "enough facts to state a claim to relief that is plausible on its face" (Twombly, 2007). The standards for filing a complaint vary from state to federal courts, or from one state to another. A typical class action complaint contains the allegations, the background details about both the plaintiff and the defendant, and the facts.

Allegations: In a complaint, allegations are statements of claimed facts. These statements are only considered allegations until they are proven. An allegation can be based on information and belief if the person making the statement is unsure of the facts. In the complaint, allegations can appear twice: once as a summary at the beginning and once in more detail later. There is usually a reference to the act that the plaintiff's attorney claims to have been violated in the allegations.

Background Details: The complaint contains background sections such as the plaintiff's history, class definitions, the defendant's history, and details about the platform/service in which the allegations took place.

Plaintiff's Facts: The plaintiff's facts or "factual background", are statements that can be proven and are often backed up with references and event dates. Note that the plaintiff's facts are written by the plaintiff lawyers.

Facts Summary: The facts summary or "factual description", are the summary of the accepted facts by the court and are written by the clerks or judges. The facts summary is usually more condensed in higher courts. Most previous LJP tasks used facts of this type. Since in this paper we consider complaints as input, when "facts" are mentioned we refer to the plaintiff's facts.

Case Description: The case description is written by the court clerks or judges and usually includes the header, the facts, the considerations, and the rulings.

Class Action Outcomes

Table 1 shows the outcomes possible in class action cases. In the following, we briefly describe each of the outcomes.

Settled: "Settling a case" refers to resolving a dispute before the trial ends.

Uncontested Dismissal: Without any opposition from the parties, the case is dismissed and closed.

Motion to Dismiss: The case was dismissed by the court following the defendant's formal request for a court to dismiss the case.

Outcome	Bin. Label	# Examples (%)
Settled	win	5234 (48.64%)
Other - Plaintiff	win	58 (00.52%)
Uncontested Dismissal	lose	4544 (42.23%)
Motion to Dismiss	lose	755 (07.01%)
Other - Defendant	lose	170 (01.56%)

Table 1: This table shows the original outcome together ruled by the court with the frequency and the final binarized label we map it to.

3 Related Work

LJP is an important and well-studied task in legal NLP. Cui et al. (2022) subdivide LJP into five subtasks: (a) In the *Article Recommendation Task*, systems predict relevant law articles for a given case (Aletras et al., 2016; Chalkidis et al., 2019; Ge et al., 2021). (b) The goal of the *Charge Prediction Task*, mainly studied in China, is to predict the counts the defendant is charged for based on the facts of the case (Zhong et al., 2018; Hu et al., 2018; Zhong et al., 2020). (c) In the *Prison Term Prediction Task*, systems predict the prison time for the defendant as ruled by the judge (Zhong et al., 2018; Chen et al., 2019). (d) In the *Court View Generation Task*, systems generate court views (explanation written by judges to interpret the judgment decision) (Ye et al., 2018; Wu et al., 2020). (e) In the *Plea Judgment Prediction Task*, systems predict the case outcome based on the case's facts (Niklaus et al., 2021b; Şulea et al., 2017; Lage-Freitas et al., 2022; Long et al., 2019; Ma et al., 2021; Strickson and De La Iglesia, 2020; Malik et al., 2021a; Alali et al., 2021). Since our work belongs to the PJP category, in the following, we elaborate more on the related work in this area.

Civil Law Niklaus et al. (2021b) released a trilingual (German, French, Italian) Swiss dataset from the Federal Supreme Court of Switzerland. They use the facts summary as input and predict a binary output: approval or dismissal of the plaintiff's pleas for approx. 85K decisions. Şulea et al. (2017) released a dataset of approx. 127K French Supreme Court cases. As input, they used the entire case description and not only the facts summary, presumably making the task considerably easier and a possible explanation for their high performance on the dataset. As output, they consider up to 8 classes of decisions ruled by the court. Lage-Freitas et al.

(2022) released a dataset comprising roughly 4K cases from a Brazilian State higher court (appellate court). They predicted three labels from the entire case description (written by the judges/clerks). Jacob de Menezes-Neto and Clementino (2022) release a large dataset of over 765K cases from the 5th Regional Federal Court of Brazil. They investigate a binary prediction task (whether the previous decision was reversed or not) using the entire case description as input. Long et al. (2019) studied the LJP task on 100K Chinese divorce proceedings considering three types of information as input: applicable law articles, fact description, and plaintiffs' pleas. Their model predicts a binary output. Ma et al. (2021) released a dataset comprising 70.5K civil cases (private lending) from China. They consider the more realistic task of inputting the plaintiff's complaints (together with debate data) instead of the easier facts summary used by most previous works. As output, their models predict three classes (reject, partially support and support). Similarly, our work also studies the more realistic (and challenging) use case of using the plaintiff's pleas as input instead of the heavily processed facts.

Common Law Strickson and De La Iglesia (2020) released a dataset of 5K cases from the UK's highest court of appeal. As input, they consider the case description and their models predict two labels (allow vs. dismiss). Malik et al. (2021a) study a dataset of 35K Indian Supreme Court cases in English. They use the case description as input and predict a binary outcome (accepted vs. rejected). Alali et al. (2021) study a dataset of 2.4K US Supreme Court decisions. Their models used the facts summary as input and predicted a binary output (first party won vs. second party won). In contrast, our dataset is $\sim$ 5 times larger and is specialized to the rare subset of class action cases.

Apart from Ma et al. (2021), the PJP task based on plaintiff's complaints has not been studied before. In contrast, most previous works studied textual input originating from the case description written by the court.

4 Dataset Description

In this section, we describe the dataset origin and statistics in detail. Additionally, we elaborate on the dataset construction process and the variants we produced.

Figures 2a and 2b show the distribution of cases across the most frequent states and courts in the dataset, respectively. Note that the origin of these class action lawsuits is very diverse, both across states and across courts. Not surprisingly, population-rich states like California, Florida, and New York lead the list. However, while California is more than double in population (39.5M vs. 20.2M as of April 2021), the number of class action lawsuits has the inverse relationship ($\sim$ 3K from New York and $\sim$ 1.8K from California). We assume that the complicated filing system in California could be a reason for this disparity[4].

4.1 Plaintiff's Pleas Instead of Facts Summary

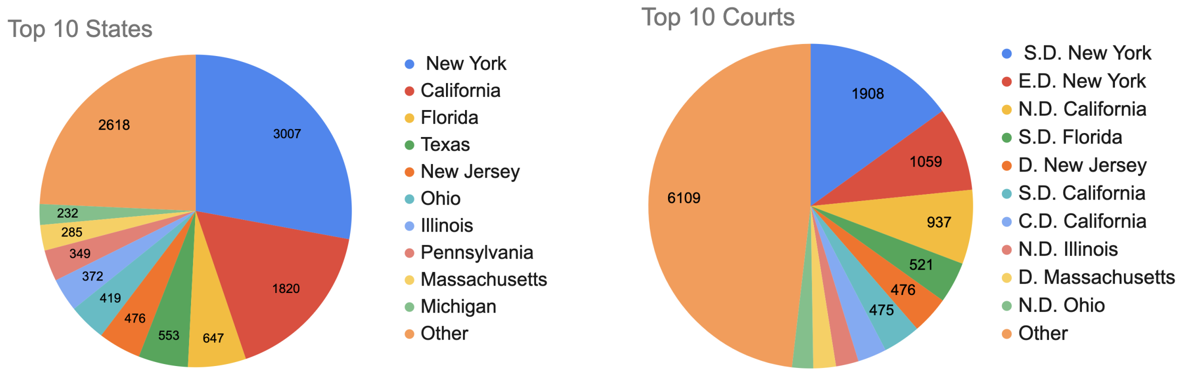
Condensing and extracting the relevant information from plaintiffs' pleas and court debates is a large part of the judge's work (Ma et al., 2021). This results in a condensed description of a case's facts. Most previous works consider this condensed description written by the judicial body (judges and clerks) as input. However, since a lot of qualified time has been spent on writing these descriptions, naturally, it makes the LJP task easier when using the court-written facts as input. Ma et al. (2021) were the first to consider the original plaintiff's pleas as input on Chinese data. In this work, to the best of our knowledge, we are the first to consider this harder task in the common law system (US class action cases in our case) and in the English language in general.

We do not consider the background details because our models might easily overfit on very specific data. In contrast, our goal was to create a dataset, where models need to focus on case-specific details to solve the task instead of being allowed to consider company-specific information such as number of employees or the area of business. We also disregard the introduction, containing metadata about the judge and the plaintiff.

4.2 Dataset Construction

To extract the plaintiffs' facts and allegations from each case, we manually reviewed hundreds of cases from different courts and different states to learn the structure of the document in each court to build a rule-based regex extraction system that detects the relevant text spans in each complaint. To summarize, constructing the dataset posed many technical difficulties due to the diverse nature of the

[4]Each court has its format of filing, and even courts within the same county do not usually use the same complaint filing format.



(a) Top 10 most frequent states (b) Top 10 most frequent courts

Figure 2: Distribution of cases across states and courts.

complaint documents. At the preprocessing stage, we perform text cleaning, including removing some irrelevant text sections that our system incorrectly matched and removing duplicate sections.

4.3 Label Distribution

In this work we consider the task of binary legal judgment prediction. To do so, we simplified the labels. We used Table 1 to map the outcomes to either *win* or *lose* (for the plaintiff). After binarization the dataset is almost balanced with 5,469 (50.8%) *lose* cases and 5,290 (49.2%) *win* cases. Therefore, in our experiments, we just report the accuracy to keep it simple and make the scores more easily interpretable.

4.4 Dataset Variants

We experimented with different variants of the dataset to study the effect of the different parts of the text. We deliberately focused our attention more on the allegations because the facts contain a lot of repetitive content and are noisier than the allegations (many paragraphs only contain citations). Additionally, the facts contain many citations to laws, which are less relevant to the case's outcome according to domain experts (the facts are more generic and less case-specific than the allegations).

Full Text

The *Full Text* dataset combines the plaintiff's facts and the allegations but also disregards any background details. We concatenated the facts at the beginning and added the allegations parts to create one input text. We observe in Figure 3a that this dataset is rather long – almost 2700 tokens on average – with 10% of cases longer than 5400 tokens.

Unified Allegations

The *Unified Allegations* dataset consists of all case's allegations (mentioned in the complaint) concatenated together to form one input text . Approx. 2K documents did not contain any allegations (based on our extraction regexes), reducing the dataset size from 10.8K to 8.8K documents. The allegations make up a bit less than half of the full text complaint, as shown in Figure 3b (mean of ∼1,100 tokens and percentile 90 at ∼2,400 tokens).

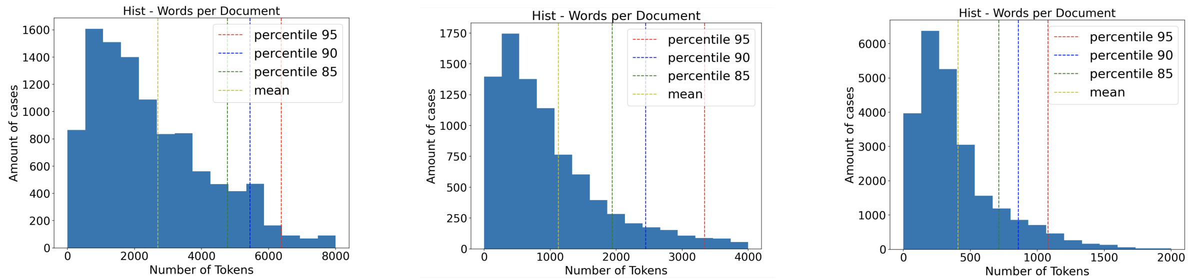
Separated Allegations

The *Separated Allegations* dataset considers each allegation as a separate sample, increasing the size from 8.8K to 25K documents. We considered this dataset to test whether the entire context is necessary. Figure 3c shows the length distribution for individual allegations. Surprisingly, even a single allegation can reach up to 2,000 tokens (∼ 4-5 pages of continuous text). However, most allegations (95%) are not longer than roughly 2 pages (1,100 tokens) with the average at 400 tokens.

5 Experiments

5.1 Experimental Setup

For all experiments, we truncated the text to the model's maximum sequence length (2,048 for Longformer and BigBird, 512 otherwise), unless otherwise specified. All experiments have been performed on the binarized labels (win or lose). We ran the experiments with 5-fold cross-validation and averaged across 5 random seeds. For more details regarding hyperparameter tuning and preprocessing, please refer to Appendix A.



(a) Full Text (b) Unified Allegations (c) Separated Allegations

Figure 3: Histograms for the three dataset variants (number of tokens calculated using bert-base-uncased tokenizer).

5.2 Methods

We compared the following pretrained transformer models: BERT (Devlin et al., 2019), LegalBERT (Chalkidis et al., 2020) (pretrained on diverse English legal data from Europe and the US with a domain-specific tokenizer), CaseLawBERT (Zheng et al., 2021) (pretrained on 37GB of US state and federal caselaw with a domain specific tokenizer), LegalRoBERTa[5] (continued pretraining from RoBERTa checkpoint on 4.6 GB of US caselaw and patents), BigBird (Zaheer et al., 2021) and Longformer (Beltagy et al., 2020). For all models, we used the publicly available base checkpoints on the Huggingface hub[6]. We ran our experiments with the Huggingface transformers library (Wolf et al., 2020) available under an Apache-2.0 license.

5.3 Results

Results are reported in the $mean_{\pm std}$ format averaged accuracy across 5 random seeds. Table 2 shows the main results. We observe that the setup considering the entire text is harder than when we only consider the allegations (best Full Text model is at $\sim$ 63% and worst allegations model is at $\sim$ 65%). These findings confirm our hypothesis, that the allegations encode more useful information than the facts (see Section 4.4) (the facts are often at the beginning of the complaints; thus the models on the Full Text dataset are likely to see mostly facts because of the truncation).

In line with previous findings (Chalkidis et al., 2021b, 2020; Zheng et al., 2021), models with legal pretraining outperform BERT also in our datasets (Unified Allegations and Separated Allegations). However, for LegalBERT the difference is small (only 0.5% above BERT). The models pretrained mostly or exclusively on US caselaw (LegalRoBERTa or CaseLawBERT respectively) perform better (up to 2% better than BERT), presumably because our dataset also originates from the US. CaseLawBERT achieves a much higher difference to BERT on the CaseHOLD task (4.6 F1) (Zheng et al., 2021) and on SCOTUS (7.6 macro-F1) (Chalkidis et al., 2021b). Both of these tasks are based on the same data as has been used in the pre-training of LegalRoBERTa and CaseLawBERT, whereas the complaints in our dataset are unseen by all models during pre-training. We suspect that this different data is the reason for the legal models not outperforming BERT as strongly as has been observed in other datasets.

Method	Accuracy
Full Text (trunc. to 2048 tokens)	
Longformer	$62.87_{\pm 2.06}$
BigBird	$63.26_{\pm 3.40}$
Unified Allegations (trunc. to 512 tokens)	
BERT	$65.06_{\pm 1.67}$
LegalBERT	$65.57_{\pm 0.26}$
CaseLawBERT	$65.87_{\pm 0.60}$
LegalRoBERTa	$65.95_{\pm 0.98}$
Separated Allegations (trunc. to 512 tokens)	
BERT	$64.98_{\pm 1.08}$
LegalBERT	$65.57_{\pm 0.62}$
CaseLawBERT	$66.82_{\pm 0.78}$
LegalRoBERTa	$65.97_{\pm 0.88}$

Table 2: Longformer and BigBird used a maximum sequence length of 2,048 tokens. All other models used 512 tokens. For all datasets, we truncated the text to fit the maximum sequence length.

6 Error Analysis

Neural Networks (NNs) and their latest incarnation, Transformers (Vaswani et al., 2017), work very well across a wide range of tasks, especially if

[5] https://huggingface.co/saibo/legal-roberta-base
[6] https://huggingface.co/models

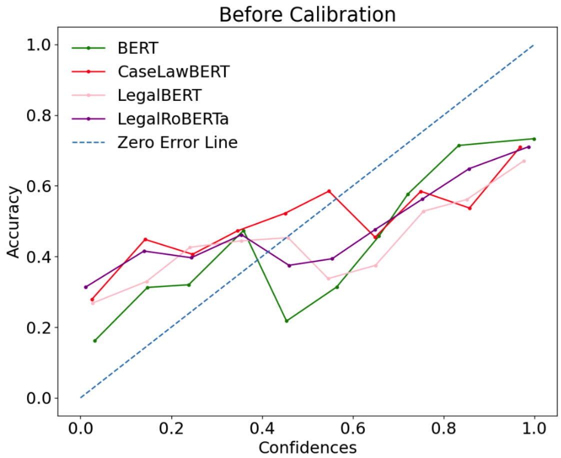
the tasks involve more "complicated" data like text or images. However, in contrast to traditional Machine Learning (ML) methods such as Linear Regression, they are not interpretable out-of-the-box. Neural Networks need additional methods to make them explain themselves better to humans. A rich body of literature investigates how to make NNs and especially Transformers more interpretable (Ribeiro et al., 2016; Sundararajan et al., 2017; Lundberg and Lee, 2017; Dhamdhere et al., 2018; Serrano and Smith, 2019; Bai et al., 2021). Interpretability is especially important in high-stakes domains such as law or medicine.

In the following two sections, we analyze our models using the two interpretability methods Calibration and IG to get a better understanding of their inner workings.

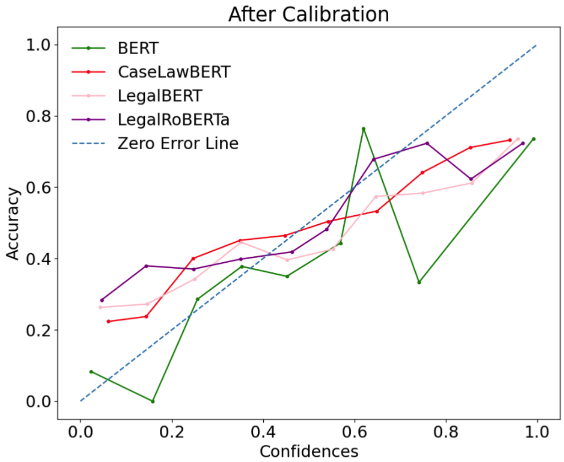
6.1 Calibration

In this section, we investigate to what extent our models are calibrated out-of-the-box and "calibratable". Calibration is a first step towards understanding whether the model output can be trusted (Guo et al., 2017b; Desai and Durrett, 2020): how aligned are the confidence scores with the actual empirical likelihoods? Thus, if the model assigns 60% probability to a label, then this label should be correct in 60% of cases if the model is calibrated. So, even if the model itself is a black-box, a calibrated model at least gives an indication whether it knows when it is wrong. This information can be very valuable when deploying models in the real world because it allows us to discard predictions where the model is below some certainty threshold. Well calibrated models are especially important in domains with high potential downside for users, such as predictive tools for court cases.

In this work, we follow Desai and Durrett (2020) by employing TS (Guo et al., 2017b) for calibrating our models using the netcal library[7] (Küppers et al., 2020) available under an Apache License 2.0 license. We show calibration plots in Figure 4 for BERT and the legal models on the Unified Allegations dataset and aggregated scores in Table 5 in Appendix B.3. We observe that the legal models are less calibrated than BERT before, but better calibrated after TS. So TS seems to calibrate domain-specific models better than general models. When comparing the calibration of our models with



(a) Before Calibration



(b) After Calibration

Figure 4: Calibration on the Unified Allegations dataset.

the calibration of models from the literature (Desai and Durrett, 2020), we note that our models are less calibrated overall (further away from the zero-error-line and higher ECE scores), both out-of-the-box and after applying TS. We hypothesize that the generally lower accuracy on our hard dataset also makes the models less calibrated, especially in the areas of high (> 0.8) and low (< 0.2) confidence. To the best of our knowledge, in legal NLP we are the first to perform such an analysis.

6.2 Integrated Gradients

We conduct a qualitative analysis of the LegalBERT model using IG[8] (Sundararajan et al., 2017) and show an illustrative example in Figure 5. We observe that the model focuses most on "flsa" an acronym for Fair Labor Standards Act[9] regulating

[7]https://github.com/fabiankueppers/calibration-framework

[8]https://github.com/cdpierse/transformers-interpret#sequence-classification-explainer

[9]https://www.dol.gov/agencies/whd/flsa

minimum wage and overtime among others. Further, the model focuses on "work" and "wages" possibly signaling a (limited) understanding of the connections between those concepts. Future work may investigate explainability of Pretrained Language Models (PLMs) in more detail on the LJP task.

7 Human Expert Annotations

Malik et al. (2021b) collected predictions for the judgment outcome of Indian Supreme Court cases from five legal experts. The experts agreed with the judges in 94% of the cases, on average. Note, however, that they have access to both the facts summary and the court's considerations. Their best model, XLNet + BiGRU, only achieves an accuracy of 78%. Contrarily, Jacob de Menezes-Neto and Clementino (2022) find that all their models outperform 22 highly skilled experts on LJP in Brazilian Federal Courts using the entire case description for prediction.

We asked legal experts (employees of our company) and US law students in their final year, to predict the judgment outcome of 200 randomly selected examples in our Full Text dataset. Note that they only had access to the facts and allegations from the plaintiff's pleas (same as our models), and not to the court case written by the judge. So, their task was much more difficult than the one posed to the annotators by Malik et al. (2021b) and Jacob de Menezes-Neto and Clementino (2022). In our task, participants (whether models or human experts) basically need to estimate how the court is going to decide based only on the plaintiff's pleas. For each document, our legal experts had to answer whether they think the plaintiff would win or lose the case. Furthermore, they also had to indicate their confidence level for being correct (from 1 – very unsure – to 5 – very sure). We made sure that the annotators did not look for any additional information regarding the complaint (e.g., news articles about the outcome or further information on different legal platforms) so that their answer is based only on the input text presented on the annotation platform. Figure 6 in Appendix C presents a screenshot of the annotation platform we used.

On the entire dataset sample (200 examples), the human experts achieve an accuracy of 53%. When we filtered out the samples where the human experts were not confident (confidence score 1, 2 or 3), they achieved an accuracy of 60%. The entire results for the human experts are shown in Appendix B.4 in Table 6. We also trained and evaluated a Longformer model for comparison with the human predictions. We randomly split our remaining dataset into 6,877 train and 1,851 validation examples. Surprisingly, the Longformer model outperforms the human expert predictions both on the entire annotated test dataset (63% vs. 53% Accuracy) and the dataset filtered for high human confidence (67% vs. 60% Accuracy). In contrast to the human experts, the Longformer model only had access to the first 2,048 tokens of the case. While the human performance increases more than the Longformer performance on the high-confidence dataset, the Longformer model also has a higher performance, suggesting that these cases are easier to predict.

The task proposed in our dataset seems very challenging, given that human experts face great challenges in solving it. Interestingly, on the Indian dataset the humans clearly outperform the models, whereas in the Brazilian dataset it is reversed, similar to our results. Note that lawyers are often specialized in very narrow domains (legal areas). The cases in our dataset may be very diverse, and thus a generic model might be better suited for this task than specialized human experts. Future work may investigate this finding in more detail.

Figure 1 shows the calibration plot on the Full Text dataset, comparing Longformer before and after calibration with the human confidence scores. We observe that Longformer is already well calibrated in comparison to the human experts. Using TS, the Expected Calibration Error (ECE) of Longformer can be reduced from 5.14 to 2.34, whereas the ECE of the human experts lies at 17.5. Again, as mentioned in Section 6.1, the lower accuracy of the humans might explain their worse calibration compared to Longformer.

8 Conclusions and Future Work

Answers to the Research Questions

RQ1: *To what extent is it possible to determine the winner of US class action cases using only the textual part of the complaints (without metadata)?* It is possible, to some extent, to determine the winner of US class action cases using only the textual part of the complaints. Our best model achieves an accuracy of 66.8% (LegalRoBERTa) on the datasets using only the allegations. However, as this number shows, there is still a lot of room for improvement.

Predicted label = 1: Case Won

[CLS] plaintiff hereby real ##leg ##es and incorporates paragraphs 1 through 43 of this complaint , as if fully set forth herein . defendants failed to pay over ##time wages to plaintiff and other similarly situated employees for all time worked in excess of forty (40) hours in individual work weeks in violation of the flsa , 29 u . s . c . § 201 . for example , during the week beginning april 4 , 2016 , plaintiff worked approximately fifty - six (56) hours for defendants . during the week beginning may 16 , 2016 , 9 plaintiff worked approximately fifty - four (54) hours for defendants . plaintiff was not paid a rate of one and one - half times his regular rate of pay for all time worked in excess of forty (40) in these weeks and all other weeks he worked over forty (40) hours . during the course of their employment with defendants , plaintiff and others similarly situated driver ##s were not exempt from the maximum hour provisions of the flsa , 29 violation of the fair labor standards act [UNK] over ##time wages (collective action under 29 u . s . c . § 216 (b)) [SEP]

Figure 5: Analysis using Integrated Gradients (IG)

RQ2: *To what extent can we use Temperature Scaling (TS) to better calibrate our models?* Similar to Natural Language Inference, Paraphrase Detection and Commonsense Reasoning tasks (Desai and Durrett, 2020), we also find that in the PJP task, TS helps in calibrating pretrained transformers. In our best model, TS led to a decrease in ECE scores from 28 to 2.

RQ3: *To what extent can expert human lawyers solve the proposed task?* Expert human lawyers perform better than chance on a randomly selected dataset of 200 samples and can increase their accuracy from 53% to 60% when they are confident in their decision. However, they are still outperformed by a Longformer model having access to only the first 2,048 tokens in both scenarios.

Conclusions

We release a challenging new dataset of class action lawsuits for the more realistic PJP task (where the input is based on the complaints instead of the further processed facts summary written by the judge) in the US, a jurisdiction with the common law system. Additionally, we calibrated our models using TS and found that despite the relatively low accuracy (66% for the best model), relatively low ECE scores around 2 can be achieved. Finally, we find that our Longformer model is 10% more accurate than the human experts on our dataset despite having only access to the first 2,048 tokens of the case.

Limitations

Our best model achieves an accuracy of 66%. This may suggest that either the task posed in this dataset is very hard, or we did not optimize our models enough. The results achieved by the human experts suggests that the former is the case. However, we believe much more work is needed here.

Although we did some first efforts to interpret our model's outputs using Calibration and IG, the literature knows a host of other explainability methods (Molnar, 2022). We leave a more thorough qualitative analysis involving domain experts and explainability methods for future work.

Our experiments were performed only on relatively short input spans (512 tokens for allegations, and 2048 for full text). Longformer or BigBird support input spans until 4096 tokens. Another possibility is the use of hierarchical models, as employed for example by Niklaus et al. (2022); Dai et al. (2022) that can also easily scale to 4096 tokens given the right hardware. With 4096 tokens, we could fully encode all allegations and almost 80% percent of the full texts. We leave these investigations to future work.

Future Work

Since the legal models outperformed BERT only to a small margin, we suspect that further pretraining (Gururangan et al., 2020) on in-domain data might further enhance the performance. Additionally, in future work, we plan to study the domain-specific PJP and whether domain-specific models are better than generic model or human experts.

Large PLMs have proved to be very strong few shot learners in many tasks (Brown et al., 2020; Chowdhery et al., 2022). The use of such models may bring performance boosts also in our studied task. We leave experimentation using different prompting strategies for future work (Arora et al., 2022; Wei et al., 2022; Suzgun et al., 2022).

We discovered through our analysis using IG that some legal domains have a strong correlation to a particular label. To produce complaints with a higher success likelihood in court, future studies may examine the linguistic structure of successful allegations.

Ethics Statement

The goal of this research is to achieve a better understanding of LJP to broaden the discussion and aid practitioners in developing better technology for both legal experts and non-specialists. We believe that this is a crucial application area, where research should be done (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize legal data, making it more accessible to end-users, while also highlighting (informing the audience on) the various multi-aspect deficiencies seeking a responsible and ethical (fair) deployment of legal-oriented technology.

In this direction, we study how we can best build our dataset to maximize accuracy of our models on the task. Additionally, we study the inner workings of the models using Integrated Gradients and make sure that our models are calibrated. A well calibrated model outputs confidence probabilities in line with actual likelihoods, thus giving the users the possibility of discarding low-confidence predictions or at least treating them with caution.

Lawyers often perform the LJP task by giving their clients advice on how high the chances for success are in court for specific cases. Given the complaint documents, we were able to show in this work that our models outperformed human experts in this task.

But, like with any other application (like content moderation) or domain (e.g., medical), reckless usage (deployment) of such technology poses a real risk. According to our opinion, comparable technology should only be used to support human specialists (legal scholars, or legal professionals).

Acknowledgements

We thank all the anonymous reviewers for their insightful comments. We thank the two employees at Darrow for the annotation of the dataset.

References

Mohammad Alali, Shaayan Syed, Mohammed Alsayed, Smit Patel, and Hemanth Bodala. 2021. JUSTICE: A Benchmark Dataset for Supreme Court's Judgment Prediction. ArXiv:2112.03414 [cs].

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93. Publisher: PeerJ Inc.

I. Angelidis, Ilias Chalkidis, and M. Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *JURIX*.

Simran Arora, Avanika Narayan, Mayee F. Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask Me Anything: A simple strategy for prompting language models. ArXiv:2210.02441 [cs].

Bing Bai, Jian Liang, Guanhua Zhang, Hao Li, Kun Bai, and Fei Wang. 2021. Why Attentions May Not Be Interpretable? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pages 25–34, New York, NY, USA. Association for Computing Machinery.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*. ArXiv: 2004.05150.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021a. MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv:2109.00904 [cs]*. ArXiv: 2109.00904.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559 [cs]*. ArXiv: 2010.02559.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael James Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021b. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. SSRN Scholarly Paper ID 3936759, Social Science Research Network, Rochester, NY.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. 2019. Charge-Based Prison Term Prediction with Deep Gating Network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6362–6367, Hong Kong, China. Association for Computational Linguistics.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. Association for Computing Machinery.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv:2204.02311 [cs]*. ArXiv: 2204.02311.

Junyun Cui, Xiaoyu Shen, Feiping Nie, Z. Wang, Jinglong Wang, and Yulong Chen. 2022. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *ArXiv*, abs/2204.04859.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting Transformer-based Models for Long Document Classification. *arXiv:2204.06683 [cs]*. ArXiv: 2204.06683.

Shrey Desai and Greg Durrett. 2020. Calibration of Pre-trained Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Kedar Dhamdhere, Mukund Sundararajan, and Qiqi Yan. 2018. How Important Is a Neuron? ArXiv:1805.12233 [cs, stat].

Jidong Ge, Yunyun huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. 2021. Learning Fine-grained Fact-Article Correspondence in Legal Cases. ArXiv:2104.10726 [cs].

Ingo Glaser and Florian Matthes. 2020. Classification of German Court Rulings: Detecting the Area of Law. page 10.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017a. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On Calibration of Modern Neural Networks. Number: arXiv:1706.04599 arXiv:1706.04599 [cs].

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv:2004.10964 [cs]*. ArXiv: 2004.10964.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. ArXiv:2103.06268 [cs].

Zikun Hu, Xiang Li, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2018. Few-Shot Charge Prediction with Discriminative Legal Attributes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 487–498, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Elias Jacob de Menezes-Neto and Marco Bruno Miranda Clementino. 2022. Using deep learning to predict outcomes of legal appeals better than human experts: A study with data from Brazilian federal courts. *PLOS ONE*, 17(7):e0272287.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A Corpus for Automatic Summarization of US Legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. 2020. Multivariate confidence calibration for object detection. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting Brazilian Court Decisions. *PeerJ Computer Science*, 8:e904. Publisher: PeerJ Inc.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-Grained Named Entity Recognition in Legal Documents. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, volume 11702, pages 272–287. Springer International Publishing, Cham.

Shangbang Long, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Automatic Judgment Prediction via Legal Reading Comprehension. In *Chinese Computational Linguistics*, Lecture Notes in Computer Science, pages 558–572, Cham. Springer International Publishing.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pedro Henrique Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pages 313–323, Cham. Springer International Publishing.

Luyao Ma, Yating Zhang, Tianyi Wang, Xiaozhong Liu, Wei Ye, Changlong Sun, and Shikun Zhang. 2021. Legal Judgment Prediction with Multi-Stage CaseRepresentation Learning in the Real Court Setting. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 993–1002. ArXiv:2107.05192 [cs].

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021a. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021b. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online. Association for Computational Linguistics.

Christoph Molnar. 2022. *Interpretable Machine Learning*, 2 edition.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021a. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021b. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. ArXiv:2209.12325 [cs].

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. ArXiv:2206.10883 [cs].

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal Judgement Prediction for UK Courts. In *Proceedings of the 2020 The 3rd International Conference on Information Science and System*, ICISS 2020, pages 204–209, New York, NY, USA. Association for Computing Machinery.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. ArXiv:1703.01365 [cs].

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them. ArXiv:2210.09261 [cs].

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

550 U.S. at 570 Twombly. 2007. Bell atlantic corp. v. twombly. *Justia*.

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2022. Design and Implementation of German Legal Decision Corpora. pages 515–521.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marco Wrzalik and Dirk Krechel. 2021. GerDaLIR: A German Dataset for Legal Information Retrieval. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 123–128, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiquan Wu, Kun Kuang, Yating Zhang, Xiaozhong Liu, Changlong Sun, Jun Xiao, Yueting Zhuang, Luo Si, and Fei Wu. 2020. De-Biased Court's View Generation with Causality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 763–780, Online. Association for Computational Linguistics.

Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. 2018. Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864, New Orleans, Louisiana. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big Bird: Transformers for Longer Sequences. *arXiv:2007.14062 [cs, stat]*. ArXiv: 2007.14062.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. *arXiv:2104.08671 [cs]*. ArXiv: 2104.08671 version: 3.

Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal Judgment Prediction via Topological Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549, Brussels, Belgium. Association for Computational Linguistics.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively Questioning and Answering for Interpretable Legal Judgment Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1250–1257. Number: 01.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria. INCOMA Ltd.

A Additional Training Details

A.1 Hyperparameter Tuning

We randomly split the data into 70% train, 15% validation and 15% test split. We searched the learning rate in {1e-6, 5e-5, 1e-5} and had the best results with 1e-5. We searched dropout in {0, 0.001, 0.1, 0.2} and finally chose 0. We searched the batch size in {16, 32, 64} and chose 16. Where GPU memory was not sufficient, we used gradient accumulation for a total batch size of 16. We searched the activation function in {Relu, SoftMax, LeakyRelu} and chose SoftMax. We searched weight decay in {0, 0.1} and found 0 to perform best. We used AMP mixed precision training and evaluation to reduce costs. We used early stopping on the validation loss with patience 2. If early stopping was not invoked, we trained for a maximum of 10 epochs. We used an AWS EC2 G5 instance with 4 CPU cores, 16 GB RAM and one NVIDIA A10G GPU (24 GB of GPU memory)

A.2 Preprocessing

We experimented with the following preprocessing methods: (a) removing punctuation; (b) removing numerals; (c) stemming; (d) lemmatization; and (e) entity masking (e.g., "Plaintiff James won would receive 30% from the 3 million compensation fund" → "PERSON won would receive PERCENT from the MONEY compensation fund"). We found that only stemming improved the results.

Method	Max Seq Len	Accuracy
Full Text		
Longformer	2048	$63.64_{\pm 0.72}$
BigBird	2048	$62.00_{\pm 1.08}$
Separated Allegations		
BERT	512	$64.82_{\pm 1.73}$
CaseLawBERT	512	$66.06_{\pm 0.84}$
LegalBERT	512	$64.57_{\pm 1.89}$
LegalRoBERTa	512	$65.41_{\pm 1.09}$

Table 3: Longformer and BigBird used a maximum sequence length of 2,048 tokens. All other models used 512 tokens. For all datasets, we filtered out the rows larger than the maximum sequence length.

A.3 Training Times

On the Unified Allegations dataset, training took approximately one hour for all the investigated models. On the Separated Allegations dataset, it took approximately two hours per model. On the Full Text dataset, it took approximately six hours for Longformer and approximately eight hours for BigBird. All training times are counted for five folds and one random seed on an AWS EC2 G5 instance with 4 CPU cores, 16 GB RAM and one NVIDIA A10G GPU (24GB of GPU memory).

A.4 Library Versions

We used the following libraries and associated versions: python 3.8, transformers 4.17.0, xgboost 1.5.2, torch 1.11.0+cu113, tokenizers 0.12.1, spacy 3.2.3, scikit-learn 1.1.1, pandas 1.3.4, numpy 1.20.3, netcal 1.2.1, nltk 3.6.5, optuna 2.10.1, matplotlib 3.4.3.

B Additional Results

B.1 Filtering the Datasets

In Table 3 we show results for the Filter setup, where we filtered out texts containing more tokens than the maximum sequence lengths of the models used. We note that the results don't change significantly in comparison to Table 2 (Truncation setup).

B.2 XGBoost

Table 4 shows the results for using XGBoost (Chen and Guestrin, 2016) on top of the embeddings instead of simple linear layers as it is reported in Table 2. We observe that this more sophisticated classification layer does not improve results.

Method	Max Seq Len	Accuracy
Full Text		
BERT	512	$60.40_{\pm 0.90}$
LegalBERT	512	$61.79_{\pm 1.13}$
CaseLawBERT	512	$60.65_{\pm 0.32}$
LegalRoBERTa	512	$60.37_{\pm 0.66}$
Longformer	2048	$59.96_{\pm 1.24}$
BigBird	2048	$60.98_{\pm 0.70}$
Unified Allegations		
BERT	512	$62.08_{\pm 0.71}$
LegalBERT	512	$63.01_{\pm 0.60}$
CaseLawBERT	512	$62.22_{\pm 0.59}$
LegalRoBERTa	512	$62.32_{\pm 1.12}$
Longformer	512	$61.7_{\pm 0.82}$
BigBird	512	$61.13_{\pm 1.02}$
Separated Allegations		
BERT	512	$63.19_{\pm 0.49}$
LegalBERT	512	$64.17_{\pm 0.44}$
CaseLawBERT	512	$63.81_{\pm 0.67}$
LegalRoBERTa	512	$64.52_{\pm 0.30}$
Longformer	512	$64.65_{\pm 0.40}$
BigBird	512	$63.38_{\pm 0.31}$

Table 4: We fed the embeddings of the transformer models into an XGBoost (Chen and Guestrin, 2016). For all datasets, we truncated the text to fit the maximum sequence length.

B.3 Calibration Results

Table 5 shows the detailed aggregated ECE scores together with the optimal temperature and the accuracy on the Unified Allegations dataset.

B.4 Human Results

Table 6 shows the results of the human experts on the 200 randomly selected examples.

C Annotation Platform

Figure 6 shows a screenshot of the annotation platform our human experts used.

D Example Complaint

Figures 7 and 8 show an example of a complaint present in the dataset.

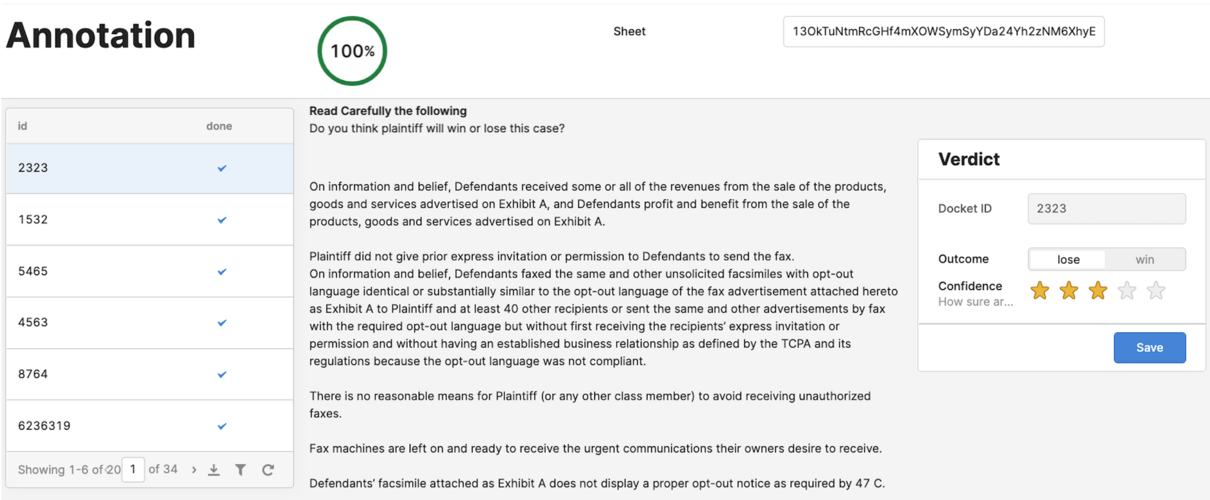


Figure 6: The platform for the human annotations.

Method	Opt. Temp.	ECE Before	ECE After	Accuracy
BERT	$0.19_{\pm 0.03}$	$23.44_{\pm 3.20}$	$5.06_{\pm 1.96}$	$65.06_{\pm 1.67}$
CaseLawBERT	$0.20_{\pm 0.03}$	$25.67_{\pm 2.32}$	$2.59_{\pm 0.90}$	$65.57_{\pm 0.60}$
LegalBERT	$0.22_{\pm 0.02}$	$24.78_{\pm 1.13}$	$3.06_{\pm 1.78}$	$65.87_{\pm 0.26}$
LegalRobertaBase	$0.13_{\pm 0.02}$	$28.02_{\pm 2.16}$	$1.92_{\pm 0.85}$	$65.95_{\pm 0.98}$

Table 5: Calibration results on the Unified Allegations dataset. The text was always truncated to fit the model's maximum sequence length of 512 tokens. Opt. Temp. abbreviates the optimal temperature used for calibrating the models.

	Precision	Recall	F1-score	# Examples
All Results				
lose	49.41	45.65	47.45	92
win	56.52	60.18	58.29	108
accuracy	-	-	53.50	200
High Confidence				
lose	75.00	37.50	50.00	24
win	54.54	85.71	66.66	21
accuracy	-	-	60.00	45

Table 6: Results of the human experts on the 200 randomly selected cases. Under High Confidence we show the results for only the examples where the human experts rated their confidence at 4 or 5 out of 5.

Case: 1:20-cv-00846 Document #: 1 Filed: 02/05/20 Page 1 of 15 PageID #:1

IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF ILLINOIS
EASTERN DIVISION

ANTHONY HALL, on behalf of himself and all others similarly situated,

Plaintiff,

vs.

CLEARVIEW AI, INC., and. CDW GOVERNMENT LLC;

Defendants.

Case No. 20-cv-00846

Jury Demanded

CLASS ACTION COMPLAINT

Plaintiff Anthony Hall, on behalf of himself and a putative class ("Plaintiff" or "Hall"), brings this Class Action Complaint against Defendants Clearview AI, Inc ("Clearview"); CDW Government, LLC ("CDW") and alleges the following:

Introduction

1. A New York Times article published on January 18, 2020 introduced Americans to the then relatively unknown company Clearview AI, Inc. The article described a dystopian surveillance database, owned and operated by a private company and leased to the highest bidder.

2. Clearview AI's database includes the photographs, and personal and private data, including names, home addresses, and work addresses, of millions of Americans. Clearview acquired the billions of data points by "scraping" or harvesting the data from publicly available internet-based platforms such as Facebook, Instagram, and Twitter.

Case: 1:20-cv-00846 Document #: 1 Filed: 02/05/20 Page 2 of 15 PageID #:1

3. But Clearview's database is unique – it has run every one of the 3 billion photographs it has acquired through facial recognition software to extract and index the unique biometric data from each face. The database thus also contains the biometric identifiers and information of millions of Americans. Any private citizen can be identified by uploading a photo to the database. Once identified, the end-user then has access to all of the individual's personal details that Clearview has also obtained.

4. A second article published in the Chicago Sun-Times on January 29, 2020 revealed that the Chicago Police Department was using Clearview's surveillance database to aid in law enforcement operations.

Jurisdiction

5. This Court has jurisdiction under 28 U.S.C. § 1332(d)(2), the Class Action Fairness Act ("CAFA") because there are 100 or more members of the class, the parties and putative class members are minimally diverse and the aggregate amount in controversy is greater than $5,000,000.

6. This Court has personal jurisdiction over Clearview because they conduct a substantial amount of business here which forms the basis of Plaintiffs' claims. Clearview has made their surveillance database, which contains the private and personal data and biometric information of thousands of Illinois residents, available to Chicago Police department. All defendants' violations of Illinois law are based on and arise from their contacts with the state and its residents. The court has personal jurisdiction over CDW because they are an Illinois company headquartered in Illinois.

7. Venue is proper here under 28 U.S.C. § 1391(b)(2) because a substantial amount of the acts and omissions giving rise to the claims occurred in Illinois.

Figure 7: These are the first two pages from an example complaint.

Case: 1:20-cv-00846 Document #: 1 Filed: 02/05/20 Page 13 of 15 PageID #:1

80. Plaintiff and the Class seek:

a. $1,000 for the Plaintiff and each member of the class for each and every separate negligent violation;

b. $5,000 for the Plaintiff and each member of the class for each and every separate intentional or reckless violation;

c. punitive damages;

d. costs, expenses, and reasonable attorneys' fees;

e. and, any other relief this court deems proper.

COUNT III – ILLINOIS CONSUMER FRAUD AND UNFAIR BUSINESS PRACTICES ACT – CLEARVIEW AND CDW

81. At all times relevant, Defendants were engaged in trade or commerce in the state: Clearview and CDW leased, sold, or otherwise provided, for profit, access to the surveillance database to agencies within Illinois such as the CPD.

82. At all times relevant, Plaintiff and members of the class were consumers within the meaning of ICFA.

83. Defendants practice of unauthorized scraping or harvesting of Plaintiff's and the Class members' photos, videos, private and personal information, and its conversion into biometric information and identifiers to add to their surveillance database is an unfair practice.

84. This practice has caused substantial injury and harm to Plaintiff and the members of the Class. It has also forced the Plaintiff to retain counsel to force Clearview to comply with BIPA and redress other violations of state law.

85. Plaintiff and the Class seek:

a. actual damages;

Case: 1:20-cv-00846 Document #: 1 Filed: 02/05/20 Page 14 of 15 PageID #:1

b. punitive damages;

c. costs, expenses, and reasonable attorneys' fees;

d. and, any other relief this court deems proper.

COUNT IV – CONVERSION – CLEARVIEW AND CDW

86. Plaintiff and each Class member have a personal property right in their biometric information and identifiers.

87. Defendants assumed control over the biometric information and identifiers of Plaintiff and the Class with their knowledge or authorization. Defendants' actions impaired Plaintiff and Class members' exclusive right to control their property.

88. Plaintiff and the Class seek:

a. the greater of actual damages or the profits gained by CDW and Clearview from the conversion of Plaintiff and Class members property;

b. punitive damages;

c. and, any other relief this court deems proper.

Jury Demand

Plaintiff demands a trial by jury.

February 5, 2020

[Signature Page Follows]

Figure 8: These are the last two pages from an example complaint.

Can we Pretrain a SotA Legal Language Model on a Budget From Scratch?

Joel Niklaus and **Daniele Giofré**
Thomson Reuters Labs, Zug, Switzerland
firstname.lastname@thomsonreuters.com

Abstract

Even though many efficient transformers have been proposed, only few such models are available for specialized domains. Additionally, since the pretraining process is extremely costly in general – but even more so as the sequence length increases – it is often only in reach of large research labs. One way of making pretraining cheaper is the Replaced Token Detection (RTD) task, by providing more signal during training compared to MLM, since the loss can be computed over all tokens. In this work, we train Longformer models with the efficient RTD task on long-context legal data to showcase that pretraining efficient LMs is possible using less than 12 GPU days. We evaluate the trained models on challenging summarization tasks requiring the model to summarize complex long texts. We find that both the small and base models outperform their baselines on the in-domain BillSum and out-of-domain PubMed tasks in their respective parameter range. We publish our models as a resource for researchers and practitioners.

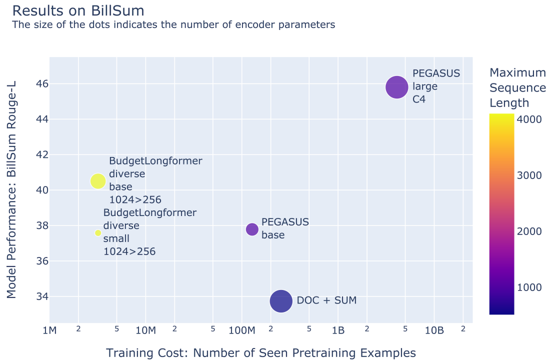


Figure 1: Results on BillSum (log-scaled x-axis)

1 Introduction

Pretrained transformer models have achieved excellent performance across various Natural Language Processing (NLP) tasks such as Text Classification (TC), Named Entity Recognition (NER), Question Answering (QA) and summarization (Devlin et al., 2019; Yang et al., 2020; He et al., 2021; Zhang et al., 2020a).

Transfer learning is to a large extent responsible for this success (Howard and Ruder, 2018). Usually, transformer models are pretrained in a self-supervised way on large unlabeled corpora (Devlin et al., 2019; Radford et al., 2018). Pretraining is very resource intensive (especially for large models), thus making it costly and only available for large organizations (Sharir et al., 2020). The Masked Language Modeling (MLM) task has been very successful, with many models adopting the task in pretraining (Devlin et al., 2019; Liu et al., 2019; Beltagy et al., 2020; Zaheer et al., 2021). Since typically only 15% of the tokens are masked, the loss can be computed for those tokens only.

Clark et al. (2020) introduced the Replaced Token Detection (RTD) task, enabling loss computation on all tokens for efficient training. On the GLUE benchmark (Wang et al., 2018), ELECTRA matches RoBERTa (Liu et al., 2019) and XLNet (Yang et al., 2020) using 1/4 their compute. Although ELECTRA's training strategy seems very promising, to the best of our knowledge, only few works have adopted the RTD task so far (He et al., 2021; Kanakarajan et al., 2021).

On another note, domain-specific pretraining has been shown to improve downstream performance in many domains such as law (Chalkidis et al., 2020; Xiao et al., 2021), biology (Lee et al., 2019), scientific articles (Beltagy et al., 2019), clinical documents (Li et al., 2022), or even code (Chen et al., 2021a). Despite the vast amount of legal text and the importance of training legal models for downstream tasks, there has yet to be domain-specific pertaining coupled with the RTD task for law.

Pretraining on legal documents is especially challenging, given that legal documents tend to span multiple pages (ranging from 10s to 100s of pages, which translates to tens of thousands tokens). This

Proceedings of The Fourth Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 158–182
July 13, 2023 ©2023 Association for Computational Linguistics

is incompatible with current transformer architectures (Vaswani et al., 2017) as they often prohibit efficient processing of sequences longer than 512 tokens on current hardware due to the quadratic time and memory requirement of the attention mechanism. To solve this problem, a rich body of research investigates how transformers can be adapted to efficiently process longer input (Tay et al., 2020b; Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2021; Roy et al., 2021; Kitaev et al., 2020; Tay et al., 2021; Lee-Thorp et al., 2021).

Longformer (Beltagy et al., 2020) is one of these efficient transformer architectures for long sequences, leveraging windowed and global attention. So far, to the best of our knowledge, there does not yet exist a public Longformer model pretrained on English legal data[1], although Xiao et al. (2021) have proven the effectiveness of the Longformer in dealing with long legal text in many Chinese-related tasks. This work aims to fill this gap.

To test the ability to grasp long-distance dependencies in the text, we mainly evaluated our models on the task of automatic (abstractive) summarization. It consists of capturing the most important concepts/ideas from the (long) document and then rewriting it in a shorter passage in a grammatical and logically coherent way (Chen et al., 2019).

In particular, we used the BillSum dataset (Kornilova and Eidelman, 2019), as a domain-specific summarization task, and the PubMed dataset (Cohan et al., 2018), to evaluate the model's ability outside the legal context (i.e., in the biomedical context). On BillSum, we achieve a new state-of-the-art (SOTA) (see Figure 1) in our parameter range. On Pubmed, we obtain comparable metrics even though the Language Model (LM) has only been pretrained on legal data and the tokenizer is also optimized for legal data (see Figure 2).

We emphasize that this performance was achieved with minimal pretraining due to the combination of the RTD task and the Longformer infrastructure making our LM very attractive from the perspective of building costs. For example, our model saw 3.2M examples during pretraining, while RoBERTa (Liu et al., 2019) or PEGASUS-large (Zhang et al., 2020a) saw 4.1B examples (nearly 1300x more). For reference, RoBERTa was trained for 1024 GPU days (>42x more than our base model), while our small and base models only used 12 and 24 GPU days respectively (16GB NVIDIA V100 GPUs for all models).[2]

Contributions

The contributions of this paper are three-fold:

- We train and release a new model pretrained on recently published curated English legal text (Henderson et al., 2022), capable of handling input spans longer than 512 tokens out of the box.
- Using Longformer and RTD, dubbed Budget-Longformer, we achieve a new SOTA on BillSum and PubMed compared to models of the same size. Our small model even outperforms a transformer base model (Vaswani et al., 2017) containing almost 4 times more encoder parameters (110M vs. 29M). On BillSum it performs on par with a PEGASUS base model (Zhang et al., 2020a) whose encoder is also almost 4 times larger and has been pretrained specifically for the abstractive summarization task in mind.
- We verified that pretraining with the RTD task is suitable for down-stream summarization tasks by evaluating our model on an out-of-domain benchmark (PubMed), obtaining comparable results with summarization-specific architectures.

Main Research Questions

In this work, we pose and examine three main research questions:
RQ1: *Is it possible to train a LM with domain (e.g. legal) expertise efficiently from scratch, reducing costs?*
RQ2: *How does our model compare with other models on the challenging legal domain-specific BillSum summarization benchmark?*
RQ3: *How well does our model compare with other models on the biomedical out-of-domain PubMed summarization benchmark?*

2 Related Work

Domain-Specific Language Models

Previous work showed that domain-specific pretraining achieves promising results on datasets of specialized domains such as law (Chalkidis et al., 2020; Xiao et al., 2021), biology (Lee et al., 2019), scientific articles (Beltagy et al., 2019), clinical

[1]On the web there is a model based on Longformer in the legal domain, but it offers no model card (https://huggingface.co/saibo/legal-longformer-base-4096). Also, concurrent to our work, Mamakas et al. (2022) trained legal Longformer models, but they are private. Additionally, concurrently, Hua et al. (2022) trained Reformer (Kitaev et al., 2020) models with the RTD task on legal data.

[2]Although Zhang et al. (2020a) do not report the compute used, we expect it to be similar to RoBERTa.

documents (Li et al., 2022), or even code (Chen et al., 2021a).

Gururangan et al. (2020) show that continued pretraining on a RoBERTa checkpoint on biomedical data, scientific articles in computer science, and reviews, clearly improves downstream performance in the respective domain-specific datasets. The effect was less pronounced on news domain datasets, presumably because RoBERTa has seen many news articles during pretraining already.

Long Document Processing

In the past few years, a vast amount of research has been devoted to addressing the problem of quadratic time and memory complexity associated with the dense attention mechanism (Vaswani et al., 2017), practically limiting the maximum sequence length severely (often to 512 tokens) (Tay et al., 2020b; Child et al., 2019; Beltagy et al., 2020; Zaheer et al., 2021; Roy et al., 2021; Kitaev et al., 2020; Tay et al., 2021; Lee-Thorp et al., 2021). These research works have given rise to a new class of transformers, referred to as sparse transformers or efficient transformers (Tay et al., 2020b). Reducing the cost associated with the computation of the dense attention matrix while maintaining the same performance is the core idea behind efficient transformers. This is often achieved by introducing sparsity in the attention matrix in a variety of ways that may be fixed pattern such as local (windowed) attention (Child et al., 2019; Beltagy et al., 2020), global attention (Zaheer et al., 2021) or learnable patterns such as routing attention (Roy et al., 2021) and LSH attention (Kitaev et al., 2020) or a random pattern (Zaheer et al., 2021; Tay et al., 2021). Recently, Lee-Thorp et al. (2021) proposed to use Fourier transforms instead of the attention layer. Tay et al. (2020b) provide a comprehensive list of efficient transformers and the detailed description of their attention mechanism. (Tay et al., 2020a) proposed a series of tasks designed for testing the capabilities of these different models suitable for longer inputs. However, this so-called "Long Range Arena" considers mostly artificial tasks, with the goal of evaluating the models independent of any pretraining.

Efficient Pretraining

ELECTRA-style pretraining (Clark et al., 2020) has been shown to reduce training cost substantially, while matching the performance of SOTA LMs. ELECTRA leverages a smaller generator model (discarded after pretraining), that changes some tokens. The larger discriminator model (used for down-stream tasks) must predict for each token if it was changed by the generator or not, similar to how Generative Adversarial Networks (GANs) are trained (Goodfellow et al., 2014). This enables the loss to be relevant for every token, leading to much faster and thus more efficient training.

3 Datasets

3.1 Pile of Law

Henderson et al. (2022) recently released a large-scale English corpus suitable for pretraining LMs. It contains 256 GB of diverse legal text in English from various jurisdictions and judicial bodies including for example bills, court decisions and contracts from the US, Canada, and Europe even though the focus clearly lies on US data. While there are 28 US datasets available (253.25 GB or 99%), there is only 1 Canadian dataset[3] (243 MB or 0.09%), 3 European datasets[4] (2.3 GB or 0.9%), and 2 international datasets[5] (212 MB or 0.08%). The non-US datasets only cover the categories "Legal Case Opinions and Filings", "Laws" and "Conversations", but do not cover categories "Legal Analyses", "Contracts / Business Documents" and "Study Materials", whereas the US data is much more diverse and covers all categories.

3.2 BillSum

Kornilova and Eidelman (2019) introduced a legislative summarization dataset covering 21K US bills from 1993 to 2018. It is challenging due to the technical nature and complex structure of the bills. Additionally, the bills are rather long, ranging from 5K to 20K characters ($\sim$ 1K to 4K tokens[6]) with their summaries being up to 5K characters ($\sim$ 1K tokens) long (see Appendix C for more details).

3.3 PubMed

Cohan et al. (2018) introduced another challenging summarization dataset in a specialized domain (scientific articles from the biomedical domain). It includes 133K scientific papers together with their abstracts in English. The papers are 3K words long on average and the summaries (abstracts) 200

[3] Canadian Court Opinions (ON, BC)

[4] European Court of Human Rights Opinions, EUR-LEX and European Parliament Proceedings Parallel Corpus

[5] World Constitutions and U.N. General Debate Corpus

[6] Our experiments show that using our tokenizer, one token corresponds to 5.33 characters on average.

words. Thus, similar to the BillSum dataset, this dataset is well suited as a test bed for methods capable of long document summarization. Note, that in this dataset, the domain is vastly different from the legal domain (see Appendix C for more details).

4 BudgetLongformer

In the legal domain, it is especially important that models can handle long input. So far, there does not exist an English legal model capable of handling more than 512 tokens. Since many tasks in legal NLP are formulated as TC problems, a hierarchical architecture has been used frequently to process long documents (Chalkidis et al., 2019; Niklaus et al., 2021, 2022, 2023). This simple hierarchical architecture, however, cannot be easily adapted to solve the more complex sequence-to-sequence tasks like token classification or summarization because it compresses the long input sequence into a single token. For this reason, in this work, we pretrain a more versatile Longformer model. To make pretraining more affordable, we trained the well-proven Longformer model (Beltagy et al., 2020) with the RTD task proposed by Clark et al. (2020).

4.1 Longformer

We opted for the Longformer method over other efficient transformer architectures because it seems to work robustly[7] and is heavily used in the literature (Xiao et al., 2021; Dai et al., 2022; Maroudas et al., 2022). Longformer (Beltagy et al., 2020) proposed three sparse attention mechanisms: Sliding Window Attention, Dilated Sliding Window Attention and Global + Sliding Window. We follow their recommendations and use the Global + Sliding Window attention mechanism because we pretrain an encoder-only model.

4.2 Replaced Token Detection

Inspired by GAN training (Goodfellow et al., 2014), the RTD task adapts this training framework to NLP. The drawback of training with MLM is that the loss can only be computed for the masked tokens (usually 15%). With RTD training, a smaller generator model (usually 1/3 the size of the discriminator) solves the MLM task. The discriminator receives the predictions of the generator and determines for each token, whether it is original or changed by the generator. This leads to the loss being computed for each token for the discriminator, thus transporting more information per forward-pass and leading to more efficient training.

[7] 164 models on huggingface hub as of January 3rd 2023

PileOfLaw Subset	Dataset Size	# Words	# Documents
caselaw			
CL Opinions	**59.29GB**	**7.65B**	**3.39M**
diverse			
Total	**73.04GB**	**8.91B**	**2.1M**
CL Opinions	8.74GB	1.13B	500K
CL Docket Entries	17.49GB	1.80B	500K
U.S. State Codes	6.77GB	829.62M	157
U.S. Code	0.27GB	30.54M	43
EUR-Lex	1.31GB	191.65M	106K
Edgar Contracts	7.26GB	0.97B	500K
Atticus Contracts	31.2GB	3.96B	488K

Table 1: The datasets used for pretraining our models. CL is short for Court Listener

5 Experimental Setup

In this section, we describe how we set up the experiments. For all experiments, we used the huggingface transformers library (Wolf et al., 2020) available under an Apache 2.0 license and AMP mixed precision training and evaluation to reduce costs and GPU memory.

5.1 Tokenizer

We trained a byte-level BPE tokenizer (Wang et al., 2019) akin to Beltagy et al. (2020) with a large 64K token vocabulary to encode complex legal language well. We trained the tokenizer using the huggingface tokenizers library[8] on the entire PileOfLaw training split ($\sim$ 192GB, $\sim$ 22.5B tokens, $\sim$ 7.5M documents), covering a wide array of English (mostly US) legal texts without preprocessing/cleaning due to the high-quality data.

5.2 Pretraining

Henderson et al. (2022) have experienced difficulties when the language model was trained on the entire PileOfLaw. We believe that the highly imbalanced dataset concerning text types (contracts, court decisions, legislation, etc.) could have been a reason for the training instability.[9] This led us to do a sanity check by training only on caselaw first and then to subselect only the most important and largest subsets of the PileOfLaw for training the diverse model, leading to stable pretraining (see Section 6). On the contrary, on the summarization tasks, the diverse model – which includes more

[8] https://github.com/huggingface/tokenizers

[9] However, the large model size could also explain the training instability.

lexical and layout diversity of documents – turns out to perform better and train more robustly.

We trained the *caselaw* models on the training subset of "Court Listener Opinions" from the PileOfLaw (59.3 GB, 7.65B words, 3.39M documents). The *diverse* models were trained on caselaw ("Court Listener Opinions" & "Court Listener Docket Entries"), legislation ("US Code", "State Codes" & "EURLEX") and contracts ("Atticus Contracts" & "EDGAR Contracts"). To balance the training data, we limited the number of documents to 500K (this affects Court Listener Opinions, Court Listener Docket Entries and EDGAR Contracts (see Table 1 for more details). Our validation set consisted of 1000 randomly selected examples from the respective training set.[10] To maximally use the available data, we concatenated all the examples and cut them off in slices of the model's maximum sequence length (4096) – in batches of 1000 examples with multiprocessing to speed up data preparation. We dropped the last slice, since it will not contain 4096 tokens.

We trained both a small (29M parameters) and a base (159M parameters) model for each configuration (caselaw and diverse data). To reach 100K steps it took 68 hours (a bit less than 3 days) for the small model and 135 hours (a bit more than 5 days) for the base model on 4 16GB NVIDIA V100 GPUs. The achieved training and evaluation losses are shown in Table 7 in Appendix A. Interestingly, we find that the diverse models achieve lower training and evaluation losses. Please find more training details in Appendix A. Due to budget constraints, we trained for a maximum of 200K steps. Surprisingly, lower pretraining loss from 200K-step models did not transfer to downstream tasks. We hypothesize that a larger batch size might lead to improvements when training longer.

5.3 Downstream Benchmarks

For downstream finetuning, we paired our pretrained encoder model with a randomly initialized BART-base decoder model (Lewis et al., 2020).[11] For BillSum, we set the maximum input length to 1024 and the maximum target length to 256 to save compute. However, many summaries get cut off at 256 tokens. This is why we took our best model and trained it with maximum input length 4096 and maximum target length 1024 (see results in Table 5 and examples in Table 12). For PubMed, we set the maximum input length to 4096 and the maximum generation length to 512. Due to high training costs, we only trained our models with one random seed (42). Our models contain 29M (small) and 159M (base) parameters in the encoder and 96M parameters in the decoder, resulting in a total of 125M (small) and 255M (base) parameters.

5.4 Ablation Studies

We run two ablation studies on the BillSum dataset, testing the influence of the pretraining corpus and the number of pretraining steps. To reduce computational costs, we set the maximum input and generation lengths to 1024 and 128 respectively.

# Steps	Size	Rouge-1 ↑	Rouge-2 ↑	Rouge-L ↑
100K	small	51.62	30.84	40.22
200K	small	49.02	27.02	36.98
100K	base	56.10	36.50	45.17
200K	base	55.30	35.47	44.30

Table 2: Models pretrained on caselaw only.

Corpus	Size	Rouge-1 ↑	Rouge-2 ↑	Rouge-L ↑
caselaw	small	51.62	30.84	40.22
diverse	small	53.61	33.54	42.50
caselaw	base	56.10	36.50	45.17
diverse	base	54.87	35.63	44.21

Table 3: Models pretrained for 100K steps.

Pretraining Steps

Though train and evaluation losses decrease steadily with more pretraining steps (see Table 7), surprisingly, models trained longer underperform on the BillSum benchmark (see Table 2). We hypothesize the low pretraining batch size caused fast convergence to a local optimum, inhibiting further progress. Consequently, we use the 100K steps model checkpoints.

Pretraining Corpus

In total, we trained 4 models (small and base each on the caselaw and diverse corpora). In Table 3 we perform an ablation on the pretraining corpus. The results are inconclusive, with the diverse corpus outperforming for the small models and the caselaw corpus outperforming for the base models. The caselaw models were unstable during finetuning and even failed completely for some learning

[10] We used such a small validation set to save compute.

[11] Interestingly, the randomly initialized decoder yielded better results than when we used the weights from the pretrained huggingface checkpoint at https://huggingface.co/facebook/bart-base.

rates. Together with the fact that the diverse models reached lower pretraining losses (see Table 7), we focus on the diverse models for our experiments.

We acknowledge the necessity of more ablations. Because of limited compute, we opted for the safest and cheapest choices instead of ablating them (e.g. windowed and global attention, RTD pretraining task). Additionally, we put a focus on providing our models as a resource for further research in this area and for practitioners in the field of legal NLP. We thus leave further ablations for future work (w.r.t. pretraining task, more general domain corpora, efficient transformer method, etc.).

6 Results

In this section, we present results for the BillSum and PubMed datasets, conducting error analysis on generated summaries. Table 4 compares models in detail. All further experiments utilize models trained on the diverse dataset.

6.1 BillSum

We achieve a new SOTA on BillSum in the small and base parameter range and outperform models with almost 12 times more encoder parameters and others having seen more than 1200 times more pretraining examples. The results on BillSum are presented in Figure 1 and Table 5.

We observe that even our small diverse model clearly exceeds the baseline of the original article (DOC + SUM), even though their model is based on BERT-large, containing almost 12 times more encoder parameters and pretrained for 10x more steps. Even more surprisingly, our small diverse model is on par with the PEGASUS-base model (Zhang et al., 2020a) (37.58 vs. 37.78 Rouge-L), pretrained using the Gap-Sentences task specifically designed for abstractive summarization. PEGASUS-base contains almost 4 times more encoder parameters and has seen 40 times more training examples during pretraining (128M vs. 3.2M; see Table 4). Most surprisingly, it even outperforms an LED large model[12] (37.58 vs. 34.23 Rouge-L) using a much longer input length (16384 vs. 1024), containing more than 8 times as many encoder parameters (257M vs. 29M) and having seen more than 1200 times more examples during pretraining.

By scaling up our model to the base size and increasing the maximum input and generation length

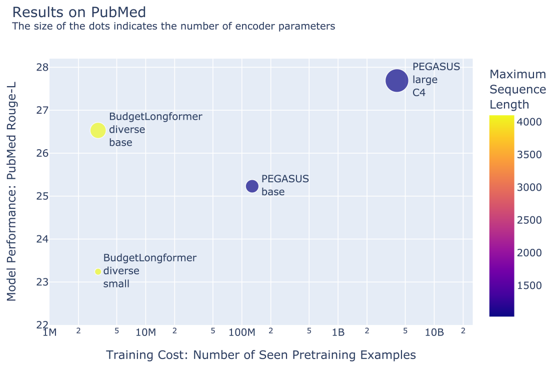


Figure 2: Results on PubMed (log-scaled x-axis)

to 4096 and 1024 tokens respectively, we even approach the performance of PEGASUS-large (43.23 vs. 45.8 Rouge-L). PEGASUS-large has seen three orders of magnitude more training examples during its pretraining in comparison to our model (4.1B vs. 3.2M) and contains almost twice as many encoder parameters (301M vs. 159M).

To conclude, it appears that pretraining with the RTD on (high-quality) in-domain data can be an effective and computationally cheap alternative to a summarization-specific model trained on web text (i.e. PEGASUS). Whether the gain is due to in-domain pretraining or the RTD task is inconclusive, and we leave these experiments for future work.

6.2 PubMed

We achieve a new SOTA on PubMed in the small and base parameter range and almost reach the performance of a PEGASUS large model pretrained with a summarization-specific task. The results on PubMed are presented in Figure 2 and Table 6.

Similar to the results on BillSum, our small model clearly outperforms the transformer-base model (23.24 vs. 19.02 Rouge-L) and approaches the PEGASUS-base model (23.24 vs. 25.2 Rouge-L) despite not being specifically pretrained for summarization and having seen significantly fewer examples during pretraining (3.2M vs. 128M). Similar again, our base model outperforms PEGASUS-base (26.53 vs. 25.23 Rouge-L) and almost reaches the performance of PEGASUS-large (26.53 vs. 27.69 Rouge-L) while having seen 1280 times fewer examples during pretraining (3.2M vs. 4.1B).

Our model is pretrained on the narrower domain of legal text, rather than broader C4 data used by PEGASUS. Furthermore, our model and tokenizer had no exposure to medical data in pretraining. This, combined with the high quality of legal data used in pretraining, may explain our model's good out-of-domain performance, similar to the findings

[12] https://huggingface.co/Artifact-AI/led_large_16384_billsum_summarization

Model Name	Source	P. Steps	P. BS	# P. Examples	# Enc. Par	# Dec. Par	MaxSeqLen	Vocab
DOC + SUM	(Kornilova and Eidelman, 2019)	1000K	256	256M	340M	–	512	30K
Transformer base	(Zhang et al., 2020a)	–	–	–	159M	187M	1024	96K
PEGASUS base	(Zhang et al., 2020a)	500K	256	128M	159M	187M	1024	96K
PEGASUS large (C4)	(Zhang et al., 2020a)	500K	8192	4096M	301M	368M	1024	96K
LED large	(Beltagy et al., 2020)	500K	8192	4096M	257M	254M	16384	50K
LongT5 xl	(Guo et al., 2022)	1000K	2048	2048M	1224M	1626M	16384	32K
BudgetLongformer small	ours	100K	32	3.2M	29M	96M	4096	64K
BudgetLongformer base	ours	100K	32	3.2M	159M	96M	4096	64K

Table 4: Comparison of the evaluated models. For more information on the baselines, refer to the cited papers. (Abbreviations: P.: Pretraining, BS: Batch Size, Enc.: Encoder, Dec.: Decoder, Par: Parameters.)

Model	Size	MaxInLen	MaxGenLen	Rouge-1 ↑	Rouge-2 ↑	Rouge-L ↑
BudgetLongformer	small	1024	256	**49.85**	**29.63**	**37.58**
Transformer	base	512	256	44.05	21.30	30.98
PEGASUS	base	512	256	51.42	29.68	37.78
BudgetLongformer	base	1024	256	52.70	32.97	40.50
BudgetLongformer	base	4096	1024	**55.45**	**36.68**	**43.23**
DOC + SUM	large	512	512	40.80	23.83	33.73
PEGASUS (C4)	large	1024	256	**57.20**	**39.56**	**45.80**
LED	large	16384	1024	47.84	26.34	34.23

Table 5: Results on BillSum. Best results per model size are in bold.

of Taylor et al. (2022). Even though our pretraining data is out-of-domain PubMed – whereas C4, likely contains medical data – compared to PEGASUS, our models perform similarly on PubMed as on BillSum. This makes us believe the gains stem mainly from the RTD pretraining task.

Krishna et al. (2022) find that pretraining on the downstream corpus can achieve similar results as pretraining on a large upstream corpus, significantly cutting costs. Finetuning a small model on BillSum cost us approx. half-day of a 16GB V100 GPU. Pretraining the small model for 100K steps cost approx. 12 GPU days[13]. Pretraining and finetuning a smaller model with the RTD task on a task specific corpus might be a suitable alternative to finetuning a larger general model, yielding similar performance with shorter inference time and costs.

6.3 Error Analysis

We conducted an error analysis by manually inspecting 25 random summaries. Example summaries are shown in Appendix D.

Coherence The inspected summaries were well-structured and emulated the specific style of the reference summaries in the respective domains[14].

[13]For the base model the numbers are approx. double

[14]In future work, we will corroborate these findings by performing human evaluations with domain experts.

Consistency We find the summaries mostly factually aligned with the source.[15] However, sometimes it copies formulas from the source text, but then mixes up numbers.[16]

Fluency Generally, we find the summaries to be fluent[17] and grammatically correct.

Relevance In general, the model summaries contain important content from the source document. However, we find repetitions to be a repeating issue in both BillSum and PubMed summarization. In the BillSum task, the model occasionally uses the same start of the sentence multiple times instead of providing a longer list [18]. It correctly imitates the lists often given in BillSum summaries, but then seems to struggle with continuing lists to more entries. Other times it manages well to formally continue the lists, but repeats list items. In the PubMed task, in one particular summary, a phrase

[15]e.g. "imaging guidance improved the accuracy of intra - articular injections of the knee (96.7% versus 81.0%, p < 0.001) and shoulder (97.3% versus 65.4%, p < 0.001)"

[16]e.g. "[a1c (%) = [0.021 mbg (mg / dl] + 4.3, r = 0.92) + 4.3, r = 0.58]"

[17]Repetitions are discussed in "Relevance"

[18]e.g. "**amends the agricultural marketing act of 1946 to** terminate the authority of the secretary of agriculture (usda) to: (1) livestock processing plant processing plant slaughter, and (2) slaughtering plant slaughter. **amends the agricultural marketing act of 1946 to**: (1) revise minimum reporting requirements; and (2) revise reporting requirements"

Model	Size	MaxInLen	MaxGenLen	Rouge-1 ↑	Rouge-2 ↑	Rouge-L ↑
BudgetLongformer	small	4096	512	**34.98**	**13.56**	**23.24**
Transformer	base	512	256	33.94	7.43	19.02
PEGASUS	base	512	256	39.98	15.15	25.23
BudgetLongformer	base	4096	512	**41.16**	**18.15**	**26.53**
PEGASUS (C4)	large	1024	256	**45.49**	**19.90**	**27.69**
LongT5	xl	16384	512	**50.23**	**24.76**	**46.67**

Table 6: Results on PubMed. Best results per model size are in bold.

gets repeated 10 times. Even in a high scoring example (Rouge1: 62.2, RougeL: 48.5, 464 tokens summary length), a sentence is repeated three times. Here, in contrast to BillSum, the repetitions are also occurring on a lower level.[19]

Generally, the problems are similar in the BillSum and the PubMed tasks; however, they are less pronounced in the in-domain BillSum dataset.

7 Conclusions and Future Work

7.1 Answers to Main Research Questions

RQ1: *Is it possible to train a LM with domain (e.g. legal) expertise efficiently from scratch, reducing costs?* Yes, this work demonstrates the feasibility of pretraining a domain-expertise LM from scratch with minimal compute, matching performance of methods exposed to three orders of magnitude more pretraining examples. Particularly when a high-performing large teacher model is unavailable, our method is advisable.

RQ2: *How does our model compare with other models on the challenging legal domain-specific BillSum summarization benchmark?* Our LMs compare favorably to baselines on the challenging domain-specific summarization benchmark BillSum, necessitating long input processing. Our small model outperforms the larger PEGASUS-base, and our base model almost reaches the performance of the larger PEGASUS-large. Both baselines have been pretrained with much more compute and data, and additionally with a pretraining task crafted specifically for summarization.

RQ3: *How well does our model compare with other models on the biomedical out-of-domain PubMed summarization benchmark?* Our results on the out-of-domain PubMed summarization benchmark show that our models compare favorably to baselines. Again, our small model outperforms PEGASUS-base and our base model approaches PEGASUS large.

7.2 Conclusion

In this work, we show that we can successfully pretrain Longformer models with the RTD task on a Budget. Using very little pretraining, we can achieve SOTA performance on the challenging legal summarization task BillSum, outperforming PEGASUS, that has been pretrained specifically for summarization. Our model even outperforms PEGASUS on the out-of-domain PubMed dataset involving biomedical research articles. To sum up, we present a simple and extremely cheap way of pretraining a long-context LM in cases without the availability of a large teacher model.

7.3 Future Work

Future work could test our models on further legal downstream benchmarks such as LexGLUE (Chalkidis et al., 2021), ClassActionPrediction (Semo et al., 2022), CUAD (Hendrycks et al., 2021) or MultiLexSum (Shen et al., 2022). Additionally, one can test whether the out-of-domain results hold on other out-of-domain summarization datasets, such as BigPatent (Sharma et al., 2019) or ArXiv (Cohan et al., 2018). Future work could further scale up the models in terms of batch size, pretraining steps, parameter count and data size to test what further gains can be achieved. Additionally, to further save compute and enhance models, one could explore warm-starting ELECTRA pretraining from existing checkpoints.. The difficulty, of course, lies in getting a suitable generator and discriminator, trained with the same tokenizer. One possible setup might be Longformer-base as the generator and Longformer-large as the discriminator. Finally, one can investigate the use of other efficient transformers with the RTD task.

[19] e.g. "hemoglobin glycated hemoglobin (hba1c)"

Limitations

ELECTRA-style training has the disadvantage of the setup being slightly more complicated, requiring a generator and a discriminator. Additionally, the generator should be smaller than the discriminator to ensure stable training. This makes it difficult to warm start from available checkpoints, since two models of different sizes are required. Often, small models are not released, which makes it difficult to warm-start base models using the RTD task. We leave the direction of warm starting a large discriminator with a base generator to future work.

Except for EUR-LEX (1.31 GB or 1.8% of our diverse dataset), our models have only seen US data during the pretraining phase. So, while these models are expected to work well on US data or datasets with similar content such as heavily influenced by the US or mainly common-law based, legal data from Europe for example is expected to look very different (mainly civil-law based except for the UK) and often translated from the original European languages. Thus, our models are not expected to transfer well to such kind of data.

Because of insufficient compute, we were not able to scale up our models in terms of parameter size, batch size and number of pretraining steps. So while we can show that our approach scales well from the small to the base model, it is unknown if this continues to even larger model sizes. Although it is expected to produce better results, we do not know if using a higher batch size and more pretraining steps boosts performance significantly. Additionally, the lacking compute budget made evaluating on more and especially large datasets like BigPatent impossible. Therefore, we cannot give any conclusions at this point to whether our results are robust across a wide range of datasets.

So far, we did not evaluate our summarization models using newer reference-based metrics such as BERTScore (Zhang et al., 2020b) or BARTScore (Yuan et al., 2021), or reference-free metrics such as SUPERT (Gao et al., 2020) or Semantic Distribution Correlation (SDC) (Chen et al., 2021b). However, our baselines used ROUGE only, requiring us to rerun experiments for comparison using newer scores, straining our low compute budget.

So far, we did not have the resources to conduct a thorough human expert evaluation of the quality of our summarization outputs. Such an evaluation would be needed for production systems and for better comparison of models. However, it also requires highly educated medical experts (for PubMed) or lawyers with specific expertise in US bills (for BillSum) respectively, and thus a prohibitively high amount of resources.

For comparing the efficiency of pretraining, number of FLOPs would probably be best. We compared the models' efficiency based on the number of seen examples during pretraining, due to ready availability (most papers report batch size and number of steps, but few papers report FLOPs).

Ethics Statement

Pretraining language models is a very compute-heavy process and thus leaves a large carbon footprint (Strubell et al., 2019; Patterson et al., 2021). Our method makes significantly reduces the compute requirements and thus the carbon footprint.

As with any large LM there is the risk of it producing biased or unfair output. Researchers using the model should put into place respective safeguards to identify biased and/or toxic language.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv:1903.10676 [cs]*. ArXiv: 1903.10676.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*. ArXiv: 2004.05150.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559 [cs]*. ArXiv: 2010.02559.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael James Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. SSRN Scholarly Paper ID 3936759, Social Science Research Network, Rochester, NY.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021a. Evaluating Large Language Models Trained on Code. *arXiv:2107.03374 [cs]*. ArXiv: 2107.03374.

Wang Chen, Piji Li, and Irwin King. 2021b. A Training-free and Reference-free Summarization Evaluation Metric via Centrality-weighted Relevance and Self-referenced Redundancy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 404–414, Online. Association for Computational Linguistics.

Yangbin Chen, Yun Ma, Xudong Mao, and Qing Li. 2019. Multi-Task Learning for Abstractive and Extractive Summarization. *Data Science and Engineering*, 4(1):14–23.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. *arXiv:1904.10509 [cs, stat]*. ArXiv: 1904.10509.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*. ArXiv: 2003.10555.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting Transformer-based Models for Long Document Classification. *arXiv:2204.06683 [cs]*. ArXiv: 2204.06683.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Yang Gao, Wei Zhao, and Steffen Eger. 2020. SUPERT: Towards New Frontiers in Unsupervised Evaluation Metrics for Multi-Document Summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1347–1354, Online. Association for Computational Linguistics.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. ArXiv:1406.2661 [cs, stat].

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. LongT5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv:2004.10964 [cs]*. ArXiv: 2004.10964.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543 [cs]*. ArXiv: 2111.09543.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. ArXiv:2207.00220 [cs].

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. 2021. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. ArXiv:2103.06268 [cs].

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber. 2022. LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension. ArXiv:2212.08204 [cs].

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:Pretrained Biomedical text Encoder using Discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. *arXiv:2001.04451 [cs, stat]*. ArXiv: 2001.04451.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A Corpus for Automatic Summarization of US Legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Kundan Krishna, Saurabh Garg, Jeffrey P. Bigham, and Zachary C. Lipton. 2022. Downstream Datasets Make Surprisingly Good Pretraining Corpora. ArXiv:2209.14389 [cs].

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, page btz682. ArXiv: 1901.08746.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. FNet: Mixing Tokens with Fourier Transforms. *arXiv:2105.03824 [cs]*. ArXiv: 2105.03824.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yikuan Li, Ramsey M. Wehbe, Faraz S. Ahmad, Hanyin Wang, and Yuan Luo. 2022. Clinical-Longformer and Clinical-BigBird: Transformers for long clinical sequences. *arXiv:2201.11838 [cs]*. ArXiv: 2201.11838.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing Long Legal Documents with Pre-trained Transformers: Modding LegalBERT and Longformer. ArXiv:2211.00974 [cs].

Stelios Maroudas, Sotiris Legkas, Prodromos Malakasiotis, and Ilias Chalkidis. 2022. Legal-Tech Open Diaries: Lesson learned on how to develop and deploy light-weight models in the era of humongous Language Models. ArXiv:2210.13086 [cs].

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. ArXiv:2301.13126 [cs].

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon Emissions and Large Neural Network Training. *arXiv:2104.10350 [cs]*. ArXiv: 2104.10350.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. page 12.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient Content-Based Sparse Attention with Routing Transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68. Place: Cambridge, MA Publisher: MIT Press.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Or Sharir, Barak Peleg, and Yoav Shoham. 2020. The Cost of Training NLP Models: A Concise Overview. ArXiv:2004.08900 [cs].

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities. ArXiv:2206.10883 [cs].

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. ArXiv:1906.02243 [cs].

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2021. Synthesizer: Rethinking Self-Attention in Transformer Models. *arXiv:2005.00743 [cs]*. ArXiv: 2005.00743.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020a. Long Range Arena: A Benchmark for Efficient Transformers. *arXiv:2011.04006 [cs]*. ArXiv: 2011.04006.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020b. Efficient Transformers: A Survey. *arXiv:2009.06732 [cs]*. ArXiv: 2009.06732.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A Large Language Model for Science. ArXiv:2211.09085 [cs, stat].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Changhan Wang, Kyunghyun Cho, and Jiatao Gu. 2019. Neural Machine Translation with Byte-Level Subwords. ArXiv:1909.03341 [cs].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, 2:79–84.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*. ArXiv: 1906.08237.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. *arXiv:2106.11520 [cs]*. ArXiv: 2106.11520.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2021. Big Bird: Transformers for Longer Sequences. *arXiv:2007.14062 [cs, stat]*. ArXiv: 2007.14062.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *arXiv:1912.08777 [cs]*. ArXiv: 1912.08777.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*. ArXiv: 1904.09675.

A Hyperparameters and Training Details

Model	Data	# Steps	Train Loss	Eval Loss
small	caselaw	50K	14.61	15.78
small	caselaw	100K	13.93	15.07
small	caselaw	150K	13.63	14.77
small	caselaw	200K	13.38	14.49
small	diverse	50K	13.75	12.70
small	diverse	100K	12.78	11.66
small	diverse	150K	12.28	11.29
small	diverse	200K	12.05	11.03
base	caselaw	50K	12.40	13.76
base	caselaw	100K	11.67	12.99
base	caselaw	150K	11.31	12.58
base	caselaw	200K	11.02	12.27
base	diverse	50K	10.70	10.01
base	diverse	100K	9.86	9.22
base	diverse	150K	9.42	8.79
base	diverse	200K	9.20	8.56

Table 7: Training and Evaluation losses for the different trained models. Note that these losses are the addition of the loss of the generator and the loss of the discriminator. Since the loss of the discriminator is much smaller, it is scaled by a factor of 50 to stabilize training.

In this section, we present additional details regarding training and the chosen hyperparameters.

A.1 Pretraining

We pretrained our models with batch size 128 and learning rate 5e-4 and 3e-4 for the small and base models respectively. We used a Longformer attention window of 256. As described in by Clark et al. (2020), we used 10000 warm up steps and a 4 and 3 times smaller generator than the discriminator in the small and base version respectively. In contrast to Clark et al. (2020), we reduced the generator's depth (number of hidden layers) instead of its width (embedding size, hidden size and intermediate size). We used a MLM probability of 25% for the generators. The pretraining losses are shown in Table 7.

A.2 Downstream Benchmarks

We finetuned on the summarization datasets using early stopping on the validation set with patience of 3 epochs. We used a batch size of 32 and learning rate of 7e-5 after tuning in {5e-4, 9e-5, 7e-5, 5e-5, 3e-5, 1e-5}. We used the bart-base default config for num_beams (4) and no_repeat_ngram_size (3).

Overall, we found the diverse models to be more robust in finetuning with less failed runs and typically higher performance.

A.3 Compute Costs

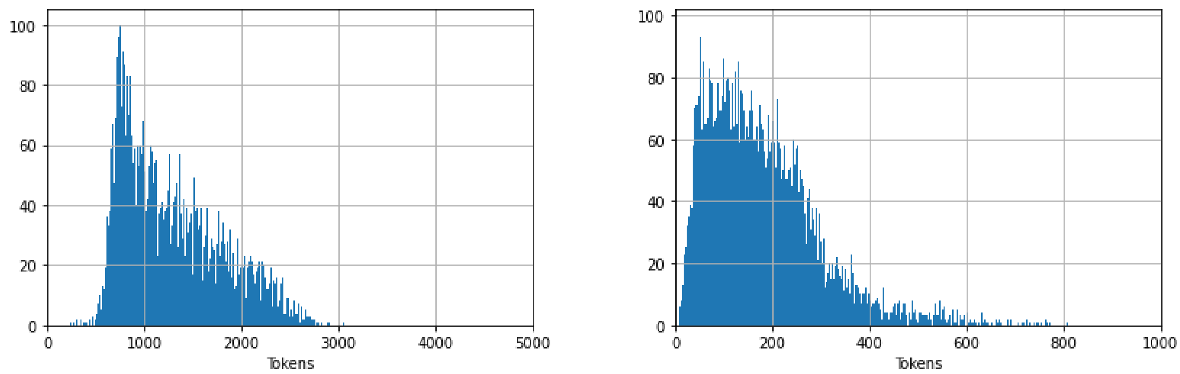
For running the pretraining, we used an AWS p3.8xlarge instance with 4 16GB NVIDIA V100 GPUs. Training the four models to 200K steps each, took approx. 36 days or 144 GPU days in total (almost. 6 days and almost 12 days for the small and base models respectively). Previous debug runs additionally consumed approx. 12 GPU days. For running the finetuning experiments, we used an AWS p3.16xlarge instance with 8 16GB NVIDIA V100 GPUs. Running the BillSum, and PubMed experiments including debugging and hyperparameter tuning took approximately 25 and 7 GPU days in total respectively. Putting it all together, we trained our models for 176 16GB NVIDIA V100 GPU days.

B Library Versions

We used the following versions to the libraries in a pip requirements.txt format:
datasets==2.4.0
huggingface-hub==0.9.0
nltk==3.7
pandas==1.3.5
rouge-score==0.1.2
scikit-learn==1.0.2
scipy==1.7.3
tokenizers==0.12.1
torch==1.12.1
tqdm==4.64.0
transformers==4.21.1

C Data Details

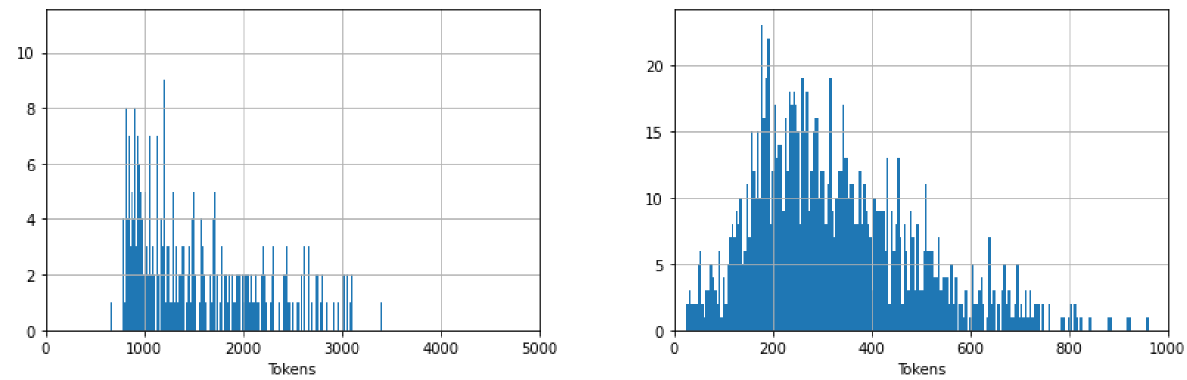
We used our own tokenizer to calculate the number of tokens. In Tables 3, and 4 we show the data length distributions for the BillSum train and test splits. In Tables 5, 6, and 7 we show the data length distributions for the PubMed train, validation and test splits.



(a) **Input Text**
Mean: 1289, Median: 1166
75-Quant: 1644, 95-Quant: 2290, Max: 3055

(b) **Summary**
Mean: 179, Median: 157
75-Quant: 240, 95-Quant: 398, Max: 808

Figure 3: Histograms for the BillSum training set (18949 samples).



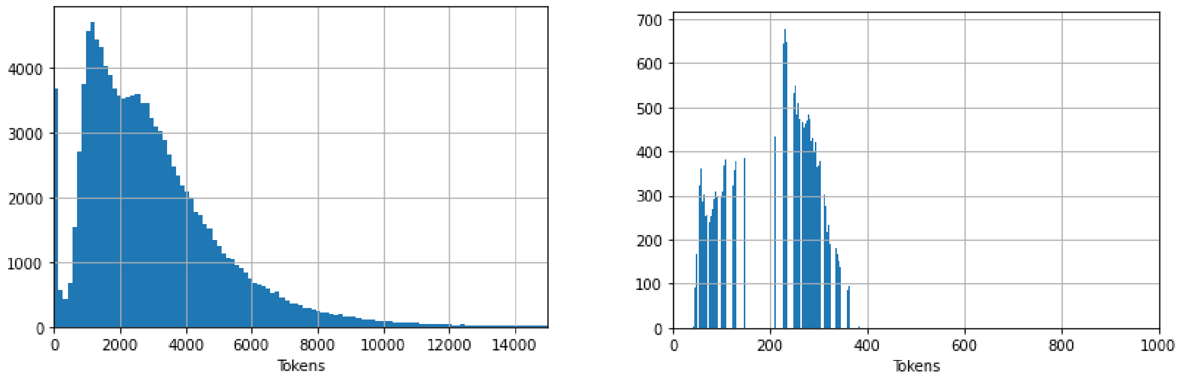
(a) **Input Text**
Mean: 1284, Median: 1164
75-Quant: 1629, 95-Quant: 2288, Max: 2957

(b) **Summary**
Mean: 179, Median: 156
75-Quant: 239, 95-Quant: 394, Max: 787

Figure 4: Histograms for the BillSum test set (3269 samples).

D Examples

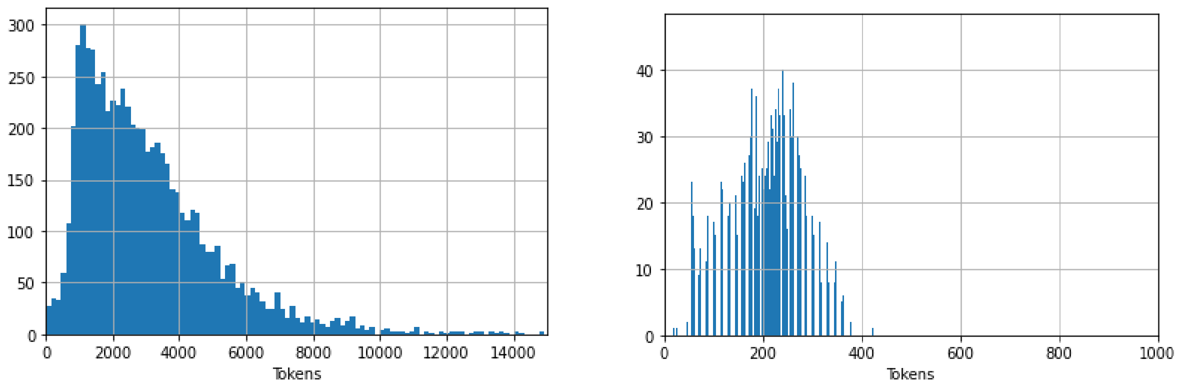
Example summaries are displayed in Tables 8, 9, 10, 11, 12, 13, 13, 15, and 16. Since the documents are very long sometimes, we truncated them to the first 2500 characters. We sorted the examples by RougeL scores and show the bottom 5%, bottom 25%, top 75% and top 95% percentile.



(a) **Input Text**
Mean: 3044, Median: 2572
75-Quant: 3996, 95-Quant: 7057, Max: 109759

(b) **Summary**
Mean: 202, Median: 208
75-Quant: 262, 95-Quant: 326, Max: 391

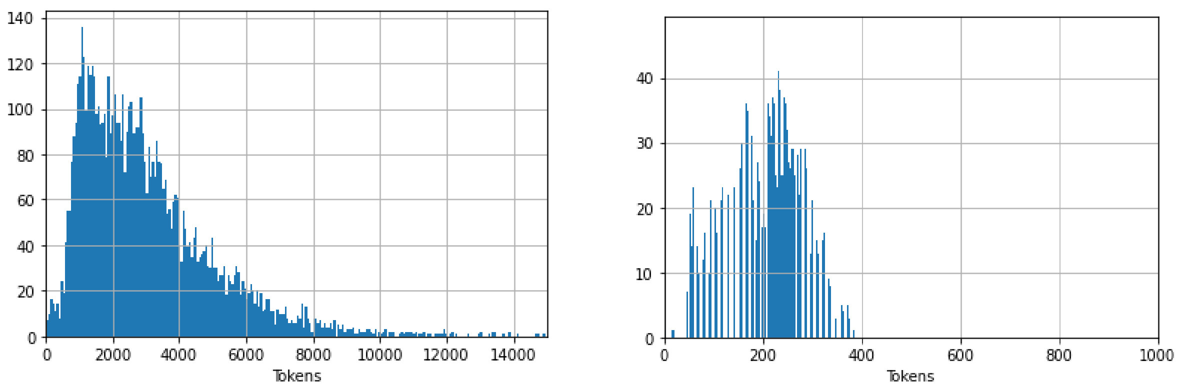
Figure 5: Histograms for the PubMed train set (119924 samples).



(a) **Input Text**
Mean: 3112, Median: 2609
75-Quant: 4011, 95-Quant: 6968, Max: 119269

(b) **Summary**
Mean: 203, Median: 209
75-Quant: 263, 95-Quant: 330, Max: 518

Figure 6: Histograms for the PubMed validation set (6633 samples).



(a) **Input Text**
Mean: 3093, Median: 2596
75-Quant: 3964, 95-Quant: 6985, Max: 48750

(b) **Summary**
Mean: 205, Median: 213
75-Quant: 265, 95-Quant: 329, Max: 506

Figure 7: Histograms for the PubMed test set (6658 samples).

Bottom 5% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the “Child Citizenship Act of 2000”. TITLE I–CITIZENSHIP FOR CERTAIN CHILDREN BORN OUTSIDE THE UNITED STATES SEC. 101. AUTOMATIC ACQUISITION OF CITIZENSHIP FOR CERTAIN CHILDREN BORN OUTSIDE THE UNITED STATES. (a) In General.–Section 320 of the Immigration and Nationality Act (8 U.S.C. 1431) is amended to read as follows: “children born outside the united states and residing permanently in the united states; conditions under which citizenship automatically acquired “Sec. 320. (a) A child born outside of the United States automatically becomes a citizen of the United States when all of the following conditions have been fulfilled: “(1) At least one parent of the child is a citizen of the United States, whether by birth or naturalization. “(2) The child is under the age of eighteen years. “(3) The child is residing in the United States in the legal and physical custody of the citizen parent pursuant to a lawful admission for permanent residence. “(b) Subsection (a) shall apply to a child adopted by a United States citizen parent if the child satisfies the requirements applicable to adopted children under section 101(b)(1).”. (b) Clerical Amendment.–The table of sections of such Act is amended by striking the item relating to section 320 and inserting the following: “Sec. 320. Children born outside the United States and residing permanently in the United States; conditions under which citizenship automatically acquired.”. SEC. 102. ACQUISITION OF CERTIFICATE OF CITIZENSHIP FOR CERTAIN CHILDREN BORN OUTSIDE THE UNITED STATES. (a) In General.–Section 322 of the Immigration and Nationality Act (8 U.S.C. 1433) is amended to read as follows: “children born and residing outside the united states; conditions for acquiring certificate of citizenship “Sec. 322. (a) A parent who is a citizen of the United States may apply for naturalization on behalf of a child born outside of the United States who has not acquired citizenship automatically under section 320. The Attorney General shall issue a certificate of citizenship to such parent upon proof, to the satisfaction of the Attorney General, that the following conditions have been fulfilled: “(1) At least one parent ...
Gold	Provides for issuance of a certificate of naturalization for a child born outside of the United States when the following conditions are met: (1) at least one parent is a U.S. citizen who has been present in the United States for not less than five years, at least two of which were after having attained the age of 14, or who has a citizen parent meeting such requirements; (2) the child is under 18 years old; and (3) the child is residing outside the United States in the legal and physical custody of the citizen parent, is temporarily and lawfully present in the United States, and is maintaining such lawful status. Applies such provision to an adopted child meeting certain definitional requirements who is adopted by a U.S. citizen parent. Title II: Protections for Certain Aliens Voting Based on Reasonable Belief of Citizenship - Amends the Immigration and Nationality Act respecting unlawful voting or false U.S. citizenship claims by permanent resident aliens under 16 years old having natural or adoptive U.S. citizen parents, to provide exceptions from certain provisions regarding deportability, moral character, inadmissability or related criminal penalties.
Model	table of contents: title i: citizenship for certain title ii: immigration and naturalization provisions title i: citizenship for certain children born outside the united states - amends the immigration and nationality act (ina) to revise naturalization requirements with respect to child born outside the united states. (sec. 102) amends the immigration and nationality act to revise requirements with respect to: (1) naturalization as a citizen of the united states; (2) naturalization of a u.s. citizen; (3) citizenship; (4) citizenship; (5) naturalization service; and (6)
Metrics	Rouge1: 35.9, Rouge2: 15.54, RougeL: 22.56, RougeLsum: 20.51, Summary length (tokens): 129
Bottom 25% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the “Effective Terrorists Prosecution Act of 2006”. SEC. 2. DEFINITION OF UNLAWFUL ENEMY COMBATANT. Paragraph (1) of section 948a of title 10, United States Code (as enacted by the Military Commissions Act of 2006 (Public Law 109-366)), is amended to read as follows: “(1) Unlawful enemy combatant.–The term ‘unlawful enemy combatant’ means an individual who directly participates in hostilities as part of an armed conflict against the United States who is not a lawful enemy combatant. The term is used solely to designate individuals triable by military commission under this chapter.”. SEC. 3. DETERMINATION OF UNLAWFUL ENEMY COMBATANT STATUS BY COMBATANT STATUS REVIEW TRIBUNAL NOT DISPOSITIVE FOR PURPOSES OF JURISDICTION OF MILITARY COMMISSIONS. Section 948d of title 10, United States Code (as enacted by the Military Commissions Act of 2006 (Public Law 109-366)), is amended– (1) by striking subsection (c); and (2) by redesignating subsection (d) as subsection (c). SEC. 4. EXCLUSION FROM TRIAL BY MILITARY COMMISSION OF STATEMENTS OBTAINED BY COERCION. Section 948r of title 10, United States Code (as enacted by the Military Commissions Act of 2006 (Public Law 109-366)), is amended by striking subsections (c) and (d) and inserting the following new subsection (c): “(c) Exclusion of Statements Obtained by Coercion.–A statement obtained by use of coercion shall not be admissible in a military commission under this chapter, except against a person accused of coercion as evidence that the statement was made.”. SEC. 5. DISCRETION OF MILITARY JUDGE TO EXCLUDE HEARSAY EVIDENCE DETERMINED TO BE UNRELIABLE OR LACKING IN PROBATIVE VALUE. Section 949a(b)(2)(E)(ii) of title 10, United States Code (as enacted by the Military Commissions Act of 2006 (Public Law 109-366)), is amended by striking “if the party opposing the admission of the evidence demonstrates that the evidence is unreliable or lacking in probative value” and inserting “if the military judge determines, upon motion by counsel, that the evidence is unreliable or lacking in probative value”. SEC. 6. DISCRETION OF MILITARY JUDGE TO TAKE CERTAIN ACTIONS IN EVENT THAT A SUBSTITUTE FOR CLASSIFIED EXCULPATORY EVIDENCE IS INS...
Gold	Effective Terrorists Prosecution Act of 2006 - Amends federal armed forces provisions enacted by the Military Commissions Act of 2006 to, among other things: (1) exclude from military commission (commission) trials statements obtained by coercion; (2) allow a commission military judge to exclude hearsay evidence determined to be unreliable or lacking in probative value; (3) provide for review of commission decisions by the U.S. Court of Appeals for the Armed Forces rather than the Court of Military Commission Review; (4) revise generally provisions concerning the implementation of treaty obligations with respect to the U.S. prosecution of enemy combatants; (5) restore habeas corpus rights for individuals detained by the United States; and (6) provide for expedited judicial review of provisions of the Military Commissions Act of 2006.
Model	effective terrorists prosecution act of 2006 - amends federal armed forces law to revise the definition of " unlawful enemy combatant" to include an individual who directly participated in hostilities as part of an armed conflict against the united states who is not a lawful enemy combatant. amends the military pay reform and reform act of 2006 to provide that a complaint obtained by mail or the military court of appeals for the u.s. military court of appeals for the armed forces who is not a lawful enemy combatant. directs the u.s. military department of defense to review the record in each case, except against a person aggrieved by prosecution or
Metrics	Rouge1: 51.64, Rouge2: 24.64, RougeL: 33.8, RougeLsum: 37.56, Summary length (tokens): 129
Top 75% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the “Census Address List Improvement Act of 1994”. SEC. 2. ADDRESS INFORMATION REVIEWED BY LOCAL GOVERNMENTS. (a) In General.–Chapter 1 of title 13, United States Code, is amended by adding after section 15 the following new section: “Sec. 16. Address information reviewed by States and local governments “(a) The Secretary, to assist efforts to ensure the accuracy of censuses and surveys under this title, shall– “(1) publish standards defining the content and structure of address information which States and local units of general purpose government may submit to the Secretary to be used in developing a national address list; “(2)(A) develop and publish a timetable for the Bureau to receive, review, and respond to submissions of information under paragraph (1) before the decennial census date; and “(B) provide for a response by the Bureau with respect to such submissions in which the Bureau specifies its determinations regarding such information and the reasons for such determinations; and “(3) be subject to the review process developed under section 3 of the Census Address List Improvement Act of 1994 relating to responses pursuant to paragraph (2). “(b)(1) The Secretary– “(A) shall provide officials who are designated as census liaisons by a local unit of general purpose government with access to census address information for the purpose of verifying the accuracy of the address information of the Bureau for census and survey purposes; and “(B) together with such access, should provide an explanation of duties and obligations under this title. “(2) Access under paragraph (1) shall be limited to address information concerning addresses within the local unit of general purpose government represented by the census liaison or an adjacent local unit of general purpose government. “(3) The Bureau should respond to each recommendation made by a census liaison concerning the accuracy of address information, including the determination (and reasons therefor) of the Bureau regarding each such recommendation. “(4) For the purposes of paragraph (1), in a case in which a local unit of general purpose government is within another local unit of general purpose government and is not independent of the enclosing unit, the census liaison shall be ...
Gold	Census Address List Improvement Act of 1994 - Directs the Secretary of Commerce to: (1) publish standards defining the content and structure of address information which States and local governments may submit to the Secretary to be used in developing a national address list; (2) develop and publish a timetable for the Bureau of the Census to receive, review, and respond to the submitted information before the decennial census date; (3) provide for a response by the Bureau that specifies its determinations regarding such information and the reasons for such determinations; and (4) be subject to the review process developed under this Act relating to such responses. Directs the Secretary to provide officials who are designated as census liaisons by local governments with access to census address information for the purpose of verifying the accuracy of the Bureau’s address information for census and survey purposes and together with such access, provide an explanation of duties and obligations under this Act. Limits such access to the addresses within the local government represented by the census liaison or an adjacent local government. Requires the Bureau to respond to each recommendation made by a census liaison concerning the accuracy of address information, including the determination (and reasons therefor) of the Bureau regarding each such recommendation. Prohibits a census liaison from using information made available under this Act for purposes other than the purposes specified in this Act. Makes provisions that require, with exceptions, that such information be treated as confidential applicable to local government census liaisons. Imposes a fine and up to five years’ imprisonment on whoever being or having been a census liaison wrongfully discloses such information. Requires: (1) the Administrator of the Office of Information and Regulatory Affairs, acting through the Chief Statistician, to develop an appeals process for those States and local governments which desire to appeal determinations of the Bureau; and (2) the Postal Service to provide to the Secretary for use by the Bureau such address, address- related, and point of postal delivery information, including postal delivery codes, determined by the Secretary to be appropriate for any census or survey being conducted by the Bureau.
Model	census address list improvement act of 1994 - directs the secretary of transportation to: (1) publish standards amending the content and structure of address information which states and local governments may submit to the secretary to the secretary to the secretary for such a national address list; (2) provide for a response by a census liaison with respect to such reports; and (3) provide for a response by a census liaison with respect to such reports. requires the secretary to: (1) provide officials who are designated as census liaison with access to census information; and (2) report annually to the congressional committees on the accuracy of
Metrics	Rouge1: 62.26, Rouge2: 41.9, RougeL: 50.0, RougeLsum: 52.83, Summary length (tokens): 129
Top 95% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the “National Geologic Mapping Reauthorization Act of 1996”. SEC. 2. FINDINGS. Congress finds that– (1) in enacting the National Geologic Mapping Act of 1992 (43 U.S.C. 31a et seq.), Congress found, among other things, that– (A) during the 2 decades preceding enactment of that Act, the production of geologic maps had been drastically curtailed; (B) geologic maps are the primary data base for virtually all applied and basic earth-science investigations; (C) Federal agencies, State and local governments, private industry, and the general public depend on the information provided by geologic maps to determine the extent of potential environmental damage before embarking on projects that could lead to preventable, costly environmental problems or litigation; (D) the lack of proper geologic maps has led to the poor design of such structures as dams and waste- disposal facilities; (E) geologic maps have proven indispensable in the search for needed fossil fuel and mineral resources; and (F) a comprehensive nationwide program of geologic mapping is required in order to systematically build the Nation’s geologic-map data base at a pace that responds to increasing demand; (2) the geologic mapping program called for by that Act has not been fully implemented; and (3) it is time for this important program to be fully implemented. SEC. 3. REAUTHORIZATION AND AMENDMENT. (a) Definitions.–Section 3 of the National Geologic Mapping Act of 1992 (43 U.S.C. 31b) is amended– (1) by striking “As used in this Act:” and inserting “In this Act:”; (2) by redesignating paragraphs (2), (3), (4), and (5) as paragraphs (3), (4), (5), and (6), respectively; (3) by inserting after paragraph (1) the following: “(2) Association.–The term ‘Association’ means the Association of American State Geologists.”; and (4) in each paragraph that does not have a heading, by ...
Gold	National Geologic Mapping Reauthorization Act of 1996 - Amends the National Geologic Mapping Act of 1992 to establish a national cooperative geologic mapping program between the U.S. Geological Survey and State geological surveys. Establishes a geologic mapping advisory committee to advise the Director of the U.S. Geological Survey on planning and implementation of the geological mapping program. Authorizes appropriations.
Model	national geologic mapping reauthorization act of 1996 - amends the national geologic mapping act of 1992 to establish a national cooperative geologic mapping program within the united states geological survey (usgs) to be administered and administered through the association. establishes a national cooperative geologic mapping program between the united states geological survey and the association. authorizes appropriations.
Metrics	Rouge1: 74.14, Rouge2: 56.14, RougeL: 65.52, RougeLsum: 67.24, Summary length (tokens): 69

Table 8: Examples of the BillSum dataset using the model billsum-1024-128 small diverse

Bottom 5% example (Sorted by rougeL)	
Document	SECTION 1. NATIONAL GUARD SUPPORT FOR BORDER CONTROL ACTIVITIES. (a) Operation Jump Start.– (1) In general.–Not fewer than 6,000 National Guard personnel shall continue to be deployed along the international border between the United States and Mexico under Operation Jump Start until the date on which the Federal Government has achieved operational control of such border (as defined in section 2(b) of the Secure Fence Act of 2006 (Public Law 109- 367)). (2) Exemption.–National Guard personnel deployed pursuant to paragraph (1) shall not be included in the calculation to determine compliance with– (A) limits on end strength; or (B) limits on the number of National Guard personal that may be placed on active duty for operational support. (b) Support Authorized.–Chapter 1 of title 32, United States Code, is amended by inserting after section 112 the following: "Sec. 112a. Border control activities "(a) Funding Assistance.–The Secretary of Defense may provide funds to the Governor of a State who submits to the Secretary a State border control activities plan satisfying the requirements of subsection (c). Such funds shall be used for the following: "(1) The pay, allowances, clothing, subsistence, gratuities, travel, and related expenses, as authorized by State law, of personnel of the National Guard of that State used, while not in Federal service, for the purpose of border control activities. "(2) The operation and maintenance of the equipment and facilities of the National Guard of that State used for the purpose of border control activities. "(3) The procurement of services and equipment, and the leasing of equipment, for the National Guard of that State used for the purpose of border control activities. However, the use of such funds for the procurement of equipment may not exceed $5,000 per item, unless approval for procurement of equipment in excess of that amount is granted in advance by the Secretary of Defense. "(b) Use of Personnel Performing Full-Time National Guard Duty.– (1) Under regulations prescribed by the Secretary of Defense, personnel of the National Guard of a State may, in accordance with the State...
Gold	Provides for continued National Guard support for border control activities along the U.S.- Mexico border under Operation Jump Start until operational control of such border has been achieved. Authorizes the Secretary of Defense to provide funding to a state that submits to the Secretary of State a state border control activities plan that meets certain requirements. Limits the number of National Guard that may be so deployed.
Model	(this measure has not been amended since it was passed by the senate on december 21, 2011. the summary of that version is repeated here.) states that not fewer than six,000, national guard personnel shall continue to be deployed along the international border between the united states and mexico under operation jump start up to six,000 shall not be included in the annual calculation to determine compliance with: (1) limits on end strength or limits on the number of national guard personal that may be placed on active duty for operational support; and (2) the pay, allowances, and benefits of the national guard of a state shall
Metrics	Rouge1: 46.33, Rouge2: 14.86, RougeL: 22.6, RougeLsum: 32.77, Summary length (tokens): 129
Bottom 25% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the "International Commerce Enhancement Act". SEC. 2. REFERENCES. Except as otherwise specifically provided in this Act, whenever in this Act an amendment or repeal is expressed as the amendment or repeal of a section or other provision, the reference shall be considered to be made to a section or other provision of the Arms Export Control Act (22 U.S.C. 2751 et seq.). SEC. 3. FOREIGN AND NATIONAL SECURITY POLICY OBJECTIVES AND RESTRAINTS. (a) Value of Defense Articles and Services.–Section 3(d) (22 U.S.C. 2753(d)) is amended in paragraphs (1) and (3)(A)– (1) by striking "$14,000,000" each place it appears and inserting "$25,000,000"; and (2) by striking "$50,000,000" each place it appears and inserting "$85,000,000". (b) Transfers With Respect to NATO and Major Non-NATO Countries.– Section 3(d) (22 U.S.C. 2753(d)) is amended– (1) in paragraph (2)– (A) in subparagraph (A), by striking "Except as provided in subparagraph (B), unless" and inserting "Unless"; and (ii) in subparagraph (B) to read as follows: "(B) Subparagraph (A) shall not apply in the case of a proposed transfer to the North Atlantic Treaty Organization, or any member country of such Organization, Japan, Australia, or New Zealand."; and (iii) in subparagraph (C), by striking "or (B)"; and (2) in paragraph (3)– (A) in the second sentence of subparagraph (A), by striking "shall be submitted" and all that follows through "unless the President" and inserting "shall be submitted at least 30 calendar days before such consent is given in the case of a transfer to a country other than a country which is a member of the North Atlantic Treaty Organization, Japan, Australia, or New Zealand, unless the President"; (B) in the third sentence of subparagraph (A), by striking "(thus waiving the requirements of clause (i) or (ii), as the case may be, and of subparagraph (B))"; and (C) in subparagraph (B)– (i) by striking ...
Gold	International Commerce Enhancement Act - Amends the Arms Export Control Act to increase threshold values of major defense equipment or defense articles or related training or other defense services whose transfer or lease to foreign countries would require a presidential certification to Congress. Exempts from specified congressional oversight requirements any such transfers to North Atlantic Treaty Organization (NATO) countries, Japan, Australia, or New Zealand.
Model	international commerce enhancement act - amends the arms export control act to: (1) increase the amount of defense articles and services from $10 million to $10 million to $10 million the value of defense articles and services (currently, $10 million); (2) prohibit the transfer of defense articles or services to the north atlantic treaty organization (nato); and (3) prohibit the transfer of defense articles or services from the north atlantic treaty organization (nato) defense articles or services. repeals the requirement that the transfer of defense articles or services from defense articles or services from the north atlantic treaty (nato)
Metrics	Rouge1: 44.72, Rouge2: 21.38, RougeL: 34.78, RougeLsum: 38.51, Summary length (tokens): 129
Top 75% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the "Family Education Reimbursement Act of 2005". SEC. 2. FAMILY EDUCATION REIMBURSEMENT ACCOUNTS. (a) Establishment.–The Secretary of Education, in consultation with the Secretary of Health and Human Services, shall– (1) establish a Family Education Reimbursement Account Program under which, at the direction of the parent of each displaced student who signs up under subsection (d), the Secretary provides reimbursement to enable the student or preschool-age child to attend the school or preschool program of his or her parent's choice during the 2005-2006 school year; (2) of the amount available to carry out this section for fiscal year 2006, use not more than one third of one percent of such amount for administrative expenses, including outreach, support services, and dissemination of information; and (3) contract with a nongovernmental entity to administer and operate the program. (b) Reimbursement.– (1) In general.–In carrying out this section, the Secretary– (A) shall allow the parent of the participating displaced student to select the school or preschool program to be attended by the student during the 2005- 2006 school year; (B) at the direction of the parent, shall provide reimbursement to that school or preschool program on a quarterly basis; and (C) in the case of a public school, may provide such reimbursement to the appropriate local fiscal agent for the school. (2) Amount.–In providing reimbursement under paragraph (1), the Secretary shall– (A) determine the amount of reimbursement to a school or preschool program based on the number of weeks during which the participating displaced student attended the school or preschool program during the preceding quarter; (B) subject to subparagraph (C), provide the same amount of reimbursement to each school and preschool program for each week of attendance by one participating displaced student; (C) not provide reimbursement...
Gold	Family Education Reimbursement Act of 2005 - Directs the Secretary of Education to establish a Family Education Reimbursement Account Program for families of students displaced by Hurricane Katrina or Hurricane Rita. Provides for reimbursing parents for costs of such students or preschool-age children attending schools or preschool programs, chosen by the parents, during the 20052006 school year. Requires the Secretary to make a contract with a nongovernmental entity to administer and operate the program.
Model	family education reimbursement act of 2005 - directs the secretary of education to establish a family education reimbursement account program under which, at the direction of the parent of each displaced student who signs up under the 2005- 2006 school year, the secretary of education shall: (1) provide reimbursement to the student or preschool child to attend the school or breakfast program of his or her parent's choice during the 2005- 2006- 2006 school year; (2) contract with a nongovernmental entity to administer and operate the program; and (3) contract with a nongovernmental entity to administer the program. requires the secretary to develop and implement a web-
Metrics	Rouge1: 63.74, Rouge2: 44.44, RougeL: 52.75, RougeLsum: 61.54, Summary length (tokens): 129
Top 95% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the "Pharmaceutical Products Price Equity Act". SEC. 2. FINDINGS. The Congress finds the following: (1) Modern life enhancing and life sustaining drugs are the products of decades of scientific research and practice. (2) The refining and production of products by American pharmaceutical enterprises must be appropriately compensated. (3) The pricing of American pharmaceutical products varies greatly from country to country with the corporate owners of such products voluntarily assuming no losses by always selling their products above the cost of production. (4) Considering the vital life and death nature of pharmaceutical products, it is appropriate that the United States Government, in harmony with other nations belonging to the World Trade Organization, should engage in the reasonable regulation of the prices of pharmaceutical products. SEC. 3. REASONABLE LIMITS ON PROFITS OF PHARMACEUTICAL PRODUCTS. (a) Presidential Authority.–The President shall issue such orders and regulations, and establish such procedures and reporting requirements, as the President determines to be appropriate to ensure that no pharmaceutical product (as defined by the President in such regulation or order) is sold to any consumer in the United States at a price that is more than 6 percent above the average retail price at which such pharmaceutical product is sold in the 5 most industrialized, free-market countries, other than the United States, as determined by the President. (b) Delegation.–The President may delegate the performance of any function under this section to any officer of any Federal department or agency who has been appointed by the President, by and with the consent of the Senate. (c) Confidentiality of Information.–All information reported to or otherwise obtained by any person exercising any authority under this section which contains or relates to a trade secret or other matter referred to in section 1905 of title 18, United States Code, shall be considered confidential for purposes of that section, except that such information may be disclosed to any other officer or employee of the United States involved in carrying out this section solely for the purpose of carrying out, and enforcing compliance with, this section. (d) Subpoena Powe...
Gold	Pharmaceutical Products Price Equity Act - Directs the President to issue orders and regulations, and establish procedures and reporting requirements, to ensure that no pharmaceutical product is sold to any consumer in the United States at a price that is more than six percent above the average retail price at which such pharmaceutical product is sold in the five most industrialized, free-market countries, other than the United States. Sets forth enforcement and civil liability provisions.
Model	pharmaceutical products price equity act - directs the president to issue orders and regulations to ensure that no pharmaceutical product is sold to any consumer in the united states at a price more than six percent above the average retail price at which such product is sold in the five most industrialized, free-market countries, other than the united states. authorizes the president to delegate the performance of any such function to any federal department or agency who has been appointed by the president, by and with the consent of the senate, by and with the consent of the senate.
Metrics	Rouge1: 70.11, Rouge2: 63.95, RougeL: 68.97, RougeLsum: 68.97, Summary length (tokens): 113

Table 9: Examples of the BillSum dataset using the model billsum-1024-128 base diverse

Bottom 5% example (Sorted by rougeL)	
Document	.–(1) For purposes of subsection (a)(2) and this subsection, the term ‘joint resolution’ means only a joint resolution introduced by a qualifying Member specified in paragraph (2) after the date on which the report of the President under subsection (a)(1) is received by the Congress– “(A) the matter after the resolving clause of which is as follows: ‘That the Congress hereby concurs in the certification of the President relating to deployment of a National Missile Defense system as submitted to Congress pursuant to section 4(b) of the National Missile Defense Act of 1999.’; “(B) which does not have a preamble; and “(C) the title of which is as follows: ‘Joint resolution relating to deployment of a National Missile Defense system.’. “(2) For purposes of this subsection, a qualifying Member described in this paragraph is– “(A) in the case of the House of Representatives, the majority leader or minority leader of the House of Representatives or a Member of the House of Representatives designated by the majority leader or minority leader; and “(B) in the case of the Senate, the majority leader or minority leader of the Senate or a Member of the Senate designated by the majority leader or minority leader. “(3) The provisions of paragraphs (3) through (8) of section 4(c) of the National Missile Defense Deployment Criteria Act of 2001 shall apply to a joint resolution under this subsection in the same manner as to a joint resolution under such section.”. SEC. 4. LIMITATION ON OBLIGATION OF FUNDS FOR PROCUREMENT FOR NATIONAL MISSILE DEFENSE SYSTEM. (a) Limitation.–No funds appropriated to the Department of Defense for procurement may be obligated for the National Missile Defense system unless– (1) the President submits to Congress a report concerning testing of the National Missile Defense system against countermeasures that includes a certification described in subsection (b); and (2) a joint resolution concurring in the President’s certification in such report is enacted as provided for in this section. (b) Presidential Certification.–A certification described in this subsection is a certification by the President that– (1) an adequate testing program for the National Missile ...
Gold	National Missile Defense Deployment Criteria Act of 2001 - Amends the National Missile Defense Act of 1999 to allow deployment of a national missile defense system (system) only if: (1) the system is technologically feasible; (2) system cost in relation to other Department of Defense (DOD) priorities will not lead to an overall reduction in national security by reducing resources available for other defense priorities; (3) the system will not diminish overall U.S. national security; (4) the system will not threaten to disrupt relations with U.S. nuclear allies, U.S. European allies, Russia, the People’s Republic of China, and other nations; and (5) the threat of a long-range ballistic missile attack from a nation of concern is clearly demonstrated.Prohibits the President from directing DOD to deploy a system unless and until: (1) the President certifies to Congress that the above deployment conditions have been met; and (2) a joint resolution is enacted concurring in the President’s certification.Prohibits DOD procurement funds from being obligated for a system unless: (1) the President certifies to Congress that adequate system tests have been undertaken to meet identified threats against countermeasures; and (2) a joint resolution is enacted concurring in the President’s certification.Requires the Secretary of Defense to direct the Ballistic Missile Defense Organization to: (1) include specified system countermeasures in system ground and flight testing conducted before the system becomes operational; and (2) determine the extent to which the exoatmospheric kill vehicle and the system can reliably discriminate between warheads and such countermeasures.
Model	prohibits funds appropriated to the department of defense (dod) for procurement from being obligated for the national missile defense system unless the president certifies to congress that: (1) an adequate testing program for the system is in place to meet the threats identified in the report; and (2) an adequate ground and flight testing of the system has been conducted against the system that are likely to be used against the system and that other countries have or are likely to acquire.
Metrics	Rouge1: 40.69, Rouge2: 16.67, RougeL: 20.0, RougeLsum: 20.0, Summary length (tokens): 94
Bottom 25% example (Sorted by rougeL)	
Document	TITLE I–FEDERAL AIRPORTS SECURITY ENHANCEMENT ACT SEC. 101. SHORT TITLE. This title may be cited as the “Federal Airports Security Enhancement Act”. SEC. 102. ESTABLISHMENT OF AIRPORT SECURITY COMMITTEES. The Act of July 5, 1994 (49 U.S.C. 44935), is amended– (1) by striking section 44901 subparagraph (b) and inserting the following: “SEC. 103. EMPLOYMENT STANDARDS AND TRAINING.”. (2) by striking section 44935 subparagraph (b) and inserting the following: “(a) Review and Recommendations.–The Administrator of the Federal Aviation Administration shall establish Security Committees at each airport location to be composed of representatives of the air carriers, airport operators, other interested parties and at least one representative from the Federal Protective Service, the Federal Bureau of Investigation, The Federal Aviation Administration and one member from each local jurisdiction that the airport may be located in or that may have jurisdictional authority for the airport facility. Each Airport Security Committee shall meet at least quarterly and shall make recommendations for minimum security countermeasures to the Administrator. The Federal Protective Service shall have primary responsibility for conducting on an ongoing basis security surveys and formulating recommendations to the Security Committee. The Administrator shall prescribe appropriate changes in existing procedures to improve that performance.”. SEC. 103. SCREENING PASSENGERS AND PROPERTY. The Act of July 5, 1994 (49 U.S.C. 44935), is amended by striking section 44901, subparagraph (a), and inserting the following: “(a) General Requirements.–The Administrator of the Federal Aviation Administration shall prescribe regulations requiring screening of all passengers and property that will be carried in a cabin of an aircraft in air transportation or intrastate air transportation. The screening must take place before boarding and be carried out by a weapon detecting facility or procedure used or operated by an employee or agent of the Federal Protective Service. The Administrator– “(1) shall require that sufficient Federal Police Officers are posted at airport facilities to provide patrol duties during all hours of operations as well as supervise screening personnel; “(2) shall maintain sufficient numbers of Special Agents to provid...
Gold	Federal Airports Security Enhancement Act - Amends Federal aviation law to direct the Administrator of the Federal Aviation Administration (FAA) to establish at each airport a Security Committee which shall make recommendations for minimum security counter-measures. Requires the Administrator, on the basis of such recommendations, to prescribe appropriate changes to improve the performance of existing airport security procedures.Requires the screening of passengers and property that will be carried in a cabin of an aircraft to be carried out by Federal Protective Service employees or agents. (Currently, screening is carried out by employees or agents of an air carrier, interstate air carrier, or foreign air carrier).Authorizes the Administrator of the General Services Administration (GSA) to appoint police officers and special agents (currently, special policemen and nonuniformed special policemen) for the policing of all Federal buildings (including buildings under the control of the GSA). Sets forth certain additional powers of such officers and agents, including the authority to carry firearms and to police areas adjacent to Federal property.Establishes the Federal Protective Service as a separate operating service of the GSA. Calls for at least 1,000 full-time equivalent Service police officers to be assigned to areas outside of airport operations. Requires the Commissioner of the Service to prescribe minimum employment and training standards to be applied in the contracting of security personnel for the policing of buildings and areas controlled by the United States and GSA. Authorizes GSA to recover airport security costs from the FAA.
Model	table of contents: title i: federal airports security enhancement act title ii: miscellaneous provisions general federal airports security enhancement act - title i: federal airports security enhancement - amends the federal aviation act of 1992 to direct the administrator of the federal aviation administration (faa) to prescribe regulations requiring screening of all passengers and property that will be carried in a port of aircraft in air transportation or intrastate air transportation. (sec. 102) directs the administrator to prescribe regulations requiring screening of all passengers and property that will be carried out by the federal protective service, the federal bureau of investigation (fbi), the federal bureau of investigation (fbi), and one member from each local jurisdiction that the aircraft may be located in or that may have jurisdictional authority for the airport of an aircraft in air transportation or intrastate air transportation. (sec. 103) directs the administrator to prescribe regulations requiring screening of all passengers and property that will be carried out by a weapon detection facility or procedure used or operated by an employee or agent of the federal protective service. (sec. 103) authorizes the administrator to enter into agreements with state and local law enforcement authorities to obtain authority for, jointly with state and local law enforcement authorities. (
Metrics	Rouge1: 52.44, Rouge2: 22.84, RougeL: 29.7, RougeLsum: 47.8, Summary length (tokens): 256
Top 75% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the “Patent and Trademark Office Authorization Act of 2002”. SEC. 2. AUTHORIZATION OF AMOUNTS AVAILABLE TO THE PATENT AND TRADEMARK OFFICE. (a) In General.–There are authorized to be appropriated to the United States Patent and Trademark Office for salaries and necessary expenses for each of the fiscal years 2003 through 2008 an amount equal to the fees estimated by the Secretary of Commerce to be collected in each such fiscal year, respectively, under– (1) title 35, United States Code; and (2) the Act entitled “An Act to provide for the registration and protection of trademarks used in commerce, to carry out the provisions of certain international conventions, and for other purposes”, approved July 5, 1946 (15 U.S.C. 1051 et seq.) (commonly referred to as the Trademark Act of 1946). (b) Estimates.–Not later than February 15, of each fiscal year, the Undersecretary of Commerce for Intellectual Property and the Director of the Patent and Trademark Office (in this Act referred to as the Director) shall submit an estimate of all fees referred to under subsection (a) to be collected in the next fiscal year to the chairman and ranking member of– (1) the Committees on Appropriations and Judiciary of the Senate; and (2) the Committees on Appropriations and Judiciary of the House of Representatives. SEC. 3. ELECTRONIC FILING AND PROCESSING OF PATENT AND TRADEMARK APPLICATIONS. (a) Electronic Filing and Processing.–Not later than December 1, 2004, the Director shall complete the development of an electronic system for the filing and processing of patent and trademark applications, that– (1) is user friendly; and (2) includes the necessary infrastructure to– (A) allow examiners and applicants to send all communications electronically; and (B) allow the Office to process, maintain, and search electronically the contents and history of each application. (b) Authorization of Appropriations.–Of amounts authorized under section 2, there are authorized to be appropriated to carry out subsection (a) of this section not more than $50,000,000 for each of fiscal years 2003 and 2004. Amounts made available under this subsection shall...
Gold	Patent and Trademark Office Authorization Act of 2002 - Authorizes appropriations to the U.S. Patent and Trademark Office for salaries and expenses for FY 2003 through 2008 in an amount equal to all patent and trademark fees estimated by the Secretary of Commerce (Secretary) to be collected in each such fiscal year.(Sec. 2) Requires the Under Secretary of Commerce for Intellectual Property and the Director of the Office (Director), by February 15 of each fiscal year, to report an estimate of all fees to be collected in the next fiscal year to the chairman and ranking member of specified congressional committees.(Sec. 3) Requires the Director, by December 1, 2004, to complete the development of an electronic system for the filing and processing of patent and trademark applications that: (1) is user friendly; and (2) includes the necessary infrastructure to allow examiners and applicants to send all communications electronically, and the Office to process, maintain, and search electronically the contents and history of each application. Authorizes appropriations for FY 2003 and 2004 for development of such system.(Sec. 4) Requires the Secretary, in each of the five calendar years following the enactment of this Act, to report to specified congressional committees on the progress made in implementing the 21st Century Strategic Plan issued on June 3, 2002, and on any amendments made to it.(Sec. 5) Amends Federal patent law to provide that previous citation by or to, or consideration by the Office of, a patent or printed publication does not preclude the existence of a substantial new question of patentability in patent reexamination proceedings.(Sec. 6) Revises requirements for appeals in inter partes reexamination proceedings to allow a third-party requester to appeal to the U.S. Court of Appeals for the Federal Circuit, or be a party to any appeal taken by the patent owner, with respect to any final decision favorable to the patentability of any original or proposed amended or new claim of the patent.Allows a third-party requester to appeal a decision of the Board of Patent Appeals and Interferences.Provides that a third-party requester in an inter partes reexamination proceeding dissatisfied with the final decision in an appeal to the Board may appeal the decision only to the U.S. Court of Appeals for the Federal Circuit.
Model	patent and trademark office authorization act of 2002 - authorizes appropriations to the u.s. patent and trademark office for fy 2003 through 2008. requires the director of the patent and trademark office to: (1) complete the development of an electronic system for the filing and processing of patent and trademark applications; and (2) submit an annual report to the congressional committees on progress made in implementing the 21st century strategic plan issued under the federal patent and trademark programs.
Metrics	Rouge1: 48.99, Rouge2: 39.86, RougeL: 44.3, RougeLsum: 48.32, Summary length (tokens): 94
Top 95% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the “Guidance, Understanding, and Information for Dual Eligibles (GUIDE) Act”. SEC. 2. FINDINGS; PURPOSE. (a) Findings.–The Congress finds the following: (1) Nearly 8,800,000 Americans were eligible for benefits under the Medicare program and for medical assistance under Medicaid (dual eligible beneficiaries) in fiscal year 2005. Of these “dual eligible beneficiaries”, almost 40 percent have cognitive impairments, including Alzheimer’s disease, dementia, serious mental illnesses, and intellectual disabilities. Until December 31, 2005, dual eligible beneficiaries received outpatient prescription drug benefits through medical assistance under Medicaid. On January 1, 2006, drug coverage for dual eligibles switched from Medicaid to Medicare. (2) In 2008, 53 percent of dual eligible beneficiaries had medication access problems and of those, 27 percent experienced significant adverse clinical events. (3) Individuals with medication access issues experience significantly more adverse clinical events. Among dual eligible beneficiaries with mental illness who had medication access problems, 27 percent experienced significant adverse clinical events, which included emergency room visits and hospitalizations. (4) In total, over 1,000,000 dual eligible beneficiaries and low-income subsidy beneficiaries were automatically auto- enrolled to new benchmark prescription drug plans under part D of the Medicare program between 2006 and 2007. (5) Community providers are at the front line of helping the most vulnerable dual eligible beneficiaries obtain prescription drug coverage under the Medicare program and navigate complex enrollment and low-income subsidy eligibility requirements under such program. (b) Purpose.–It is the purpose of this bill to help low-income persons with cognitive impairments to enroll in and navigate the prescription drug benefit under the Medicare program by providing front line community providers who serve the population daily with financial assistance to conduct vigorous education and outreach and direct case management. SEC. 3. MEDICARE PRESCRIPTION DRUG OUTREACH DEMONSTRATION PROGRAM FOR DUAL E...
Gold	Guidance, Understanding, and Information for Dual Eligibles (GUIDE) Act - Directs the Secretary of Health and Human Services to establish a three-year demonstration program under which the Secretary awards grants and contracts to appropriate, qualified community programs and clinics for individuals with intellectual or developmental disabilities, or certain programs under the Public Health Services Act, to employ qualified social workers and case managers to provide one-on-one counseling about benefits under part D (Voluntary Prescription Drug Benefit Program) of title XVIII (Medicare) of the Social Security Act (SSA) to a full-benefit dual eligible individual (eligible for benefits under both Medicare and SSA title XIX [Medicaid]) who has one or more mental disabilities.
Model	guidance, understanding, and information for dual eligible beneficiaries with intellectual or developmental disabilities act - directs the secretary of health and human services (hhs) to establish a three-year demonstration program under which the secretary awards grants and contracts to qualified community programs and clinics for individuals with intellectual or developmental disabilities or such programs to provide medicare prescription drug assistance to individuals with intellectual or developmental disabilities or such programs.
Metrics	Rouge1: 60.87, Rouge2: 47.25, RougeL: 58.7, RougeLsum: 58.7, Summary length (tokens): 80

Table 10: Examples of the BillSum dataset using the model billsum-1024-256 small diverse

Bottom 5% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the "Health Coverage Tax Credit Extension Act of 2015". SEC. 2. EXTENSION AND MODIFICATION OF HEALTH COVERAGE TAX CREDIT. (a) Extension.–Subparagraph (B) of section 35(b)(1) of the Internal Revenue Code of 1986 is amended by striking "before January 1, 2014" and inserting "before January 1, 2020". (b) Coordination With Credit for Coverage Under a Qualified Health Plan.–Subsection (g) of section 35 of the Internal Revenue Code of 1986 is amended– (1) by redesignating paragraph (11) as paragraph (13), and (2) by inserting after paragraph (10) the following new paragraphs: "(11) Election.– "(A) In general.–A taxpayer may elect to have this section apply for any eligible coverage month. "(B) Timing and applicability of election.–Except as the Secretary may provide– "(i) an election to have this section apply for any eligible coverage month in a taxable year shall be made not later than the due date (including extensions) for the return of tax for the taxable year, and "(ii) any election for this section to apply for an eligible coverage month shall apply for all subsequent eligible coverage months in the taxable year and, once made, shall be irrevocable with respect to such months. "(12) Coordination with premium tax credit.– "(A) In general.–An eligible coverage month to which the election under paragraph (11) applies shall not be treated as a coverage month (as defined in section 36B(c)(2)) for purposes of section 36B with respect to the taxpayer. "(B) Coordination with advance payments of premium tax credit.–In the case of a taxpayer who makes the election under paragraph (11) with respect to any eligible coverage month in a taxable year or on behalf of whom any advance payment is made under section 7527 with respect to any month in such taxable year– ...
Gold	Health Coverage Tax Credit Extension Act of 2015 This bill extends the tax credit for health insurance costs of a taxpayer and qualifying family members through 2019. The tax credit for health insurance costs is a refundable tax credit equal to 72.5% of the cost of qualified health coverage paid by an eligible individual [defined as an individual who is receiving a trade adjustment allowance, is eligible for the alternative trade adjustment assistance program, or is over age 55 and receives pension benefits from the Pension Benefit Guaranty Corporation (PBGC)]. The bill requires a taxpayer to make an election to have the tax credit apply for any eligible coverage month during a taxable year. An eligible coverage month is a month in which an eligible individual is covered by qualified health insurance, does not have other specified coverage, and is not imprisoned. The bill also directs the Departments of the Treasury, Health and Human Services, and Labor and the PBGC to conduct a public outreach, including on the Internet, to inform individuals eligible for the tax credit for health insurance costs on the extension of such credit and the availability of the election to claim such credit retroactively for coverage months beginning after December 31, 2013.
Model	health coverage tax credit extension act of 2015 this bill amends the internal revenue code, with respect to health care coverage, to: (1) extend through 2020 the tax credit for advance payments to individuals, (2) allow advance payments of advance payments of advance payments of advance payments, and (3) extend through 2018 the tax credit for advance payments of advance payments of advance payments to individuals.
Metrics	Rouge1: 26.37, Rouge2: 11.07, RougeL: 21.25, RougeLsum: 25.64, Summary length (tokens): 82
Bottom 25% example (Sorted by rougeL)	
Document	SECTION 1. EXTENSION. (a) In General.–Chapter 5 of subtitle B of the Agricultural Marketing Act of 1946 (7 U.S.C. 1636 et seq.) is amended by adding at the end the following new section: "SEC. 260. TERMINATION OF AUTHORITY. "The authority provided by this subtitle terminates on September 30, 2010.". (b) Conforming Amendment and Extension.–Section 942 of the Livestock Mandatory Reporting Act of 1999 (7 U.S.C. 1635 note; Public Law 106-78) is amended by striking "terminate on September 30, 2005" and inserting "(other than section 911 of subtitle A and the amendments made by that section) terminate on September 30, 2010". SEC. 2. DEFINITIONS. (a) Base Market Hogs.–Section 231(4) of the Agricultural Marketing Act of 1946 (7 U.S.C. 1635i(4)) is amended to read as follows: "(4) Base market hog.–The term 'base market hog' means a barrow or gilt for which no discounts are subtracted from and no premiums are added to the base price.". (b) Boars.–Section 231(5) of such Act (7 U.S.C. 1635i(5)) is amended to read as follows: "(5) Boar.–The term 'boar' means a sexually-intact male swine.". (c) Packer of Sows and Boars.–Section 231(12) of such Act (7 U.S.C. 1635i(12)) is amended by– (1) striking subparagraph (B) and inserting the following new subparagraph: "(B) for any calendar year, the term includes only– "(i) a swine processing plant that slaughtered an average of at least 100,000 swine per year during the immediately preceding five calendar years; and "(ii) a person that slaughtered an average of at least 200,000 sows, boars, or any combination thereof, per year during the immediately preceding five calendar years; and"; and (2) in subparagraph (C)– (A) by inserting "or person" after "swine processing plant"; (B) by inserting "or person" after "plant capacity of the processing plant"; and (C) by inserting "or person" after "determining whether the processing plant". SEC. 3. REPORTING; BARROWS AND GILTS. Section 232(c) of the Agricultural Marketing Act of 1946 (7 U.S.C. 1635j(c)) is amended to read as follows: "(c) Daily Reporting; Barrows and Gilts.– "(1) Prior day report.– "(A) In general.–The corporate officers or officially d...
Gold	Amends the Agricultural Marketing Act of 1946 to extend the provisions of the Livestock Mandatory Price Reporting Act of 1999 through September 30, 2010. Amends swine price reporting provisions.
Model	amends the agricultural marketing act of 1946 to terminate the authority of the secretary of agriculture (usda) to: (1) livestock processing plant processing plant slaughter, and (2) slaughtering plant slaughter. amends the agricultural marketing act of 1946 to: (1) revise minimum reporting requirements; and (2) revise reporting requirements. amends the agricultural marketing act of 1946 and the agricultural marketing act of 1946 to: (1) revise reporting requirements; and (2) revise reporting requirements.
Metrics	Rouge1: 33.66, Rouge2: 18.18, RougeL: 31.68, RougeLsum: 29.7, Summary length (tokens): 105
Top 75% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the "Maritime Administration Authorization Act for Fiscal Year 2001". SEC. 2. AUTHORIZATION OF APPROPRIATIONS FOR FISCAL YEAR 2001. Funds are hereby authorized to be appropriated, as Appropriations Acts may provide, for the use of the Department of Transportation for the Maritime Administration as follows: (1) For expenses necessary for operations and training activities, not to exceed $80,240,000 for the fiscal year ending September 30, 2001. (2) For the costs, as defined in section 502 of the Federal Credit Reform Act of 1990, of guaranteed loans authorized by title XI of the Merchant Marine Act, 1936 (46 U.S.C. App. 1271 et seq.), $50,000,000, to be available until expended. In addition, for administrative expenses related to loan guarantee commitments under title XI of that Act, $4,179,000. SEC. 3. AMENDMENTS TO TITLE IX OF THE MERCHANT MARINE ACT, 1936. (a) Title IX of the Merchant Marine Act, 1936 (46 U.S.C. App. 101 et seq.) is amended by adding at the end thereof the following: "SEC. 910. DOCUMENTATION OF CERTAIN DRY CARGO VESSELS. "(a) In General.–The restrictions of section 901(b)(1) of this Act concerning a vessel built in a foreign country shall not apply to a newly constructed drybulk or breakbulk vessel over 7,500 deadweight tons that has been delivered from a foreign shipyard or contracted for construction in a foreign shipyard before the earlier of– "(1) the date that is 1 year after the date of enactment of the Maritime Administration Authorization Act for Fiscal Year 2001; or "(2) the effective date of the OECD Shipbuilding Trade Agreement Act. "(b) Compliance With Certain U.S.-Build Requirements.–A vessel timely contracted for or delivered pursuant to this section and documented under the laws of the United States shall be deemed to have been United-States built for purposes of sections 901(b) and 901b of this Act if– "(1) following delivery by a foreign shipyard, the vessel has any additional shipyard work necessary to receive its initial Coast Guard certificate of inspection performed in a United States shipyard; "(2) the vessel is not documented in another country before being documented under the laws of the United States; "(3)...
Gold	(Sec. 3) Amends the Merchant Marine Act, 1936 to declare that certain restrictions concerning a vessel built in a foreign country shall not apply to a newly constructed drybulk or breakbulk vessel over 7,500 deadweight tons that has been delivered from a foreign shipyard or contracted for construction in a foreign shipyard before the earlier of two specified dates. Deems U.S.-built any vessel timely contracted for or delivered and documented under U.S. law, if certain conditions are met. (Sec. 4) Directs the Secretary of State, in coordination with the Secretary of Transportation, to initiate discussions in all appropriate international forums to establish an international standard for the scrapping of vessels in a safe and environmentally sound manner. Directs the Secretary of Transportation to develop, and report to specified congressional committees on, a program for the scrapping of obsolete National Defense Reserve Fleet Vessels. Amends the National Maritime Heritage Act of 1994 to extend, through September 30, 2006, the authority of the Secretary to dispose of certain vessels in the National Defense Reserve Fleet. Requires that such vessels be disposed of in the most cost effective manner to the United States, taking into account the need for disposal, the environment, and safety concerns. Amends Federal law to authorize the expenditure of funds from the National Defense Sealift Fund for costs related to the scrapping of National Defense Reserve Fleet vessels. Names vessels in the National Defense Reserve Fleet that may be scrapped in the United States or a foreign country. (Sec. 5) Requires the Maritime Administration (in its annual report to Congress and its estimated annual budget) to state separately the amount, source, intended use, and nature of any funds (other than funds appropriated to the Administration or to the Secretary for use by the Administration) administered, or subject to oversight, by the Administration. (Sec. 6) Amends Federal maritime law to authorize the Secretary of Transportation to make a grant to a National Maritime Enhancement Institute for maritime and maritime intermodal research as if the Institute were a university transportation center. (Sec. 7) Directs the Secretary to study maritime research and technology development, and report the results, including any recommendations, to Congress. Authorizes appropriations. (Sec. 8) Authorizes the Secretary to convey all right, title, and U.S. interest in the U.S.S. GLACIER (formerly of the National Defense Reserve Fleet) to the Glacier Society, Inc., Bridgeport, Connecticut.
Model	maritime administration authorization act for fiscal year 2001 - authorizes appropriations for the department of transportation (dot) for fy 2001 for: (1) operations and training activities; (2) training activities; and (3) administrative expenses.amends the merchant marine act, 1936 to make appropriations for fy 2001 through 2001 for the maritime administration.amends the merchant marine act, 1936 to apply certain restrictions concerning a vessel located in a foreign country to a newly constructed dry or breakable vessel over seven,500 feet that has been delivered from a foreign shipyard or contracted for construction in a foreign shipyard before the earlier of: (1) one year after enactment of this act, or (2) the effective date of the international maritime administration act. directs the secretary of state in coordination with the secretary of transportation to initiate discussions in all appropriate international forums in order to establish an international standard for the scrapping of vessels in a safe and environmentally sound manner. directs the secretary of state to initiate discussions in all appropriate international forums to establish an international standard for the scrapping of vessels in a safe and environmentally sound manner.
Metrics	Rouge1: 61.19, Rouge2: 41.5, RougeL: 47.76, RougeLsum: 57.21, Summary length (tokens): 222
Top 95% example (Sorted by rougeL)	
Document	SECTION 1. SMALL BUSINESS EXPENSING PROVISIONS MADE PERMANENT. (a) Increase in Small Business Expensing Made Permanent.– (1) In general.–Subsection (b) of section 179 of the Internal Revenue Code of 1986 (relating to limitations) is amended– (A) by striking "$25,000 ($125,000 in the case of taxable years beginning after 2006 and before 2011)" in paragraph (1) and inserting "$500,000", and (B) by striking "$200,000 ($500,000 in the case of taxable years beginning after 2006 and before 2011)" in paragraph (2) and inserting "$1,000,000". (2) Conforming amendment.–Section 179(b) of such Code is amended by striking paragraph (7). (b) Expensing for Computer Software Made Permanent.–Clause (ii) of section 179(d)(1)(A) of such Code is amended by striking "and which is placed in service in a taxable year beginning after 2002 and before 2011,". (c) Inflation Adjustment.– (1) So much of subparagraph (A) of section 179(b)(5) of such Code as precedes clause (i) thereof is amended to read as follows: "(A) In general.–In the case of any taxable year beginning in a calendar year after 2009, the $500,000 and $1,00,000 amounts in paragraphs (1) and (2) shall each be increased by an amount equal to–". (2) Section 179(b)(5)(A)(ii) of such Code is amended by striking "2006" and inserting "2008". (d) Effective Date.–The amendments made by this section shall apply to taxable years ending after the date of the enactment of this Act. SEC. 2. DEDUCTION FOR PURCHASE OF DOMESTICALLY MANUFACTURED AUTOMOBILES. (a) In General.–Part VII of subchapter B of chapter 1 of the Internal Revenue Code of 1986 (relating to additional itemized deductions for individuals) is amended by redesignating section 224 as section 225 and by inserting after section 223 the following new section: "SEC. 224. DEDUCTION FOR PURCHASE OF DOMESTICALLY MANUFACTURED AUTOMOBILES. "(a) Allowance of Deduction.–In the case of an individual, there shall be allowed as a deduction an amount equal to the cost of any qualified automobile placed in service by the taxpayer during the taxable year. "(b) Limitation Per Vehicle.–The amount of the ded...
Gold	Amends the Internal Revenue Code to: (1) increase and make permanent the expensing allowance for depreciable business assets; and (2) allow a tax deduction, up to $10,000, for the purchase of a motor vehicle manufactured in the United States. Terminates such tax deduction after 2010.
Model	amends the internal revenue code to make permanent: (1) the increased expensing allowance for depreciable business assets; and (2) the tax deduction for the purchase of manufactured manufactured automobiles.
Metrics	Rouge1: 72.0, Rouge2: 46.58, RougeL: 64.0, RougeLsum: 64.0, Summary length (tokens): 40

Table 11: Examples of the BillSum dataset using the model billsum-1024-256 base diverse

Bottom 5% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the "Public Health Equity Act". SEC. 2. FINDINGS. Congress finds that– (1) all communities and individuals are entitled to protection from occupational and other exposure to substances that are hazardous to the public health; (2) hazardous substances have had a disproportionate impact on the public health of poor and ethnic minority communities and individuals, resulting in exclusion from participation in, denial of benefits under, and discrimination under, programs and activities receiving Federal financial assistance; and (3) each Federal agency has an obligation to ensure that all federally assisted programs and activities that affect human health do not directly or through contractual arrangements use criteria, methods, or practices that cause discrimination on the ground of race, color, or national origin. SEC. 3. PUBLIC HEALTH EQUITY. The Public Health Service Act (42 U.S.C. 201 et seq.) is amended by adding at the end thereof the following new title: "TITLE XXVII–PUBLIC HEALTH EQUITY "SEC. 2701. DEFINITIONS. "As used in this title: "(1) Activity; program.–The term 'program or activity' means any operation of– "(A)(i) a department, agency, special purpose district, or other instrumentality of a State or of a local government; or "(ii) the entity of such State or local government that distributes such assistance and each such department or agency (and each other State or local government entity) to which the assistance is extended, in the case of assistance to a State or local government; "(B)(i) a college, university, or other postsecondary institution, or a public system of higher education; or "(ii) a local educational agency (as defined in section 198(a)(10) of the Elementary and Secondary Education Act of 1965), system of vocational education, or other school system; "(C)(i) an entire corporation, partnership, or other private organization, or an entire sole ...
Gold	Public Health Equity Act - Amends the Public Health Service Act to add a new title, title XXVII: Public Health Equity. Prohibits, under such title, acts with discriminatory effect as well as intentionally discriminatory acts that result in disproportionate exposure to a covered substance on the basis of race, color, or national origin. Defines a covered substance to include certain hazardous materials, substances, contaminants, or chemicals listed, identified, or defined in specified laws.
Model	public health equity act - amends the public health service act to require the president to ensure that no person shall be excluded from participation in, be denied the benefits of, or being subject to discrimination under, any program or activity on the ground of race, color, or national origin. requires the president to ensure that no person shall be excluded from participation in, be denied the benefits of, or being subject to discrimination under, any program or activity on the ground of race, color, or national origin. requires the secretaries of labor, health and human services, the administrator of the health and human services, the administrator, and any other head of a federal agency with responsibility for providing federal financial assistance to a program or activity to issue regulations implementing such nondiscrimination requirements. requires such regulations to: (1) declare that no person shall be excluded from participation in, be denied the benefits of, or be subject to discrimination under, any program or activity on the ground of race, color, or national origin; and (2) address actions of programs or activities that result in disproportionate exposure to a covered substance on the basis of race, color, or national origin.
Metrics	Rouge1: 28.89, Rouge2: 20.15, RougeL: 22.96, RougeLsum: 26.67, Summary length (tokens): 239
Bottom 25% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE; REFERENCES TO TITLE 38, UNITED STATES CODE. (a) Short Title.–This Act may be cited as the "Veterans Programs Improvement Act of 2003". (b) References.–Except as otherwise expressly provided, wherever in this Act an amendment is expressed in terms of an amendment to a section or other provision, the reference shall be considered to be made to a section or other provision of title 38, United States Code. SEC. 2. INCREASE IN RATES OF DISABILITY COMPENSATION AND DEPENDENCY AND INDEMNITY COMPENSATION. (a) Rate Adjustment.–The Secretary of Veterans Affairs shall, effective on December 1, 2003, increase the dollar amounts in effect for the payment of disability compensation and dependency and indemnity compensation by the Secretary, as specified in subsection (b). (b) Amounts To Be Increased.–The dollar amounts to be increased pursuant to subsection (a) are the following: (1) Compensation.–Each of the dollar amounts in effect under section 1114. (2) Additional compensation for dependents.–Each of the dollar amounts in effect under section 1115(1). (3) Clothing allowance.–The dollar amount in effect under section 1162. (4) New dic rates.–Each of the dollar amounts in effect under paragraphs (1) and (2) of section 1311(a). (5) Old dic rates.–Each of the dollar amounts in effect under section 1311(a)(3). (6) Additional dic for surviving spouses with minor children.–The dollar amount in effect under section 1311(b); (7) Additional dic for disability.–Each of the dollar amounts in effect under subsections (c) and (d) of section 1311. (8) DIC for dependent children.–Each of the dollar amounts in effect under sections 1313(a) and 1314. (c) Determination of Increase.– (1) The increase under subsection (a) shall be made in the dollar amounts specified in subsection (b) as in effect on November 30, 2003. (2) Except as provided in paragraph (3), each such amount shall be increased by the same percentage as the percentage by which benefit amounts payable under title II of the Social Security Act (42 U.S.C. 401 et seq.) are increased effective December 1, 2003, as a result of a determination under section 215(i) of such Act (42 U.S.C...
Gold	Veterans Programs Improvement Act of 2003 - Directs the Secretary of Veterans Affairs to increase, as of December 1, 2003, the rates of veterans' disability compensation, additional compensation for dependents, the clothing allowance for certain disabled adult children, and dependency and indemnity compensation for surviving spouses and children. Makes the effective date for the award of death pension the same as that for the award of death compensation or dependency and indemnity compensation. Excludes lump-sum insurance proceeds from income for purposes of eligibility for veterans' pensions. Prohibits the payment of veterans' disability compensation for an alcohol- or drug-abuse related disability even if the the alcohol or drug abuse is secondary to a service-connected disability. Provides alternative beneficiaries for National Service Life Insurance and United States Government Life Insurance proceeds when the first beneficiary does not make a claim. Provides burial benefit eligibility for a veteran's surviving spouse who remarries following the veteran's death. Makes permanent the authority for the State cemetery grants program. Repeals the Department of Veterans Affairs Education Loan program. Includes self-employment training under the Montgomery GI Bill.
Model	veterans programs improvement act of 2003 - directs the secretary of veterans affairs, effective on december 1, 2003, to increase the rates of disability and dependency and indemnity compensation (dic) through the department of veterans affairs (va), to: (1) increase the rates of disability compensation and dependency and indemnity compensation; (2) provide for additional compensation for dependents; (3) provide for additional compensation for dependents; (4) exclude lump-sum sales of any life insurance policy or policies on a veteran for purposes of pension benefits; (5) exclude lump-sum sales of any life insurance policy or policies on a veteran for purposes of pension benefits; (6) exclude lump-sum life insurance proceeds from the determinations of annual income for pension purposes; (7) provide for alternative beneficiaries for certain veterans' life insurance policies or policies on a veteran's service-connected disability; and (8) authorize the secretary to approve a program of self-employment on-employment in the department of veterans affairs education loan program.amends the veterans' advisory committee on education to: (1) repeal the requirement that a claimant and the claimant's representative is necessary to complete an application is not received by the secretary within one year from the date of such notification; (2) make permanent the same authority for state cemetery grants program; and (3) authorize the secretary to approve a program of self-employment on-employment in the department of america known as the department of veterans affairs.
Metrics	Rouge1: 60.71, Rouge2: 29.79, RougeL: 33.88, RougeLsum: 50.82, Summary length (tokens): 297
Top 75% example (Sorted by rougeL)	
Document	SECTION 1. SHORT TITLE. This Act may be cited as the "Cameron Gulbransen Kids and Cars Safety Act of 2003". SEC. 2. EVALUATION OF DEVICES AND TECHNOLOGY TO REDUCE CHILD INJURY AND DEATH FROM PARKED OR UNATTENDED MOTOR VEHICLES. (a) In General.–The Secretary of Transportation shall evaluate– (1) devices and technologies intended to reduce the incidence of child injury and child death occurring outside of parked motor vehicles in nontraffic, noncrash events, including backing-over incidents, that are caused by such vehicles, and determining which of those methods is the most effective; and (2) currently available technology to prevent injury and death of children left unattended inside of parked motor vehicles, including injury or death due to hyperthermia, power windows, or power sunroofs. (b) Report.–The Secretary of Transportation shall submit a report on the findings and determinations of the evaluation under this section to the Congress by not later than one year after the date of the enactment of this Act. (c) Completion of Rulemaking Regarding Power Windows.–The Secretary of Transportation shall by not later than 6 months after the submission of the report under subsection (b) complete any rulemaking begun before the date of the enactment of this Act regarding power windows and power window switches. SEC. 3. DATABASE FOR TRACKING THE NUMBER AND TYPES OF INJURIES AND DEATHS IN NONTRAFFIC, NONCRASH EVENTS. (a) Establishment.–The Secretary of Transportation shall establish a database of (or modify an existing database to include), and collect data regarding, the numbers and types of injuries and deaths in nontraffic, noncrash events involving motor vehicles. (b) Included Information.–The Secretary of Transportation shall collect and include in such database the following information: (1) The types, makes, models, and model years of motor vehicles involved in nontraffic, noncrash events. (2) Whether there was an operator of each motor vehicle in such events. (3) The age of each operator of such motor vehicles. (4) The age of each individual who suffered injury or death in such events. (5) Whether each motor vehicle had technology installed to detect individuals and objects behind it. (6...
Gold	Cameron Gulbransen Kids and Cars Safety Act of 2003 - Directs the Secretary of Transportation to: (1) evaluate devices and technologies to reduce child injuries and deaths occurring outside of parked motor vehicles in non-traffic, non-crash events or inside of parked vehicles when children are left unattended; (2) establish a database of, and collect data on, the number and types of injuries and deaths in such events; (3) evaluate technologies for detecting and preventing collisions with individuals and objects behind motor vehicles; (4) prescribe safety standards to require devices for detecting individuals and objects behind motor vehicles; and (5) prescribe safety standards for power windows and power sunroofs, including requirements for child-safe switches and auto reverse technology.
Model	tamarisk kids and cars safety act of 2003 - directs the secretary of transportation (dot) to evaluate: (1) devices and technologies intended to reduce the incidence of child injury and death occurring inside distant motor vehicles in nontraffic, noncrash events, and determine which are the most effective; and (2) currently available technology to prevent injury and death of children left behind the motor vehicles. directs the secretary to: (1) establish a database of, and collect data regarding, the number and types of injuries and deaths in nontraffic, noncrash events involving motor vehicles; and (2) prescribe motor vehicle safety standards.
Metrics	Rouge1: 63.59, Rouge2: 37.21, RougeL: 50.69, RougeLsum: 49.77, Summary length (tokens): 132
Top 95% example (Sorted by rougeL)	
Document	SECTION 1. FINDINGS. The Congress finds the following: (1) As a Member of Congress from the Tenth Congressional District of Texas, as Majority Leader of the U.S. Senate, Vice- President and President of the United States, Lyndon Baines Johnson's accomplishments in the fields of civil rights, education, and economic opportunity rank among the greatest achievements of the past half century. (2) As President, Lyndon Johnson proposed, championed, led to passage, and signed into law on August 6, 1965, the Voting Rights Act of 1965, which swept away barriers impeding millions of Americans from meaningful participation in American political life. (3) On July 30, 1965, President Johnson signed into law the Social Security Amendments Act of 1965, popularly known as Medicare, which has transformed the delivery of health care in the United States and which, along with Social Security, reduced the rate of poverty among the elderly from 28.5 percent in 1966 to 9.1 percent in 2012. (4) On July 2, 1964, President Johnson secured passage and signed into law the most sweeping civil rights legislation since Reconstruction, the Civil Rights Act of 1964, which prohibits discrimination in employment, education, and public accommodations based on race, color, religion, or national origin. (5) On November 8, 1965, President Johnson signed into law the Higher Education Act, which provided need-based financial aid to students in the form of scholarships, work-study grants, and loans, and thus made higher education more accessible to populations of persons who were previously unable to attend college because of economic circumstances. (6) On October 3, 1965, President Johnson signed into law the Immigration and Naturalization Act of 1965, which transformed the Nation's immigration system by abolishing the racially based quota system that had defined American immigration policy for four decades and replaced it with a policy whose central purpose was family reunification, with a preference for immigrants with specific skill sets. (7) According to Robert A. Caro, the preeminent biographer of Lyndon Baines Johnson, with the s...
Gold	This bill directs the Speaker of the House and the President pro tempore of the Senate to arrange for the posthumous award of a Congressional Gold Medal to Lyndon Baines Johnson in recognition of his contributions to the nation, including passage of the Voting Rights Act of 1965, the Social Security Amendments Act (Medicare) of 1965, the Civil Rights Act of 1964, the Higher Education Act of 1965, and the Immigration and Naturalization Act of 1965. Requires such medal to be given to the Lyndon Baines Johnson Library and Museum following its award, where it will be available for display and research.
Model	this bill directs the speaker of the house of representatives and the president pro tempore of the senate to arrange for the posthumous award, on behalf of congress, of a gold medal to lyrics to lyrics in recognition of his contributions to the nation, including recognition of his contributions to the nation, including recognition of the landmark voting rights act of 1965, the civil rights act of 1964, the higher education act of 1965, and the immigration and naturalization act of 1965.
Metrics	Rouge1: 72.83, Rouge2: 62.64, RougeL: 68.48, RougeLsum: 68.48, Summary length (tokens): 97

Table 12: Examples of the BillSum dataset using the model billsum-4096-1024 base diverse

Bottom 5% example (Sorted by rougeL)	
Document	in the last decade the amount of data regarding micrornas (mirs) and their target genes described in the literature has expanded tremendously . the volume of information on this new group of regulators (i.e. , mirs) has complicated attempts to integrate this data within existing metabolic and signalling networks . as regulators of gene expression in addition , a single mir can potentially regulate multiple different genes at the same time , leading to complex functional outcomes . however , from another perspective , the identification of groups of genes targeted by the same mir and the clustering of these genes within individual signalling pathways represents a means to understand the cross talk between multiple signalling networks and their role in a common biological process . the focus of this review is to summarize the validated groups of mirs functionally linked to the cross talk between tgf- , notch , and wnt signalling during the common biological process of epithelial - to - mesenchymal transition (emt) . in particular , this review will address whether the documented cross talk between these three important emt - associated pathways could be further reinforced by the identification of a signature of mirs , already depicted in the literature but not yet sharpened or clearly defined in this role . in the past years , many studies have elegantly described the role of tgf- , notch , and wnt pathways in promoting emt and emt - associated disorders including fibrosis and metastatic dissemination in cancer [16] . here we identify published and validated interactions between mirs and genes involved in tgf- , notch , and wnt signalling . this led to the discovery of a signature of 30 mirs each regulating all three pathways . we then searched for additional validated genes targeted by these 30 mirs and then further clustered these into the tgf- , notch , and wnt signalling pathways . interestingly , in our attempt to identify mirs that were common to all three of these signalling pathways , we found that the 30-mir signature strongly reinforced existing evidence supporting cross talk between these three pathways during emt . in this review we used tarbase v6.0 , the largest currently available manually curated mir target gene database , which includes targets derived from specific and high throughput experiments . using tarbase v6.0 we searched the collection of manually curated , experimentally validated mir - gene intera...
Gold	epithelial - to - mesenchymal transition (emt) is a reversible process by which cancer cells can switch from a sessile epithelial phenotype to an invasive mesenchymal state . emt enables tumor cells to become invasive , intravasate , survive in the circulation , extravasate , and colonize distant sites . paracrine heterotypic stroma - derived signals as well as paracrine homotypic or autocrine signals can mediate oncogenic emt and contribute to the acquisition of stem / progenitor cell properties , expansion of cancer stem cells , development of therapy resistance , and often lethal metastatic disease . emt is regulated by a variety of stimuli that trigger specific intracellular signalling pathways . altered microrna (mir) expression and perturbed signalling pathways have been associated with epithelial plasticity , including oncogenic emt . in this review we analyse and describe the interaction between experimentally validated mirs and their target genes in tgf- , notch , and wnt signalling pathways . interestingly , in this process , we identified a signature of 30 experimentally validated mirs and a cluster of validated target genes that seem to mediate the cross talk between tgf- , notch , and wnt signalling networks during emt and reinforce their connection to the regulation of epithelial plasticity in health and disease .
Model	micrornas (mirs) are a class of small noncoding rnas that regulate multiple gene expression at the same time, leading to complex functional outcomes. recent studies have shown that mirs regulate multiple different pathways at the same time, leading to complex functional outcomes. however, from another perspective, the identification of genes targeted by the same mir and the clustering of these genes within individual signalling pathways represents a means to understand the cross talk between multiple signalling pathways and their role in a common biological process. in this review , we summarize the published studies of mir-, notch, and wnt signalling in
Metrics	Rouge1: 20.22, Rouge2: 0.0, RougeL: 11.24, RougeLsum: 14.61, Summary length (tokens): 129
Bottom 25% example (Sorted by rougeL)	
Document	mhc class ii molecules are heterodimeric cell surface glycoproteins that bind exogenously derived antigenic peptides and present them to cd4 t cells 12 . class ii and chains are translocated into the endoplasmic reticulum (er) , where they form nonamers with invariant (ii) chain 3 . ii chain prevents the binding of immunogenic peptides due to the presence of a 14amino acid domain (clip) that occupies the peptide - binding groove of / dimers 3 . after ii degradation in the endocytic pathway , the mhc - encoded molecules hla - dm (or h2-m in the mouse) and hla - do (h2-o) facilitate the removal of clip from / dimers , allowing peptide binding 456 . ii chain has been implicated in functions such as er export , endosome targeting , and even b cell maturation 37 . two alternatively spliced ii isoforms exist (p31 and p41) , distinguished by a 64-residue domain in the lumenal portion of p41 8 . the isoforms are expressed differently in various apcs and regulate the presentation of certain antigen epitopes in b cells 9 . this difference may reflect protease inhibition by the amino acid insertion in p41 , as it has been shown to inhibit the lysosomal cysteine protease cathepsin l both in vitro and in vivo 910 . therefore , ii chain may contribute to the modulation of the proteolysis in the endocytic pathway and thus modulate antigen processing indirectly 1112 . we demonstrate here that ii chain deletion leads to the lysosomal degradation of h2-mb in apcs , suggesting that ii chain is required to prevent the proteolysis of h2-m and perhaps of other proteins . this feature may help explain how ii chain expression affects t cell selection and b cell maturation independently from its effect on mhc class ii traffic 131415 . c57bl/6 (control) and ii , ii p31 1416 , class ii , and class ii / ii mice (the gift of p. marrack , university of colorado health sciences center , denver , co) were kept in a pathogen - free environment for 78 wk before killing . splenocytes were obtained as described 7 . bone marrow derived dendritic cells (dcs) were cultured as described 17 . after purification , immature dcs were characterized by immunofluorescence and processed in parallel with the lps - treated dcs . the primers used here to detect i - a , h2-m , and h2-m are identical to the primers described previously 19 . 3 10 late dcs were pulse labeled with 7.5 mci / ml of [s]methionine translabel (icn) and chased as described 17 . ...
Gold	the association of invariant (ii) chain with major histocompatibility complex (mhc) class ii dimers is required for proper antigen presentation to t cells by antigen - presenting cells . mice lacking ii chain have severe abnormalities in class ii transport , t cell selection , and b cell maturation . we demonstrate here that h2-m , which is required for efficient class ii antigenic peptide loading , is unexpectedly downregulated in splenocytes and mature dendritic cells (dcs) from ii/ mice . downregulation reflects an increased rate of degradation in ii/ cells . degradation apparently occurs within lysosomes , as it is prevented by cysteine protease inhibitors such as e64 , but not by the proteasome inhibitor lactacystin . thus , ii chain may act as a lysosomal protease inhibitor in b cells and dcs , with its deletion contributing indirectly to the loss of h2-m .
Model	mhc class ii molecules are heterodimeric cell surface glycoproteins that bind exogenously derived antigenic peptides and present them to cd4 t cells 12. type ii chains are translocated into the endoplasmic reticulum (er), where they form nonamers with invariant (ii) chain 3. type ii chains are translocated into the endoplasmic reticulum (er), where they form nonamers with invariant (ii) chain 3. type ii chains are translocated into the endoplasmic reticulum (er), where
Metrics	Rouge1: 27.67, Rouge2: 8.92, RougeL: 17.61, RougeLsum: 25.16, Summary length (tokens): 129
Top 75% example (Sorted by rougeL)	
Document	stroke , after myocardial infarction (mi) , is the second leading reason for mortality in iran as with many countries worldwide . the epidemiology of stroke has already been investigated in the american , european , african , and asian countries . no comprehensive study has yet investigated the epidemiology of stroke , particularly in mi patients , in iran , one of the largest countries in southwest asia . stroke and mi share many risk factors , most prevalent of which are smoking , dyslipidemia , type 2 diabetes , and hypertension . the risk factors for stroke and mi , especially smoking , hypertension , and dyslipidemia are highly prevalent in iran , as well . according to projections urbanism , increased life expectancy , reduction in childbirth , aging and elderly population , epidemiological changes , socioeconomic status , geographical conditions , and lifestyles such as poor diet , stress , and low mobility are the main causes of the burden of noncommunicable diseases , particularly stroke . because the determinants of stroke in different communities are various , we require knowledge about the risk factors and determinants of mortality in a community for effective planning and selection of appropriate strategies for the prevention and management of stroke and heart attack as the most important causes of death . since no comprehensive study has yet been investigated the status and mortality determinants of stroke in mi patients in iran , this study is conducted to determine and compare the determinants of mortality due to stroke in mi patients . in this retrospective cohort study , the data obtained from the mi registry of iran 's cardiovascular diseases surveillance system were analyzed . around 20,750 hospitalized patients with mi with a new presentation (hospitalized in 540 hospitals) between april , 2012 and march , 2013 were enrolled . the study was approved by the management center of noncommunicable diseases and the department of cardiovascular diseases prevention of iran 's ministry of health and medical education (approval no . the research followed the principles of the declaration of helsinki ; the researchers did not conduct any interventions on the patients , and an institutional review board approved this research . the study protocol was approved by an independent scientific review committee (at shahid beheshti university medical of sciences) (no . inclusion criteria were determined according to the world...
Gold	background : the data and determinants of mortality due to stroke in myocardial infarction (mi) patients are unknown . this study was conducted to evaluate the differences in risk factors for hospital mortality among mi patients with and without stroke history.materials and methods : this study was a retrospective , cohort study ; 20,750 new patients with mi from april , 2012 to march , 2013 were followed up and their data were analyzed according to having or not having the stroke history . stroke and mi were defined based on the world health organization 's definition . the data were analyzed by logistic regression in stata software.results:of the 20,750 studied patients , 4293 had stroke history . the prevalence of stroke in the studied population was derived 20.96% (confidence interval [ci] 95% : 20.1321.24) . of the patients , 2537 (59.1%) had st - elevation mi (stemi) . mortality ratio in patients with and without stroke was obtained 18.8% and 10.3% , respectively . the prevalence of risk factors in mi patients with and without a stroke is various . the adjusted odds ratio of mortality in patients with stroke history was derived 7.02 (95% ci : 5.429) for chest pain resistant to treatment , 2.39 (95% ci : 1.972.9) for stemi , 3.02 (95% ci : 2.53.64) for lack of thrombolytic therapy , 2.2 (95% ci : 1.662.91) for heart failure , and 2.17 (95% ci : 1.62.9) for ventricular tachycardia.conclusion:with regards to the factors associated with mortality in this study , it is particularly necessary to control the mortality in mi patients with stroke history . more emphasis should be placed on the mi patients with the previous stroke over those without in the interventions developed for prevention and treatment , and for the prevention of avoidable mortalities .
Model	background : stroke and myocardial infarction (mi) share many risk factors, most prevalent of which are smoking, dyslipidemia, type 2 diabetes, and hypertension. the aim of this study was to determine and compare the determinants of stroke in mi patients.methods:in this retrospective cohort study, the data obtained from the mi registry of iran's cardiovascular diseases surveillance system were analyzed. patients with mi were defined by the date at mi diagnosis, hospital stay, and follow - up till discharge or death (outcome). odds ratio (or) of mortality for clinical and demographic risk factors were calculated by logistic regression.results
Metrics	Rouge1: 55.88, Rouge2: 20.79, RougeL: 31.37, RougeLsum: 51.96, Summary length (tokens): 129
Top 95% example (Sorted by rougeL)	
Document	stroke commonly causes loss of motor function due to weakening of upper / lower extremity muscles1 . according to ryerson2 , use the affected upper extremity decreases because of the patient s dependency on the unaffected upper extremity for normal functions , which results in problems such as learned disuse , asymmetric postural patterns , contractures , and aggravated functional restrictions involving the affected upper extremity . therefore , to improve functions of the affected upper extremity in stroke patients , measures that maximize opportunities to use the affectedupper extremity are necessary . bilateral activities have been discussed as measures to improve the body symmetry and to reduce abnormal muscle tone3 , thereby promoting voluntary movement of the affected upper extremity4 . thus far , bilateral upper extremity coordination movements have been applied in the form of bilateral single exercises utilizing tasks such as figure imitation5 , robot arm upper extremity mechanisms6 and functional stretching7 , and bilateral complex exercises combined with the principle of motor learning , such as rhythmic acoustic8 , unaffected extremity weight addition9 , and active neuromuscular electrical stimulation in stroke patients10 . most previous studies have reported the positive effects of these exercises on motor function recovery in stroke patients . whitall et al.8 reported that when chronic hemiplegia patients underwent bilateral training to push and pull upper extremity apparatuses , including acoustic signals , their upper extremity functions were improved . in a study conducted by summer et al.7 with 12 chronic stroke patients , the movement time of the patients affected side upper extremity decreased and upper extremity functions showed more improvement when they performed tasks to hold objects by stretching both hands simultaneously then when they performed tasks to hold objects by stretching only one hand . however , lee su - jin11 advised that the tasks in those studies involved mostly gross motor exercises of the upper extremity and simple repetitive training tasks , and that they were not sufficient to improve patients grasping ability to hold objects with various sizes , shapes , and weights using the shoulder , elbow , wrist , and fingers . although studies that applied diverse functional training tasks including activities of daily living to symmetric bilateral upper ...
Gold	[purpose] the purpose of this study was to investigate the effects of task - oriented bilateral arm training and repetitive bilateral arm training on upper limb function and activities of daily living in stroke patients . [subjects] forty patients with hemiplegia resulting from stroke were divided into a task orientied bilateral arm training group (n=20) and a repetitive bilateral arm training group (n=20) . [methods] the task - oriented group underwent bilateral arm training with 5 functional tasks , and the repetitive group underwent bilateral arm training with rhythmin auditory cueing for 30 minutes / day , 5 times / week , for 12 weeks . [results] the upper limb function and the ability to perform activities of daily living improved significantly in both groups . although there were significant differences between the groups , the task - oriented group showed greater improvement in upper limb function and activities of daily living . [conclusion] we recommend bilateral arm training as well as adding functional task training as a clinical intervention to improve upper limb function activities of daily living in patients with hemiplegia .
Model	[purpose] the purpose of this study was to compare the effects of symmetric bilateral upper extremity exercises using diverse functional training tasks with the effects of repetitive bilateral upper extremity exercises including acoustic signals on the recovery of upper extremity function and daily living activities in stroke paitents. [subjects and methods] forty patients with hemiplegia resulting from stroke were randomly and equally assigned to a task - oriented bilateral arm training group (tbg : 11 males, 9 females) and a repetitive bilateral arm training group (rbg : 13 males, 7 females). both groups performed their exercises for
Metrics	Rouge1: 62.03, Rouge2: 34.59, RougeL: 50.27, RougeLsum: 60.96, Summary length (tokens): 129

Table 13: Examples of the PubMed dataset using the model pubmed-1024-128 small diverse

Bottom 5% example (Sorted by rougeL)	
Document	cdnas containing the sequence of human wt - ftl and human mutant ftl498499instc were introduced into the pet-28a(+) expression vector (novagen , emd chemicals inc .) . the cdnas were cloned between the bamhi and xhoi sites , downstream from and in - frame with the sequence encoding an n - terminal his6 tag . to eliminate the his6 tag (included in the expression vector) , the sequence of the vector was modified by introducing the recognition sequence for cleavage by factor xa before the coding sequence of the ferritin genes . pcr amplification of the ferritin cdnas was performed using the upstream primer f1 5-tgg atc cat cga agg tcg tat gag ctc cca gat t-3 and the downstream primer r1 5-tta tgc ctc gag ccc tat tac ttt gca agg-3. f1 contains the factor xa sequence (underlined) . pet-28a(+) carrying wt - ftl and mt - ftl cdnas was transformed into bl21 (de3) escherichia coli (invitrogen) . transformed cells were grown in luria broth medium (lb) containing 30 g / ml kanamycin (invitrogen) at 37 c up to an absorbance of 0.91.0 at 600 nm . bacteria were induced to overexpress recombinant proteins by adding 1 mm isopropyl thio–d - galactopyranoside (icn biotechnologies) for 12 h at 25 c . purification of recombinant wt- and mt - ftl homopolymers cells were harvested by centrifugation and frozen at -80 c . the cell pellets were suspended in 50 mm sodium phosphate , 500 mm nacl (ph 7.4) , 1 mg / ml lysozyme , and a protease inhibitor mixture (complete , roche applied science) for 30 min . bacteria were disrupted by sonication , and the insoluble material was removed by centrifugation at 21,000 g for 30 min . the soluble fraction was purified by nickel iminodiacetic acid affinity chromatography using an akta purifier system (ge healthcare) . purified protein was eluted with 250 mm imidazole in 50 mm sodium phosphate (ph 7.4) , 0.5 m nacl . recombinant proteins were diluted with 50 mm tris and 10% glycerol (v / v) down to an absorbance of 0.5 at 280 nm , and ferritins were cleaved from the his tag by digestion with factor xa protease (ge healthcare) (5 units / mg of protein) . after being dialyzed against 50 mm tris , ph 8.0 , for 18 h , proteins were further purified by anion exchange chromatography (mono q) using a linear nacl elution gradient in 50 mm tris (ph 8) . peak fractions were 95% pure based on s...
Gold	nucleotide insertions in the ferritin light chain (ftl) polypeptide gene cause hereditary ferritinopathy , a neurodegenerative disease characterized by abnormal accumulation of ferritin and iron in the central nervous system . here we describe for the first time the protein structure and iron storage function of the ftl mutant p.phe167serfsx26 (mt - ftl) , which has a c terminus altered in sequence and extended in length . mt - ftl polypeptides assembled spontaneously into soluble , spherical 24-mers that were ultrastructurally indistinguishable from those of the wild type . far - uv cd showed a decrease in -helical content , and 8-anilino-1-naphthalenesulfonate fluorescence revealed the appearance of hydrophobic binding sites . near - uv cd and proteolysis studies suggested little or no structural alteration outside of the c - terminal region . in contrast to wild type , mt - ftl homopolymers precipitated at much lower iron loading , had a diminished capacity to incorporate iron , and were less thermostable . however , precipitation was significantly reversed by addition of iron chelators both in vitro and in vivo . our results reveal substantial protein conformational changes localized at the 4-fold pore of mt - ftl homopolymers and imply that the c terminus of the mt - ftl polypeptide plays an important role in ferritin solubility , stability , and iron management . we propose that the protrusion of some portion of the c terminus above the spherical shell allows it to cross - link with other mutant polypeptides through iron bridging , leading to enhanced mutant precipitation by iron . our data suggest that hereditary ferritinopathy pathogenesis is likely to result from a combination of reduction in iron storage function and enhanced toxicity associated with iron - induced ferritin aggregates .
Model	apoferrin is a member of the nuclear receptor superfamily family of membrane proteins. apoferrin is a member of the family of apoferrin family. apoferrin is a member of the family of apoferrin family. apoferrin is a member of the family of apoferrin family. apoferrin is a member of the family of apoferrin family. apoferrin is a member of the family of apoferrin family. apoferrin is a member of the family of
Metrics	Rouge1: 13.33, Rouge2: 1.35, RougeL: 12.0, RougeLsum: 12.0, Summary length (tokens): 129
Bottom 25% example (Sorted by rougeL)	
Document	intracellular signaling pathways transmit signals of various extracellular stimuli to their cytosolic and nuclear targets in order to induce biological responses , such as proliferation , differentiation , cell death and migration . when needed , the signals are transmitted from the cytoplasm to the nucleus via translocation of one or more components of each of the signaling pathways involved . thus , after stimulation , a large number of signaling proteins are rapidly translocated to the nucleus to induce and regulate many nuclear processes . however , despite the importance of stimulated nuclear signaling , the mechanisms by which these components reach the nucleus upon stimulation have been elucidated only for a few signaling pathways . classic nuclear shuttling is mediated by an importin- complex that binds to cargoes containing a nuclear localization signal (nls) , consisting of mono- or bi - partite clusters of basic amino acids [1 - 3] . this importin- complex often acts as a housekeeping mechanism that shuttles most nuclear proteins immediately to the nucleus after their translation . the relocalization of cargoes is followed by the dissociation of the proteins from the importins upon binding to rangtp , which exports the importins back to the cytoplasm , while the cargo remains in the nucleus . however , only a limited number of signaling proteins , such as nfb and erk5 (extracellular signal - regulated kinase 5) [8 - 10] , use this machinery for their stimulated nuclear shuttle . aside from this canonical mechanism , importin- or similar karyopherins , termed -like importins , operate by binding to non - canonical nlss to mediate translocation without the assistance of importin-. the mechanism by which these components operate is much less understood and is currently under investigation . recently , several -like importins have been implicated in the stimulated nuclear translocation of signaling proteins . here , we propose that this group of importins might be responsible for the swift nuclear shuttling of many proteins following various stimuli . the rapid and robust activation of erk1/2 allows the phosphorylation and modulation of the activity of more than 300 proteins , which are localized either in the cytoplasm or the nucleus [13 - 15] . these substrates are important for the induction and regulation of cellular processes , including proliferation , differentiation , and migration amongst others [16 - ...
Gold	classic nuclear shuttling is mediated by an importin- heterodimer that binds to cargoes containing a nuclear localization signal , and shuttles most nuclear proteins immediately after their translation . aside from this canonical mechanism , kariopheryn-s or -like importins operate by binding to non - canonical nuclear localization signals to mediate translocation without the assistance of importin-. the mechanism by which these components operate is much less understood and is currently under investigation . recently , several -like importins have been implicated in the stimulated nuclear translocation of signaling proteins . here , we propose that this group of importins might be responsible for the swift nuclear shuttling of many proteins following various stimuli .
Model	the nuclear translocation signal (nl) is a key regulator of many cellular processes, including proliferation, differentiation, cell death, migration, and migration. recent studies have shown that the nuclear translocation signal (nl) is required for the induction and regulation of many nuclear processes, including proliferation, differentiation, migration, and migration. here , we discuss the role of nl in the regulation of the nuclear translocation signal (nl), as well as the mechanisms that regulate it.
Metrics	Rouge1: 33.33, Rouge2: 5.0, RougeL: 18.52, RougeLsum: 32.1, Summary length (tokens): 100
Top 75% example (Sorted by rougeL)	
Document	helicobacter pylori is a stomach bacterium that colonizes ~50% of people globally.1 h. pylori is the primary risk factor for gastric cancer the third highest cause of global cancer morbidity.2 h. pylori infection rates are highly dependent on socioeconomic status ; ~80% of those living in low socioeconomic areas of latin america , asia , and eastern europe are infected , compared with < 20% of asymptomatic caucasians in the usa.3 h. pylori infection is treatable with different regimens of antibiotics,4 and eradication of h. pylori is a recognized way to lower incidence of gastric cancer.5 however , recurrence of infection is variable,6,7 and the emergence of antibiotic resistance compromises treatment efficacy . thus , determining the best course of treatment is important to improve treatment efficacy and to reduce recurrence of h. pylori infection . unfortunately , there is no broad consensus about an optimal antibiotic therapy for the treatment of h. pylori . for example , meta - analyses of european and asian clinical data compared the standard triple therapy (amoxicillin , clarithromycin , and a proton - pump inhibitor for 714 days) with 5- or 10-day quadruple therapy regimens (adding metronidazole or tinidazole to the triple therapy) and found that quadruple therapies are both significantly more effective and cheaper than the triple therapy.810 however , we previously published a study comparing eradication therapies in seven sites of six latin american countries that showed that the 14-day triple therapy was superior to the 5-day concomitant quadruple therapy , and no different than the 10-day sequential quadruple therapy.11,12 these inconsistencies reflect localized differences in antibiotic use practices , such as the use of clarithromycin for upper respiratory infections.13 the differences in efficacy of antibiotic therapy are supported by primary antibiotic resistance data . for example , h. pylori resistance to amoxicillin varied widely between africa (65.6%) , europe (0.5%) , asia (11.6%) , and the americas (2.2%).12 even in the same region , patterns of resistance differ : within central and latin america , reported average metronidazole resistance varies from 30% in argentina to 83% in columbia , and tetracycline resistance varies from 2% in brazil to 33% in columbia.14 as such , characterizing local resistance patterns is important for selecting therapies with the highest likelihood of success . our research focus...
Gold	objectivesgastric carcinoma is the most common cancer and cause of cancer mortality in peru . helicobacter pylori , a bacterium that colonizes the human stomach , is a group 1 carcinogen due to its causal relationship to gastric carcinoma . while eradication of h. pylori can help prevent gastric cancer , characterizing regional antibiotic resistance patterns is necessary to determine targeted treatment for each region . thus , we examined primary antibiotic resistance in clinical isolates of h. pylori in lima , peru.materials and methodsh . pylori strains were isolated from gastric biopsies of patients with histologically proven h. pylori infection . primary antibiotic resistance among isolates was examined using e - test strips . isolates were examined for the presence of the caga pathogenicity island and the vaca m1/m2 alleles via polymerase chain reaction.resultsseventy-six isolates were recovered from gastric biopsies . clinical isolates showed evidence of antibiotic resistance to 1 (27.6% , n=21/76) , 2 (28.9% , n=22/76) , or 3 antibiotics (40.8%) . of 76 isolates , eight (10.5%) were resistant to amoxicillin and clarithromycin , which are part of the standard triple therapy for h. pylori infection . no trends were seen between the presence of caga , vaca m1 , or vaca m2 and antibiotic resistance.conclusionthe rate of antibiotic resistance among h. pylori isolates in lima , peru , is higher than expected and presents cause for concern . to develop more targeted eradication therapies for h. pylori in peru , more research is needed to better characterize antibiotic resistance among a larger number of clinical isolates prospectively .
Model	helicobacter pylori infection is the third most common cause of global cancer morbidity. eradication of helicobacter pylori is a recognized way to lower incidence of gastric cancer. however, there is no broad consensus about an optimal antibiotic therapy for the treatment of helicobacter pylori. we assessed primary antibiotic resistance to helicobacter pylori among 76 isolates from a cohort of patients recruited in lima, peru. primary antibiotic resistance to metronidazole, amoxicillin, tetracycline, clarithromycin, levofloxacin, and rifamp
Metrics	Rouge1: 50.0, Rouge2: 17.28, RougeL: 32.93, RougeLsum: 47.56, Summary length (tokens): 129
Top 95% example (Sorted by rougeL)	
Document	in general , brain injury can occur due to sudden and severe head strike to a hard object , which can be mild , moderate or severe (1) . the main causes of head injury include traffic accidents , falling from heights , physical violence , accidents at work , inside home accidents and during exercise incidents . however , the most important cause of head trauma in iranian population is traffic accident (2) . among the warning signs of head trauma are nausea , vomiting , dizziness , headache , blurred vision , and loss of balance , difficulty in sleeping , memory problems , tinnitus and fatigue (3) . nausea and vomiting are the most common complications after minor head trauma that in addition to severe harassment of patients increases the risk of aspiration and intracranial pressure rising . ondansetron is a serotonin 5-ht3 receptor antagonist , which connects to the peripheral and central receptors of serotonin (1) . this drug is mostly used in nausea and vomiting after chemotherapy and surgery (2) . it does not have any effect on dopamine receptors thus ; it does not have extra pyramidal effect (3) . this drug has a half - life of 2 - 7 hours and is metabolized in the liver where it changes into glucuronide and sulfate which is inactive . its most common side effects include headaches , fatigue , diarrhea , constipation , dizziness and anxiety . the recommended dose for the treatment of nausea and vomiting is 4 - 8 milligrams (4 , 5) . metoclopramide as an old antiemetic is mostly used in high doses , before chemotherapy and for nausea and vomiting caused by various reasons (6 - 8) . this drug blocks the dopamine receptors on the peripheral and central dopamine receptors and increases the movement of the upper gastrointestinal tract without increasing secretion (9 , 10) . its intravenous absorption takes about 3 minutes and the peak of its effect is about 15 minute . this drug is metabolized in the liver and its half - life is approximately 4 - 5 hours (11) . its most common side effects include dystonia < 10% , fatigue , drowsiness , and flushing . based on the above - mentioned reasons , the present study was aimed to compare the antiemetic effects of metoclopramide and ondansetron in the treatment of post head trauma nausea and vomiting . study design and setting the study was a controlled , randomized , double blind clinical trial , which was conducted in the first 6 months of 2014 in al - zahra a...
Gold	introduction : nausea and vomiting are the most common complications after minor head trauma that increases the risk of intracranial pressure rising . therefore , the present study was aimed to compare the antiemetic effects of metoclopramide and ondansetron in the treatment of post - traumatic nausea and vomiting . methods : the study was a controlled , randomized , double blind clinical trial , which was conducted in the first 6 months of 2014 in emergency department al - zahra and kashani hospitals in isfahan , iran . the patients with minor head trauma associated with nausea and vomiting were randomly divided into 2 groups : treatment with metoclopramide (10mg/2ml , slow injection) and treatment with ondansetron (4mg/2ml , slow injection) . the comparison between the 2 groups was done regarding antiemetic efficacy and side effects using spss 21 statistical software . results:120 patients with minor head trauma were distributed and studied into two groups of 60 patients (mean age 35.614.1 years ; 50.0% male) . administration of both ondansetron and metoclopramide significantly reduced the severity of nausea (p<0.001) . changes in the severity of nausea in both groups before and after the treatment revealed that nausea had been decreased significantly in both groups (p < 0.001) . the incidence of fatigue (p=0.44) , headache (p=0.58) and dystonia (p=0.06) had no significant difference in the two groups but the incidence of drowsiness and anxiety in the metoclopramide group was significantly higher (p < 0.001) . conclusion : the present study indicated that the treatment effectiveness of ondansetron and metoclopramide are similar . however , incidence of drowsiness and anxiety in the metoclopramide was considerably higher . since these complications can have adverse effects on the treatment of patients with brain injury , it is suggested that it may be better to use ondansetron in these patients .
Model	background : head injury can occur due to sudden and severe head strike to a hard object, which can be mild, moderate or severe. metoclopramide as an old antiemetic is mostly used in high doses, before chemotherapy and for nausea and vomiting caused by various reasons. the present study was aimed to compare the antiemetic effects of metoclopramide and ondansetron in the treatment of post head trauma nausea and vomiting. methods : this randomized, double blind clinical trial was conducted in the first 6 months of 2014 in isfahan, iran. the patients were randomly divided into
Metrics	Rouge1: 59.69, Rouge2: 46.56, RougeL: 51.31, RougeLsum: 58.64, Summary length (tokens): 129

Table 14: Examples of the PubMed dataset using the model pubmed-1024-128 base diverse

Bottom 5% example (Sorted by rougeL)	
Document	this study is an extension of a report on patients with type 1 diabetes at children 's hospital of new orleans (14) and was approved by the institutional review board at louisiana state university health sciences center , new orleans , louisiana . glucose data were downloaded from patient meters at each clinic visit . meter model and sampling protocols varied by patient preference and insurance provider . an average of three glucose measurements per day were recorded in a study using a similar self - monitoring protocol (7) . a1c was measured by national glycohemoglobin standardization program (ngsp)-approved immunoassays (15) at the children 's hospital (184 patients) or by commercial laboratories that presumably also used ngsp - approved methods (18 patients , including 4 low- , 7 moderate- , and 7 high - hgi subjects) . a population regression equation { a1c (%) = [0.021 mbg (mg / dl)] + 4.3 , r = 0.57 } was derived using mean mbg and mean a1c from 202 patients collected at 1,612 clinic visits as described elsewhere (14) . the same data were used to calculate hemoglobin glycation index (hgi) and to divide patients into low- , moderate- , and high - hgi groups . predicted a1c values were calculated at each clinic visit by inserting mbg into the regression equation . hgi values were calculated by subtracting predicted a1c from observed a1c measured at the same clinic visit . patients were divided into low- , moderate- , and high - hgi groups based on mean hgi tertile (33%) rank (low hgi , < 0.41 , n = 67 ; moderate hgi , 0.41 to 0.26 , n = 68 ; high hgi , > 0.26 , n = 67) . eag was calculated by inserting observed a1c into the adag linear regression equation (eag [mg / dl] = [28.7 a1c (%)] 46.7 , r = 0.92) (1) . a mean blood glucose index (mbgi) that quantifies the difference between mbg and eag was calculated by subtracting observed mbg from eag . descriptive statistics and linear regression analyses were generated using graphpad prism v. 4.03 (graphpad software , san diego , ca) . glucose data were downloaded from patient meters at each clinic visit . meter model and sampling protocols varied by patient preference and insurance provider . an average of three glucose measurements per day were recorded in a study using a similar self - monitoring protocol (7) . a1c was measured by national glycohemoglobin standardization program (ngsp)-approved immunoassays (15) at the children 's ho...
Gold	objectivethe a1c - derived average glucose study recommended reporting a1c in estimated average glucose (eag) equivalents . we compared eag with self - monitored mean blood glucose (mbg) to determine whether eag is systematically biased due to biological variation in the relationship between mbg and a1c.research design and methodsmbg and a1c were recorded from charts of 202 pediatric type 1 diabetic patients at 1,612 clinic visits . patients were divided into groups with low , moderate , or high a1c bias based on a hemoglobin glycation index (hgi).resultsthe mean sd values for mbg versus eag were as follows : total population , 194 34 vs. 196 36 mg / dl ; low - hgi group , 186 31 vs. 163 20 mg / dl ; moderate - hgi group , 195 28 vs. 193 19 mg / dl ; and high - hgi group , 199 42 vs. 230 31 mg / dl.conclusionseag underestimated mbg in low hgi patients and overestimated mbg in high hgi patients . disagreement between eag and mbg downloaded from patient glucose meters will cause confusion if eag is implemented for clinical use .
Model	objectiveto investigate the relationship between hemoglobin glycated hemoglobin (hba1c), hemoglobin glycated hemoglobin (hba1c), and hemoglobin glycated hemoglobin (hba1c) in children with type 1 diabetes.research design and methodsa1c was measured by national glycohemoglobin standardization program (ngsp)-approved immunoassays at the children's hospital of new orleans. hba1c was measured by national glycohemoglobin standardization program (ngsp)-approved immunoassays at the children's hospital of new orleans. hba1c was measured by national glycohemoglobin standardization program (ngsp)-approved immunoassays at the children's hospital of new orleans. a population regression equation [a1c (%) = [0.021 mbg (mg / dl] + 4.3, r = 0.92) + 4.3, r = 0.58] was derived using mean hba1c (%). a population regression equation [a1c (%) = [0.021 mbg (mg / dl] + 4.3, r = 0.92) + 4.3, r = 0.92] was
Metrics	Rouge1: 19.4, Rouge2: 6.06, RougeL: 10.03, RougeLsum: 13.38, Summary length (tokens): 256
Bottom 25% example (Sorted by rougeL)	
Document	the principal aim of this laboratory is the synthesis of conjugated unsaturated ketones as candidate antineoplastic agents . these compounds interact with thiols but in general , they have little or no affinity for amino and hydroxyl groups which are found in nucleic acids (1 - 3) . hence thiol alkylators may not have the genotoxic properties associated with a number of contemporary anticancer drugs (4) . however after an initial chemical insult , certain neoplasms are more vulnerable to a subsequent cytotoxic effect than various non - malignant cells (5 , 6) . hence by mounting the 1,5-diaryl-3-oxo-1,4-pentadienyl pharmacophore [ar - c = c - c(o)-c = c - ar] on heterocyclic and cycloaliphatic scaffolds , two sequential interactions with cellular thiols can take place which may be more detrimental to tumours than normal tissues . such considerations led to the development of 3,5-bis(benzylidene)-4-piperidones 1a - d which demonstrated potent cytotoxic properties with the ic50 values in the low micromolar range against human molt 4/c8 and cem t - lymphocytes as well as murine l1210 lymphocytic leukemia cells (7 , 8) . the hypothesis of sequential cytotoxicity was advanced that the 1,5-diaryl-3-oxo-1,4-pentadienyl group interacts at a primary binding site and a side chain on the piperidine nitrogen may align at an auxiliary binding site which could enhance cytotoxic potencies . in order to evaluate this hypothesis , a novel series of n - aroyl-3,5-bis(benzylidene)-4-piperidone derivatives 2 - 7 were synthesized (8) . in these compounds , the side chain contains atoms and groups capable of forming hydrogen and van der waals bonds and also an ionic bond in series 3 - 6 which could bind at an auxiliary binding site . when the potencies of the compounds in series 3 - 6 were compared with that of the analogs 1a - d which have the same aryl substituent , the ic50 values of the amides 3 - 6 were lower in 48% of the comparisons made while equipotency was noted in 35% of the cases (8) . the proposed primary and auxiliary binding sites of cytotoxic n - aroyl-3,5-bis(benzylidene)-4-piperidone analogs the objectives of the present investigation are two folds . first , to gain some insight into the physicochemical properties of the 3,5-bis(benzylidene)-4-piperidone derivatives 1 - 7 that influence cytotoxic potencies , and second , to derive predictive 2d - qsar models that will be helpful to design new analogs with improved bioactivity . t...
Gold	1-[4-(2-alkylaminoethoxy) phenylcarbonyl]-3,5-bis(arylidene)-4-piperidones are a novel class of potent cytotoxic agents . these compounds demonstrate low micromolar to submicromolar ic50 values against human molt 4/c8 and cem t - lymphocytes and murine leukemia l1210 cells . in this study , a comparative qsar investigation was performed on a series of 3,5-bis (arylidene)-4-piperidones using different chemometric tools to develop the best predictive models for further development of analogs with improved cytotoxicity . all the qsar models were validated by internal validation tests . the qsar models obtained by ga - pls method were considered the best as compared to mlr method . the best qsar model obtained by ga - pls analysis on l1210 , cem and molt4/c8 demonstrated good predictively with r2pred values ranging from 0.94 - 0.80 . molecular density , topological (x2a) and geometrical indices of the molecules were found to be the most important factors for determining cytotoxic properties .
Model	in order to evaluate the physicochemical properties of the 3,5-bis(benzylidene)-4-piperidone derivatives 1 - 7 and 2 - 7 were synthesized. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6. the compounds 1 - 7 showed potent cytotoxic property with ic50 values of 3 - 6.
Metrics	Rouge1: 21.38, Rouge2: 5.06, RougeL: 16.35, RougeLsum: 16.35, Summary length (tokens): 253
Top 75% example (Sorted by rougeL)	
Document	educational evaluation (ee) is a formal process performed to evaluate the quality of effectiveness and/or value of a program , process , goal or curriculum.12 it deals with data collection and assessment of the progress of academic programs.34 by considering some principles related to educational measurements and data collection , ee may result in a better understanding of such programs.57 during the past thirty years , theorists have presented numerous methods of evaluation . worthen and sanders2 mentioned that more than 50 different evaluation approaches has been developed in recent decades . among these , methods based on internal criteria are known as the ones that can interpret the scientific , educational , and therapeutic authenticity of different educational groups.48 this is greatly welcomed by the academic community and is widely spread to all universities in the world . that is because this method provided a scientific , appropriate , precise , timely , and valid basis regarding the interpretation of decision making system quality and programming for its promotion and development.3 such a method was successfully carried out in four medical education groups at supervisory and expansion of medical education council secretariat of ministry of health , treatment and medical education of iran in 1995.8 ee has its most effect , value , and results when it can provide needed information to individuals which are directly related , as well as those who may be benefited from its results.346 educating dental professionals consists of theoretical and practical (clinical , paraclinical , and laboratory) courses , differing in duration , and educational curriculum among different countries . it might vary from 4 years (e.g. , in india , turkey , and russia) to 6 years or more (e.g. , in iran consisting of 2 years of only basic medical sciences and 4 + years of dentistry courses) . due to numerous practical educational units in dentistry education and with regard to expensive but very critical protocols for infection control , a great deal of resources is consumed in governmental universities of iran over training every general practitioner with a degree of doctorate of dental surgery . on the other hand , the quality of dental services plays an important role in public health . considering these issues , dentistry education needs to be cost - effective in terms of optimizing its quality . to date , the pace of advancements in dentistry...
Gold	background : educational evaluation is a process which deals with data collection and assessment of academic activities progress . in this research , educational evaluation of dentistry school of tehran university of medical sciences , which trains students in undergraduate and residency courses , was studied.methods:this descriptive study was done with a model of educational evaluation in ten steps and 13 fields including purposes and mission objectives , management and organization , academic board members , students , human resources and support , educational , research , health and treatment spaces , educational , diagnostic , research and laboratory tools , educational , research , health and treatment programs and courses , process of teaching and learning , evaluation and assessment , alumni , and patients satisfaction . data were collected using observation , interviews , questionnaires , and checklists.results:results of the study were mainly qualitative and in some cases quantitative , based on defined optimal situation . the total mean of qualitative results of educational evaluation of dentistry school in all 13 fields was 55.98% which is relatively desirable . in the case of quantitative ones , results of some fields such as treatment quality of patients and education and learning of the students were relatively desirable (61.32% and 60.16% respectively).conclusion : according to the results , educational goals and missions , educational and research facilities and spaces which were identified as the weakest areas need to be considered and paid more serious attention .
Model	background : educational evaluation (ee) is a formal process performed to evaluate the quality of effectiveness and/or value of a program, process, goal or curriculum. it deals with data collection and assessment of the progress of academic programs. the objective of this descriptive cross - sectional study was to assess the quality of education provided at dental school of tehran university of medical sciences.materials and methods : in this descriptive cross - sectional study, the efficacy of provided educational programs provided at dental school of tehran university of medical sciences educational groups in 2010 through this descriptive cross - sectional study was assessed in 13 fields. the data were analyzed using spss version 20.results:the overall quality of education provided in 13 fields. the most important quality of education was the quality of education. the most important quality of education was the quality of education. the most important quality of education was the quality of education. the most important quality of education was the quality of education. the most important quality of education was the quality of education. the most important quality of education was the quality of education. the most important quality of education was the quality of education.
Metrics	Rouge1: 39.32, Rouge2: 12.68, RougeL: 27.67, RougeLsum: 34.47, Summary length (tokens): 256
Top 95% example (Sorted by rougeL)	
Document	the study population included patients over 18 years old who had an initial cabg or combined cabg and open chest aortic valve replacement (avr) , from april 1 , 1998 to october 31 , 2011 in ontario . the date of first cardiac surgery was the index date , and eligible patients were followed for 1 year with respect to major outcomes , and 5 years for mortality . preoperative data were included for 1 year prior to surgery , and outcomes for 1 year postoperatively . patients for whom sex , age , height , weight were missing , and patients living outside of ontario or of unknown residence were excluded . cardiac care network of ontario (ccn) data were used to identify baseline characteristics such as cardiac ejection fraction , number of grafts bypassed , prior myocardial infarction (mi) , emergency or elective surgery , and other co - morbidities . ccn data and the following datasets were combined from ices using deterministic linkage by unique ices key number identifiers : ontario health insurance plan , canadian institute of health information (cihi) discharge abstract database , national ambulatory care reporting system , same day surgery , and the registered persons database . patients who had undergone either isolated cabg or combined cabg / avr were selected from the cihi discharge abstract database . data for which other cardiac procedures had been performed during the same admission were excluded (eg , percutaneous coronary intervention or other valve procedures) . bmi was calculated as weight (kg)/height (m) , and patients were divided into groups : underweight (bmi < 20 kg / m) , normal weight (bmi 20.0 to 24.9 kg / m) , overweight (bmi 25.0 to 29.9 kg / m) , obese (bmi 30.0 to 34.9 kg / m) , and morbidly obese (bmi > 34.9 kg / m) , closely based on world health organization (who) and health canada guidelines.1214 the following comorbidities were assessed for presence within 1 year prior to index date : diabetes , smoking history (current or ever smoked) , peripheral vascular disease (pvd) , chronic obstructive pulmonary disease (copd) , dialysis within 1 year prior to surgery , cerebrovascular disease (cvd) , congestive heart failure (chf , from cihi) , hypertension , elective or emergent surgery , creatinine . within 30 days prior to surgery , the following cardiac characteristics were captured : ejection fraction , prior mi , left main coronary disease , and previous cabg at date o...
Gold	backgroundthe obesity paradox reflects an observed relationship between obesity and decreased morbidity and mortality , suggesting improved health outcomes for obese individuals . studies examining the relationship between high body mass index (bmi) and adverse outcomes after cardiac surgery have reported conflicting results.methods and resultsthe study population (n=78 762) was comprised of adult patients who had undergone first - time coronary artery bypass (cabg) or combined cabg / aortic valve replacement (avr) surgery from april 1 , 1998 to october 31 , 2011 in ontario (data from the institute for clinical evaluative sciences) . perioperative outcomes and 5-year mortality among pre - defined bmi (kg / m2) categories (underweight < 20 , normal weight 20 to 24.9 , overweight 25 to 29.9 , obese 30 to 34.9 , morbidly obese > 34.9) were compared using bivariate analyses and cox multivariate regression analysis to investigate multiple confounders on the relationship between bmi and adverse outcomes . a reverse j - shaped curve was found between bmi and mortality with their respective hazard ratios . independent of confounding variables , 30-day , 1-year , and 5-year survival rates were highest for the obese group of patients (99.1% [95% confidence interval { ci } , 98.9 to 99.2] , 97.6% [95% ci , 97.3 to 97.8] , and 90.0% [95% ci , 89.5 to 90.5] , respectively) , and perioperative complications lowest . underweight and morbidly obese patients had higher mortality and incidence of adverse outcomes.conclusionsoverweight and obese patients had lower mortality and adverse perioperative outcomes after cardiac surgery compared with normal weight , underweight , and morbidly obese patients . the obesity paradox was confirmed for overweight and moderately obese patients . this may impact health resource planning , shifting the focus to morbidly obese and underweight patients prior to , during , and after cardiac surgery .
Model	backgroundthe aim of this study was to investigate the relationship between body mass index (bmi) and mortality in patients undergoing coronary artery bypass grafting (cabg) or open chest aortic valve replacement (avr).methods and resultswe identified 1 year patients undergoing primary cabg or combined cabg and open chest aortic valve replacement (avr) from april 1, 1998 to october 31, 2011. patients were divided into groups : underweight (bmi < 20 kg / m2), normal weight (bmi 25.0 to 29.9 kg / m2), overweight (bmi 25.0 to 29.9 kg / m2), obese (bmi 30.0 to 29.9 kg / m2), and morbidly obese (bmi > 34.9 kg / m2). patients were followed for 1 year with respect to major outcomes, and 5 years for mortality. cox proportional hazards regression analysis was used to investigate multiple confounders on the relationship between bmi and mortality, providing hazard ratios and 95% confidence intervals (ci). patients with bmi < 20 kg / m2 were more likely to die during the 5 years of follow - up (hazard ratio [hr]
Metrics	Rouge1: 56.82, Rouge2: 26.29, RougeL: 40.91, RougeLsum: 50.0, Summary length (tokens): 256

Table 15: Examples of the PubMed dataset using the model pubmed-4096-512 small diverse

Bottom 5% example (Sorted by rougeL)	
Document	in august , 4 months before presentation , a 35-year - old white woman of scots and english descent developed reddish urine for several days followed by eruption of vesicles and blisters on the dorsal surfaces of her hands and fingers , the sides of her nose , and her upper anterior chest , knees , and legs . she worked as a landscaping contractor and noticed that lesions occurred on areas exposed to sunlight , but application of sunscreen neither diminished the rate at which new lesions appeared , nor promoted healing of older lesions . her skin was fragile in areas of the lesions and the lesions healed slowly , often with scarring . she also developed dark brown pigmentation and the growth of fine black hair over her cheeks . she consumed three glasses of wine each week and had smoked electronic cigarettes for approximately 6 months , having changed from tobacco cigarettes . she had donated three units of blood for transfusion , but none in several years . she had no menses in the 12 months before presentation due to the effects of a contraceptive vaginal ring (nuvaring ; etonogestrel / ethinyl estradiol) . a dermatologist performed a punch biopsy of two skin lesions on her left forefinger and referred her for hematology evaluation and treatment . physical examination confirmed the presence of new vesicles and bullae with erythematous bases , some as large as 1 cm in diameter , and older lesions in various stages of erosion , resolution , and scarring in the anatomical distribution described above . lesions were most prominent on the dorsal surfaces of the hands and fingers (fig . milia were scattered over areas affected with bullous lesions and were especially prominent on skin overlying finger joints . (a) before therapy , there were vesicles , bullae with erythematous bases , older lesions in various stages of erosion , resolution , scarring , and punctate milia , all typical of pct . (b) after iron depletion with phlebotomy , there were no active skin lesions , although hyperpigmentation , scarring , and milia (especially over dorsal aspects of second and third interphalangeal joints) persisted . punch biopsy specimens of skin were deposited in immunofluorescence transport medium , flash frozen , and cut for manual immunofluorescence staining . the sections were probed with fluorescein - labeled anti - human antibodies specific for igg , iga , igm , c3 , c5b-9 , and fibrinogen . complete blood count (including ...
Gold	porphyria cutanea tarda (pct) is characterized by decreased uroporphyrinogen decarboxylase activity in hepatocytes , uroporphyrin i and heptacarboxyl porphyrin iii accumulation , photosensitivity dermatitis , and increased storage iron . in women , estrogen therapy , including oral contraceptives , postmenopausal hormone replacement , and tamoxifen for breast cancer treatment , is a risk factor for pct . we report the case of a woman who presented with pct , hfe c282y homozygosity , and hepatic iron overload and was using a contraceptive vaginal ring containing ethinyl estradiol , an estrogen . we discuss this case in the context of characteristics of other persons with pct , including common hfe mutations , iron overload , and estrogen exposure .
Model	a 35-year - old white woman of scots and english descent developed reddish urine for several days followed by eruption of vesicles and blisters on the dorsal surfaces of her hands and fingers, the sides of her nose, and her upper anterior chest, knees, and legs. skin biopsy specimens revealed linear glassy iga and igg deposition along the epidermal basement membrane zone and superficial dermal blood vessels, forming a doughnut pattern. the patient was treated with a levonorgestrel - releasing intrauterine contraceptive system (skyla). eight months after presentation, she underwent implantation of a levonorgestrel - releasing intrauterine contraceptive system (skyla). eight months after presentation, she underwent implantation of a levonorgestrel - releasing intrauterine contraceptive system (skyla).
Metrics	Rouge1: 20.56, Rouge2: 0.94, RougeL: 11.22, RougeLsum: 17.76, Summary length (tokens): 176
Bottom 25% example (Sorted by rougeL)	
Document	this was a multicenter , community - based , retrospective observational study of patients with pns , ranging from 8 to 20 mm in diameter , presenting to 18 geographically representative outpatient pulmonary clinics across the united states . the study was approved at 15 sites by a central institutional review board and at three sites by local institutional review board approval . four hundred forty sites were identified based on investigator databases and claims data from a large insurance carrier whose coverage population was representative of the overall us population . of these , 77 sites expressed interest in participating , and 48 sites went on to sign confidentiality agreements . of these , 17 did not request additional information , leaving 31 sites undergoing qualification review . eighteen outpatient pulmonary clinics were chosen to participate based on the following criteria : (1) management of patients with pns , (2) availability of medical records , and (3) ability to perform data abstraction . in addition , investigators targeted enrollment of geographically diverse patients to limit the potential bias associated with differences in practice patterns and to account for variation in disease prevalence (eg , endemic mycoses) that could alter management decisions . patients were identified by querying databases (eg , billing and scheduling systems) using five international classification of diseases , ninth revision , clinical modification codes for pn (793.1 , 786.6 , 518.89 , 519.8 , 519.9) to ensure homogeneity in patient identification and inclusion . manual chart abstraction was then used to identify those who met the criteria . to minimize selection bias , the sites were not permitted to use additional codes during database query to identify patients . to ensure a systematic sample , inclusion criteria included age 40 years and 89 years at the time of nodule finding , presentation to a pulmonologist , nodule size 8 to 20 mm , and definitive diagnosis ascertained by tissue diagnosis or radiographic follow - up for 2 years . exclusion criteria included chest ct scan performed > 60 days prior to the initial visit , prior diagnosis of any cancer within 2 years of nodule detection , or incomplete chart data . patients were categorized into three groups by the most invasive procedure performed during management , as follows : surveillance (serial imaging) , biopsy (ct scan - guided transthoracic needle aspi...
Gold	background : pulmonary nodules (pns) are a common reason for referral to pulmonologists . the majority of data for the evaluation and management of pns is derived from studies performed in academic medical centers . little is known about the prevalence and diagnosis of pns , the use of diagnostic testing , or the management of pns by community pulmonologists.methods:this multicenter observational record review evaluated 377 patients aged 40 to 89 years referred to 18 geographically diverse community pulmonary practices for intermediate pns (8 - 20 mm) . study measures included the prevalence of malignancy , procedure / test use , and nodule pretest probability of malignancy as calculated by two previously validated models . the relationship between calculated pretest probability and management decisions was evaluated.results:the prevalence of malignancy was 25% (n = 94) . nearly one - half of the patients (46% , n = 175) had surveillance alone . biopsy was performed on 125 patients (33.2%) . a total of 77 patients (20.4%) underwent surgery , of whom 35% (n = 27) had benign disease . pet scan was used in 141 patients (37%) . the false - positive rate for pet scan was 39% (95% ci , 27.1%-52.1%) . pretest probability of malignancy calculations showed that 9.5% (n = 36) were at a low risk , 79.6% (n = 300) were at a moderate risk , and 10.8% (n = 41) were at a high risk of malignancy . the rate of surgical resection was similar among the three groups (17% , 21% , 17% , respectively ; p = .69).conclusions : a substantial fraction of intermediate - sized nodules referred to pulmonologists ultimately prove to be lung cancer . despite advances in imaging and nonsurgical biopsy techniques , invasive sampling of low - risk nodules and surgical resection of benign nodules remain common , suggesting a lack of adherence to guidelines for the management of pns .
Model	objectivethe aim of this study was to document the prevalence of cancer (25%) in patients with intermediate - sized pulmonary nodules referred to community pulmonologists.materials and methodsthis was a multicenter, community - based, observational study of patients with intermediate - sized pulmonary nodule who were referred to pulmonologists. patients were categorized into three groups by the most invasive procedure performed during management, as follows : surveillance (serial imaging), biopsy (ct - guided transthoracic needle aspiration [ttna] or bronchoscopy), or surgery (including mediastinoscopy, video - assisted thorascopic surgery, and/or thoracotomy). patients were categorized into three groups by the most invasive procedure performed during management, as follows : surveillance (serial imaging), biopsy (ct - guided transthoracic needle aspiration [ttna] or bronchoscopy), or surgery (including mediastinoscopy, video - assisted thorascopic surgery, and/or thoracotomy). patients were categorized into three groups by the most invasive procedure performed during management, as follows : surveillance (serial imaging), biopsy (ct - guided transthoracic needle aspiration [ttna] or bronchoscopy), or surgery (including mediastinoscopy, video - assisted thorascopic surgery, and/or thoracotomy). multivariate logistic regression was performed to identify factors associated with the use of an invasive diagnostic procedure.resultsof the 377 patients included, 283 (75%) had a nodule that was benign, and 94 (25%) had a malignant nodule. the overall accuracy of pet scanning was 74%, with a false - positive (fp) rate of 39% and a false - negative (fn) rate of 9%. the overall accuracy of pet scanning was 74%, with a false - positive (fp) rate of 39% and a false - negative (fn) rate of 9%. nodules measuring > 11 to 15 mm (n = 48) had fn and fp rates of 9% and 36%, respectively.conclusionsthe prevalence of cancer in patients with intermediate - sized nodules was 25%. the rate of surgical resection for benign disease varied from 9% to 23% in screening trials and surgical series.
Metrics	Rouge1: 45.58, Rouge2: 9.56, RougeL: 18.37, RougeLsum: 38.1, Summary length (tokens): 470
Top 75% example (Sorted by rougeL)	
Document	a total of 1,217 dead birds were shipped at 4c to the tropical medicine institute " pedro kouri " and identified by ornithology experts . brain , heart , and kidneys were removed and tested for wnv by using reverse transcription polymerase chain reaction (rt - pcr) (12) . briefly , rna was extracted by using the qiamp viral rna kit (qiagen , inc . , valencia , ca , usa) . primers wn212 (5-ttgtgttggctctcttggcgttctt-3) and wn619c (5-cagccgacagcactggacattcata-3) were used to detect viral rna . a second rt - pcr with primers wn9483 (5-cacctacgccctaaacactttcacc-3) and wn9794 (5-ggaacctgctgccaatcataccatc-3) was performed on the same rna preparation . serum specimens from horses in havana and havana province were tested for antibodies to wnv by using a competitive enzyme - linked immunosorbent assay (elisa) with monoclonal antibodies 3.1112 g and 6b6c-1 as described by blitvich et al . we tested 210 serum specimens from horses collected as part of an infectious anemia study . the immunoglobulin m (igm) test was not performed because horses were never suspected of having wnv and did not have any history of suspected viral encephalitis or other illness or symptoms . an inhibition value > 30% was used as the diagnostic criterion to identify flavivirus antibody (table 1) . * wnv , west nile virus ; elisa , enzyme - linked immunosorbent assay ; prnt , plaque reduction neutralization test ; slev , saint louis encephalitis virus . the cuban health ministry and medical services division conducted surveillance for encephalitis of unknown origin in patients > 30 years of age . serum and cerebrospinal fluid specimens were shipped at 4c to the tropical medicine institute " pedro kouri . " human sera were screened for wnv igm and igg by using commercial igm and igg elisa kits (focus technologies , cypress , ca , usa) according to manufacturer 's instructions . hemagglutination - inhibition (hi) tests were also undertaken with wnv and saint louis encephalitis virus (slev) antigen (14) . reactive serum samples were further tested by a plaque reduction neutralization test (prnt) with wnv (ny99 , ontario , canada , 2001 isolate) , slev (parton strain , american type culture collection catalog no . vr-1265) , and dengue virus (dengue 2 , ng - c strain) . prnt was performed to confirm wnv - specific antibody and was carried out as described previously (15) by using a neutral red double - overlay procedure . ho...
Gold	a surveillance system to detect west nile virus (wnv) was established in cuba in 2002 . wnv infection was confirmed by serologic assays in 4 asymptomatic horses and 3 humans with encephalitis in 2003 and 2004 . these results are the first reported evidence of wnv activity in cuba .
Model	we report the first evidence of antibodies to west nile virus (wnv) in horses and humans in cuba. antibodies were detected by plaque reduction neutralization test in 10 of 1,217 dead birds tested. these findings provide evidence that wnv and slev may co - circulate in cuba.
Metrics	Rouge1: 51.06, Rouge2: 17.39, RougeL: 31.91, RougeLsum: 46.81, Summary length (tokens): 64
Top 95% example (Sorted by rougeL)	
Document	intra - articular injections of corticosteroids have been used for several decades in the management of inflammatory and degenerative joint conditions when first - line conservative therapies such as rest , ice , and anti - inflammatory medications fail to provide adequate symptom relief . based in part on this long history of successful utilization coupled with the findings of several randomized controlled trials , consensus statements and meta - analyses have concluded that intra - articular corticosteroid injections provide short - term patient benefit and clinical efficacy for chronic knee pain.13 more recently , various injectable hyaluronic acid agents have become commercially available and have enjoyed widespread clinical acceptance as an effective treatment for knee osteoarthritis . these agents are indicated for the treatment of the pain associated with osteoarthritis of the knee in patients who have failed to respond adequately to conservative nonpharmacologic therapy and simple analgesics , eg , acetaminophen . traditionally , intra - articular injections have been performed using anatomical landmarks to identify the correct trajectory for needle placement . however , different anatomical - guided injection techniques have yielded inconsistent intra - articular needle positioning due , in large part , to the fact that the physician can not directly visualize the area of interest , and variations in anatomy are common . incorrect needle placement has been partially attributed to variable clinical outcomes.410 furthermore , inaccurate corticosteroid injections in the knee , for example , may result in post - injection pain , crystal synovitis , hemarthrosis , joint sepsis , and steroid articular cartilage atrophy , as well as systemic effects , such as fluid retention or exacerbation of hypertension or diabetes mellitus.1 therefore , identification of methods and proper training to aid in correct needle placement during these procedures is warranted . various imaging modalities can be used to improve the accuracy of intra - articular injections , including fluoroscopy , computed tomography , and magnetic resonance imaging . however , musculoskeletal ultrasound is one of the most practical because it is rapid , safe , relatively inexpensive , emits no ionizing radiation , and can be performed in the outpatient clinical setting.11,12 ultrasound utilizes high - frequency sound waves to visualize soft tissues and bony structures and is a f...
Gold	intra - articular corticosteroid and hyaluronic acid injections provide short - term symptom amelioration for arthritic conditions involving structural damage or degenerative changes in the knee . conventional palpation - guided anatomical injections frequently result in inaccurate needle placement into extra - articular tissue and adjacent structures . the purpose of this review was to determine the effect of ultrasound guidance on the accuracy of needle placement , clinical outcomes , and cost - effectiveness in comparison with anatomical landmark - guided intra - articular large joint injections , with particular emphasis on the knee . a total of 13 relevant studies were identified ; five studied the knee , seven studied the shoulder , one used both the knee and shoulder , and none studied the hip . ultrasound was used in seven studies ; the remaining studies utilized air arthrography , fluoroscopy , magnetic resonance arthrography , or magnetic resonance imaging . across all studies (using all imaging modalities and all joints) , needle placement accuracy ranged from 63% to 100% with ultrasound and from 39% to 100% with conventional anatomical guidance . imaging guidance improved the accuracy of intra - articular injections of the knee (96.7% versus 81.0% , p < 0.001) and shoulder (97.3% versus 65.4% , p < 0.001) . in particular , ultrasound guidance of knee injections resulted in better accuracy than anatomical guidance (95.8% versus 77.8% , p < 0.001) , yielding an odds ratio of 6.4 (95% confidence interval 2.914) . ultrasound guidance notably improves injection accuracy in the target intra - articular joint space of large joints including the knee . the enhanced injection accuracy achieved with ultrasound needle guidance directly improves patient - reported clinical outcomes and cost - effectiveness .
Model	objectivethe purpose of this review was to determine the effect of ultrasound guidance on the accuracy of needle placement, clinical outcomes, and cost - effectiveness in comparison with anatomical landmark - guided intra - articular large joint injections, with particular emphasis on the knee.methodsthe peer - reviewed literature was searched using medline and relevant bibliographies published in english language journals through december 31, 2011 that compared the accuracy of intra - articular injections with anatomical landmark - guided intra - articular large joint injections, with particular emphasis on the knee.resultsa total of 13 studies met the entry criteria. needle placement accuracy ranged from 63% to 100% with ultrasound and from 39% to 100% with conventional anatomical guidance. imaging guidance improved the accuracy of intra - articular injections of the knee (96.7% versus 81.0%, p < 0.001) and shoulder (97.3% versus 65.4%, p < 0.001). imaging guidance improved the accuracy of intra - articular injections of the knee (96.7% versus 81.0%, p < 0.001) and shoulder (97.3% versus 65.4%, p < 0.001). imaging guidance improved the accuracy of intra - articular injections of the knee (96.7% versus 81.0%, p < 0.001) and shoulder (97.3% versus 65.4%, p < 0.001). in particular, ultrasound guidance of knee injections resulted in better accuracy than did anatomical guidance (95.8% versus 77.8%, p < 0.001). in particular, ultrasound guidance of knee injections resulted in better accuracy than did anatomical guidance (95.8% versus 77.8%, p < 0.001). in particular, ultrasound guidance of knee injections resulted in better accuracy than did anatomical guidance (95.8% versus 77.8%, p < 0.001).conclusionthe use of imaging guidance, in particular ultrasound, improves the accuracy of intra - articular injections in large joints, including the knee. furthermore, accurate ultrasound - guided intra - articular knee injections improve clinical outcomes and lower health care costs.
Metrics	Rouge1: 62.21, Rouge2: 43.74, RougeL: 48.51, RougeLsum: 58.7, Summary length (tokens): 464

Table 16: Examples of the PubMed dataset using the model pubmed-4096-512 base diverse

MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset

Tobias Brugger*
University of Bern
Switzerland
tobias.brugger@students.unibe.ch

Matthias Stürmer
University of Bern
Switzerland
Bern University of Applied Sciences
Switzerland
matthias.stuermer@unibe.ch

Joel Niklaus*
University of Bern
Switzerland
Bern University of Applied Sciences
Switzerland
Stanford University
United States
joel.niklaus@unibe.ch

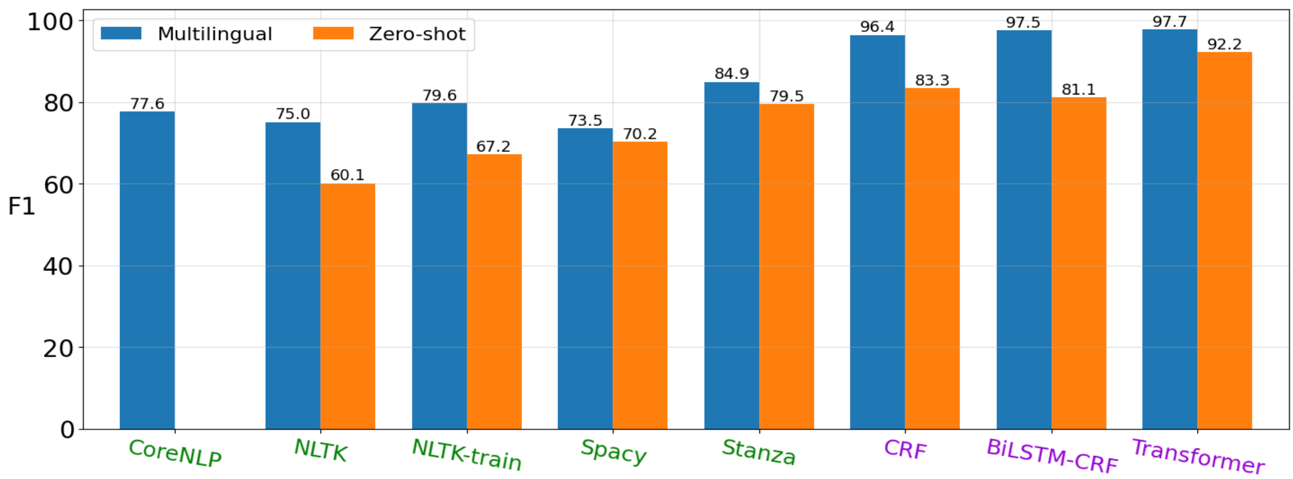


Figure 1: Mean performance comparison of baseline (green) and our multilingual models (violet) on multilingual legal text (French, Italian, Spanish, English, and German) and a zero-shot experiment on Portuguese data

ABSTRACT

Sentence Boundary Detection (SBD) is one of the foundational building blocks of Natural Language Processing (NLP), with incorrectly split sentences heavily influencing the output quality of downstream tasks. It is a challenging task for algorithms, especially in the legal domain, considering the complex and different sentence structures used. In this work, we curated a diverse multilingual legal dataset consisting of over 130'000 annotated sentences in 6 languages. Our experimental results indicate that the performance of existing SBD models is subpar on multilingual legal data. We trained and tested monolingual and multilingual models based on CRF, BiLSTM-CRF, and transformers, demonstrating state-of-the-art performance. We also show that our multilingual models outperform all baselines in the zero-shot setting on a Portuguese test set. To encourage further research and development by the community, we have made our dataset, models, and code publicly available.

*Both authors contributed equally to this research.

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06.
https://doi.org/10.1145/3594536.3595132

CCS CONCEPTS

• **Applied computing** → *Law*; Annotation; • **Computing methodologies** → **Natural language processing**; *Supervised learning*; **Machine learning**.

KEYWORDS

Sentence Boundary Detection, Natural Language Processing, Legal Document Analysis, Text Annotation, Multilingual

ACM Reference Format:
Tobias Brugger, Matthias Stürmer, and Joel Niklaus. 2023. MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3594536.3595132

1 INTRODUCTION

Recent methodological advances, e.g., transformers [34], have lead to substantial progress in quality and performance of language models as well as growth in the general field of Natural Language Processing (NLP). This trend is also evident in legal NLP, with research papers increasing drastically in recent years [14].

Not as much attention and resources have been directed to the Sentence Boundary Detection (SBD) task, being viewed as solved by some, as high baseline performances can be achieved by utilizing simple lookup methods capturing frequent sentence-terminating characters such as periods, exclamations marks and question marks combined with hand-crafted rules [26]. This approach is feasible when applied to well-formed and curated text such as news articles. Noisier domain-specific data containing differently structured text combined with the ambiguity of many sentence-terminating characters [8, 15] – e.g., the period occurring in abbreviations, ellipses, initials etc. as a non-terminating character – often overwhelm the aforementioned methods and also more complicated off-the-shelf SBD systems. This has been illustrated in a number of specific SBD applications such as user-generated content [9, 26] as well as in the clinical [20] and financial domain [7, 19].

In legal documents, the aforementioned difficulties are increased with legal text consisting of smaller parts such as paragraphs, clauses etc., making it quite different from standard text. Furthermore, sentences are long and may contain complex structures such as citations, parentheses, and lists. These structures are often utilized to convey additional information to the reader (e.g., citations referencing another text) or formatting the text in a specific way (e.g., lists emphasizing ideas or increasing the readability of long paragraphs). However, these structures or special sentences do not follow a standard sentence structure, thus posing an additional challenge to SBD systems, illustrated in several works on English [27, 29] and German [10] legal documents.

1.1 Motivation

Having a reliable SBD system is crucial for accurate NLP analysis of text. Poor SBD can result in errors propagating into higher-level text processing tasks, which hinders overall performance. For instance, the curation of the multilingual EUROPARL corpus required proper SBD to align sentences in both languages for statistical machine translation. Koehn [16] noted the difficulty of SBD as it requires specialized tools for each language, which are not readily available for all languages. Inadequate SBD weakens the performance of sentence alignment algorithms and reduces the quality of the corpus. Therefore, a high-quality SBD system, especially one customized for the legal domain, can significantly improve performance.

Another example is Negation Scope Resolution (NSR), focusing on finding negation words (e.g., "not") in sentences and their impact on surrounding words' meaning. Negations are vital in text's semantic representation, reversing proposition values. This is particularly useful in the legal domain, enabling models extracting information from documents to better understand input text meaning, such as recognizing court decisions' outcomes based on exact wording. NSR models often require data split into sentences for labeling training data and application input, making a reliable SBD system crucial. Incorrect sentence predictions by the SBD system may significantly lower input data quality and model performance. Proper SBD is also crucial in other NLP tasks such as Text Summarization, Part-of-Speech-Tagging, and Named Entity Recognition, all relevant in the legal domain.

1.2 Main Research Questions

In this work, we pose and examine three main research questions:

RQ1: What is the performance of existing SBD systems on legal data in French, Spanish, Italian, English, and German?

RQ2: To what extent can we improve upon this performance by training mono- and multilingual models based on CRF, BiLSTM-CRF, and transformers?

RQ3: What is the performance of the multilingual models on unseen Portuguese legal text, i.e., a zero-shot experiment?

1.3 Contributions

The contributions of this paper are twofold:

(1) We curate and publicly release a large, diverse, high-quality, multilingual legal dataset (see Section 3) containing over 130'000 annotated sentence spans for further research in the community.
(2) Using this dataset, we showcase that existing SBD systems exhibit suboptimal performance on legal text in French, Italian, Spanish, English, and German. We train and evaluate state-of-the-art monolingual SBD models based on Conditional Random Fields (CRF), BiLSTM-CRF and transformers, achieving F1-scores up to 99.6%. We showcase the performance and feasibility of multilingual SBD models, i.e., trained on all languages, achieving F1-scores in the higher nineties, comparable or better than our monolingual models on each aforementioned language. In a zero-shot experiment, we demonstrate that it is possible to achieve good cross-lingual transfer by testing the multilingual models on unseen Portuguese legal text. We publicly release the datasets[1], all of our monolingual and multilingual models[2] (see Section 5) as well as our code[3] for further use in the community.

2 RELATED WORK

In this section, we discuss the literature at our disposal. First, we look at works showcasing the need for more research in regard to SBD. Second, we take a look at works tackling the problem of SBD in legal text in several languages. Lastly, we investigate SBD research in other domains and present multilingual datasets in the legal domain for thoroughness.

Read et al. [26] questioned the status quo of SBD being "solved", especially in more informal language and special domains, by reviewing the current state-of-the-art SBD systems on English news articles and user-generated content. The systems were able to reach F1-scores in the higher nineties for the former, however the performance on user-generated content weakened perceptibly with scores down to the lower nineties, showcasing the need for "a renewed research interest in this foundational first step in NLP." [26]

[1] https://huggingface.co/datasets/rcds/MultiLegalSBD
[2] https://huggingface.co/models?search=rcds/distilbert-sbd and https://github.com/tobiasbrugger/MultiLegalSBD/tree/master/models
[3] https://github.com/tobiasbrugger/MultiLegalSBD

2.1 SBD in the Legal Domain

Savelka et al. [29] continued this research in the English language by curating a legal dataset, consisting of adjudicatory decisions from the United States. When testing existing systems on the dataset, they report F1-scores between 75% and 78%. Training or adapting these systems to the dataset improved their F1 score to the mid-eighties, which is still lower than their respective performance in more standard domains [26], showcasing the subpar performance of state-of-the-art SBD in the English legal domain. To improve this issue, they trained a number of CRF models as well as a model based on hand-crafted rules, reporting F1-scores of 79% for the hand-crafted model and up to 96% for the CRFs. Additionally, they developed a publicly available, comprehensive set of annotation-guidelines for sentence boundaries in legal texts which we used as a foundation for our guidelines.

Sanchez [27] experimented on the same dataset reporting an F1-score of 74% using the Punkt Model [15]; adapting it to the dataset slightly improved performance. They also trained and evaluated CRF and Neural Network (NN) models, reporting F1-scores up to 98.5% and 98.4% respectively. Our multilingual models achieve F1-scores between 95.1% and 97% on the same dataset.

Similarly, Glaser et al. [10] curated a German legal dataset, split into laws and judgements; a similar distribution is used in our work. They established a baseline performance of existing SBD systems and compared it to CRF and NN models trained on the aforementioned dataset. Their findings outline F1-scores between 70% to 78% for off-the-shelf systems, supporting the view that the performance of existing SBD system is subpar on legal data. The CRFs and NNs models achieve F1-scores up to 98.5%. However, a significant decrease in performance was reported, when applying them to previously unseen German legal texts with scores down to 81.1%. Our multilingual models showcase F1-scores between 91.6% to 97.6% on the German dataset.

2.2 SBD in Other Domains

In the financial domain, Du et al. [7] experimented with Bidirectional Long Short-Term Memory (BiLSTM) models combined with a CRF layer as well as the transformer-based model BERT [6] and compared their performance, approaching SBD as a sequence labelling task to extract useful sentences from noisy financial texts. They demonstrate that BERT significantly outperforms BiLSTM-CRFs across all evaluation metrics, including F1-scores. In their work they also underline the fact that "SBD has received much less attention in the last few decades than some of the more popular subtasks and topics in NLP."

Schweter and Ahmed [31] compared the performance of Long Short-Term Memorys (LSTMs), BiLSTMs and Convolutional Neural Networks (CNNs) to OpenNLP [4] in an SBD task on the Europarl [16], SETimes [33] and Leipzig Corpora [11] containing around 10 different languages, showcasing the use of their models as robust, language-independent SBD systems.

2.3 Multilingual Datasets in the Legal Domain

Niklaus et al. [23] present LEXTREME, a novel multilingual benchmark dataset containing 11 datasets in 24 languages, designed to evaluate natural language processing models on legal tasks. The authors assess five prevalent multilingual language models, providing a benchmark for researchers to use as a basis for comparison. Savelka et al. [30] investigate the application of multilingual sentence embeddings in sequence labeling models to facilitate transfer across languages, jurisdictions, and other legal domains. They demonstrate encouraging outcomes in allowing the reuse of annotated data across various contexts, which leads to the development of more resilient and generalizable models. Additionally, they create a vast dataset of newly annotated legal texts using these models. Chalkidis et al. [3] introduce MultiEURLEX, a multilingual and multilabel legal document classification dataset containing 65000 EU Laws. Aumiller et al. [1] present a EurLexSum, a multilingual summarization dataset curated from Eur-Lex data. Niklaus et al. [21, 24] introduce Swiss-Judgment-Prediction, a multilingual judgment prediction dataset from the Federal Supreme Court of Switzerland.

3 DATASET

We annotated sentence spans for three diverse multilingual legal datasets in French, Italian, and Spanish, each containing approximately 20,000 sentences evenly split between judgments and laws. We chose a variety of legal areas to capture a broad selection. The laws included the Constitution, part of the Civil Code, and part of the Criminal Code, with the Constitution used only for evaluation. The judgments comprised court decisions from various legal areas and sources. We also annotated a smaller Portuguese dataset with approximately 1800 sentences, divided into the same subsets as the other datasets. This dataset was used for zero-shot experiments.

Additionally, we standardized and integrated two publicly available datasets, an English collection of legal texts [29], consisting of Adjudicatory Decision from the United States as well as a German dataset [10], comprising laws and judgments, into our dataset to further increase its diversity.

Figure 2 illustrates the sentence length distribution of our dataset, showing the relative frequency of sentence length in tokens for laws and judgments, with a bin size of 5. We used an aggressive tokenizer, resulting in a larger number of tokens per sentence than usual. For clarity, we did not include sentences longer than 101 tokens, which comprised only ~2% (2634) of the sentences. Only 26 sentences were longer than 512 tokens.

For each language, we used random sampling to split the dataset into three parts: train, test and validation. The test and validation splits each contain 20% of the dataset. Every model is trained on the train split, and we report their performance on the test split. Selected statistics and information about the dataset are in Table 1.

3.1 Annotation

The human annotator was tasked with correcting the sentence-spans predicted by an automatic SBD system[5] [29] based on CRF, which was trained on data annotated using annotation guidelines by Savelka et al. [29]. This helped improve the quality and consistency of our annotations. Furthermore, a practical rule set, heavily influenced by the aforementioned guidelines, was utilized to aid the annotator in the annotation process, reducing the complexity of the task and helped provide dependable and well-founded data. The

[4]https://opennlp.apache.org/

[5]https://github.com/jsavelka/luima_sbd

Table 1: Statistics on datasets per language and subset

Language	Subset	Sentences	Tokens	# of Documents	Source
French	Judgments	9971	342469	315	Niklaus et al. [21]
	Laws	11055	334453	3	Wipolex
Italian	Judgments	10129	340041	183 + 60	Niklaus et al. [21] + Multi-Legal-Pile (MLP)
	Laws	10849	301466	3	Jus.unitn.it
Spanish	Judgments	10656	356681	20 + 84	Wipolex + MLP
	Laws	11501	229240	3	Wipolex
Portuguese	Judgments	759	20590	6	Wipolex
	Laws	1010	25947	3	Wipolex
German	Judgments	21409	506009	131	Glaser et al. [10]
	Laws	20330	484816	13	Glaser et al. [10]
English	Judgments	25899	712433	80	Savelka et al. [29]
Total	Laws & Judgments	133568	3654145	906	

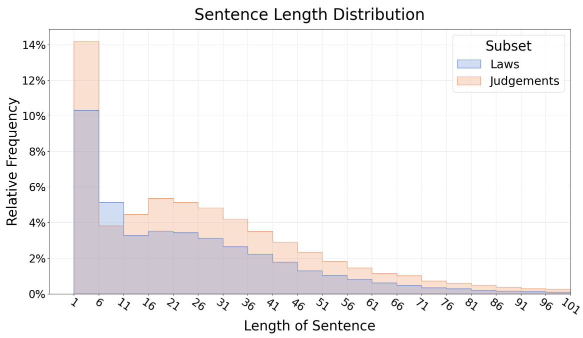


Figure 2: Sentence length distribution in tokens

rule set is outlined in Section 3.1.1, containing the most important sentence structures followed by an example.

The documents were annotated using Prodigy[6]. Because Prodigy requires pre-tokenized text, a customized tokenizer was applied to the input text, further described in Section 3.2. The decision to annotate the full sentence-span, in lieu of just the first and last token in the sequence, was made to incentivize the annotator to read the text instead of skimming it for sentence-terminating characters. To make the annotation easier, laws were split into smaller chunks with one to three articles per chunk, while judgments were only split, if they surpassed ~15000 characters since Prodigy was unable to handle longer documents.

3.1.1 Legal Sentence Structures. In this section, we briefly describe the most important sentence structures in legal text, heavily influenced by Savelka et al. [29], followed by an example in French.

Standard Sentence have subject, object and verb in the correct order and the last token in the sequence is a sentence-terminating character.
- *Il s'est établi comme ingénieur indépendant.*

Linguistically Transformed Sentence are similar to a standard sentence, but slight transformations such as changes to the word order are applied.
- *Tout porte à croire, en réalité, qu'elle est condamnée au surendettement, puis à la faillite.*

Headlines determine the structure of the text and show relatedness between parts of the document and therefore convey important information about the overall structure of the text.
- *Considérant en fait et en droit*
- *PAR CES MOTIFS*
- *DÉCLARATION*

Data fields provide the name and data of a field. This is annotated as a sentence, as for example in English "Civil Chamber: Madrid" has a similar meaning to "The civil chambers are located in Madrid".
- *Numéro d'appel: 1231/2015*

Parentheses appear frequently in legal text, often combined with citations. We annotate parentheses with the sentence they belong to. Sequences inside the parentheses are not annotated separately, as seen in the following example, containing a single sentence:
- *Ce dernier étant domicilié à l'étranger, il ne peut en effet prétendre à des mesures de réadaptation (art. 8a. 1er paragraphe. Convention de sécurité sociale entre la Suisse et la Yougoslavie du 8 juin 1962).*

Colons should not be annotated as a sentence-terminating character, unless the colon is immediately followed by a newline. The reasoning here is that a sequence ending in a colon followed by a line break usually introduce a list or block quote, which should be annotated separately to the introductory sentence.

Lists are annotated differently depending on its type. For lists with incomplete sentences as list items, often ended with a semi-colon, the whole list is annotated as a sentence. The following example consists of 2 sentences, the introductory sentence to the colon and 1° to the period.
- *Au cours du délai fixé par la juridiction pour accomplir un travail d'intérêt général, le condamné doit satisfaire aux mesures de contrôle suivantes:*
  1° Répondre aux convocations du juge de l'application des peines;
  2° (...) une affection dangereuse pour les autres travailleurs.

[6]https://prodi.gy/

However, if the list items themselves are sentences, the list number (or letter) and items are both annotated as one sentence each, the reason being that they express separate thoughts. In the example below we have 3 sentences (introductory, list number, list item).

- *Considérant en droit:*
  1.- En instance fédérale, peut seul être examiné le point de savoir si la commission de recours a exigé à bon droit de la recourante une avance de frais de 500 fr. pour la procédure de recours de première instance.

Ellipses are used to indicate when part of a sentence or part of the document are left out. The following example shows the use cases for ellipses. The first ellipsis is annotated separately, as it indicates sentences that are missing. The second ellipses indicates, that part of that single sentence was left out and is therefore not annotated separately.

- *(...) La faute de X. est d'une exceptionnelle gravité tant les faits qui lui sont reprochés (...), commis avec une certaine froideur sont insoutenables et comportent un caractère insupportable pour les victimes.*

Footnotes / Endnotes convey additional information to the reader. Indicators for end- and footnotes such as numbers or letters should always be annotated as being inside the sentence span, even if they occur after the sentence-terminating character. As an example, the sequence below is just one sentence, with "(2)" as the indicator:

- La loi ne dispose que pour l'avenir; elle n'a point d'effet rétroactif. (2)

Furthermore, endnotes appearing as numbered lists, should be annotated as following the guidelines for lists. In the example below, (2) is one sentence, followed by a normal sentence:

- (2) Le remplacement des membres du Parlement a lieu conformément aux dispositions de l'article 25.

3.2 Tokenizer

We implemented an aggressive tokenizer based on Regex to segment text into tokens, also employed in other research [10, 29]. This tokenizer was utilized for all languages. Words, numbers and special characters such as newlines and whitespace are separated into individual sequences. This was done to ensure no information (e.g., a line break indicating a sentence boundary), vital to the SBD process, was lost. An example is showcased below; tokenized whitespace is left out for clarity's sake:

- *D._ est entré à l'école le 16 juillet 1979.*
- *D | . | _ | est | entré | à | l | ' | école | le | 16 | juillet | 1979 | .*

4 EXPERIMENTAL SETUP

We conducted a series of experiments to answer our research questions posed in Section 1.2. Firstly, we compared selected existing models to establish a baseline performance. Secondly, we trained and evaluated various monolingual and multilingual models based on CRF, BiLSTM-CRF, and transformers, comparing them to baselines. Lastly, we evaluated the multilingual models' performance on unseen data in a zero-shot experiment.

4.1 Baseline Systems

We conducted a thorough evaluation of several widely used systems utilizing various technologies, including CoreNLP, NLTK, Stanza, and Spacy, which served as our baselines. In the following section, we will provide a detailed description of each system.

4.1.1 NLTK. A fully unsupervised SBD system created by Kiss and Strunk [15]. The main thought behind the system is that most falsely predicted sentence boundaries stem from periods after abbreviations. The system therefore discovers abbreviations by looking at the length, the collocational bond, internal periods and occurrences of abbreviations without an ending period of each token in the text. We test a pre-trained model as well as a model trained on our data.

4.1.2 CoreNLP. A rule-based system from the Stanford CoreNLP toolkit [18], which predicts sentence boundaries based on events like periods, question marks, or exclamation marks.

4.1.3 Stanza. A multilingual system based on a BiLSTM model [25]. We only use the first part of its NLP pipeline, the tokenizer. It addresses tokenization and sentence splitting jointly, treating it as a character sequence tagging problem, predicting if a character is the end of a token or sentence.

4.1.4 Spacy. A multilingual system [12] with pre-trained models using technologies like CNN and transformers. For our purposes, only the tokenizer and sentence splitter were used.

4.2 Our Models

Following the works presented in Section 2, we chose to test models based on CRFs, BiLSTM-CRFs and transformers. We further describe these models in the following subsections. For testing, we trained[7] and evaluated monolingual models for each language as well as multilingual models using all languages except Portuguese, once for laws, once for judgments and both types together.

4.2.1 Conditional Random Fields. The tokenizer in Section 3.2 tokenized input text, including whitespaces. Each token was translated into a list of simple features representing the token, and the features of tokens within a pre-defined window around the token were added. Window sizes for each feature varied, inspired by Glaser et al. [10] and Savelka et al. [29], as shown in Table 2. We labeled input data using the "BILOU" system following Lin et al. [17].

For training our CRF models, we used the python-crfsuite[8] implementation. We trained each model for 100 iterations, with regularization parameters 1 and $1e^{-3}$ for C1 and C2, L-BFGS as the algorithm, and including all possible feature transitions.

4.2.2 Bidirectional LSTM - CRF. A BiLSTM connects two LSTMs with opposite directions to the same output, allowing it to capture information from past and future states at the same time. The outputs of each LSTM are concatenated into a representation of each input token. For a BiLSTM-CRF model, a CRF layer is connected to the output of the BiLSTM network, using the aforementioned representation as features to predict the final label.

[7] GPU: NVIDIA GeForce RTX 3060 TI, CPU: Intel Core i5-8600K CPU @ 3.60GHz
[8] https://pypi.org/project/python-crfsuite/

Table 2: Description of CRF-Features

Feature	Description	Window
Special	Each token is categorized using the following translation: Sentence-terminating tokens as "End", opening and closing parentheses as "Open" and "Close" respectively, newline characters as "Newline", abbreviation characters as "Abbr" and the rest as "No".	10
Lowercase	The token in lowercase.	7
Length	The length of the token.	7
Signature	Each character is represented using the following translation: Lower case and upper case character are rewritten as "c" and "C" respectively, digits are written as "N" and special characters as "S".	5
Lower	Whether the first character is lower case.	3
Upper	Whether the first character is upper case.	3
Digit	Whether the token is a digit.	3

We utilized the Bi-LSTM-CRF[9] library to train our models. We used a word embedding dimension of 128, hidden dimension of 256 and a maximum sequence length of 512. The batch-size was 16 with a learning rate of 0.01 and a weight decay of 0.0001. We trained each model for 8 epochs and saved the model with the smallest validation loss. We extracted word embeddings for training from our documents. To label the training data, we utilized the "BILOU" labeling system described in Section 4.2.1. For training, gold sentences were put together into batches with a token-limit of 512 to simulate longer paragraphs.

4.2.3 Transformer. Transformers are a type of NN that utilizes self-attention mechanisms to weigh the importance of difference parts of the input when making predictions. Transformer models such as BERT use a multi-layer encoder [34] to pre-train deep bidirectional representations by jointly conditioning on both left and right context across all layers [6]. Thus, we can fine-tune transformer models to the SBD task by adding an additional output layer. In our case we used a pre-trained model[10] based on DistilBERT [28], a smaller, more lightweight version of BERT, for all languages on our SBD task.[11] We trained the models using PyTorch[12] and Accelerate[13] with the Adam optimizer for 5 epochs with a batch-size of 8 and learning rate of $2e^{-5}$.

A limitation of DistilBERT is the input length limit of 512 tokens because the runtime of the self-attention mechanism scales quadratically with the sequence length. This issue is exacerbated, since DistilBERT relies on a WordPiece Tokenizer [32], splitting the text into subwords resulting in a higher token count per sequence. Thus, to get around the 512 token-limit, each document was split into sentences using the gold annotation. Each consecutive sentence was added to a collection until the total length was as close to the token-limit as possible. Next, the model predicted the sentence boundaries for each collection. Sentences longer than 512 tokens were truncated.[14] An obvious downside to this solution is that the input text already has to be split into sentences or short sections, making it difficult to apply BERT models to unknown text.

For future work, it would be interesting to see, whether it is feasible to chain SBD models (i.e., first, apply a CRF model on the input text to split the text into sections smaller than 512 tokens and second apply a transformer based model). Another solution might be using pre-trained transformer models that support longer input text utilizing an attention mechanism scaling linearly with sequence length, such as Longformers [2].[15]

4.3 Evaluation

A characteristic of the SBD task is the inherent imbalance towards non-sentence boundary labels, as each sentence can at most have two sentence boundaries. Thus, to more accurately score our models, we used commonly utilized measures to evaluate our models - Precision (P), Recall (R) and F1-Score (F1). Although the SBD task is not yet solved in specialized domains, it is comparatively easier than other NLP tasks such as Questions Answering or Summarization. Because SBD is a pre-processing task, it is necessary to achieve higher scores to prohibit the propagation of errors into downstream tasks. Thus, we expect that state-of-the-art SBD models exhibit F1-scores in the high nineties to be useful in practice.

For the evaluation process, we let models predict the sentence spans of every document. These annotated spans are tokenized by our tokenizer (Section 3.2). Each token is then assigned a binary value, depending on whether it was a sentence boundary or not. This decouples the predicted sentence spans or boundaries from the tokenizer used, as the tokenizer of some models might designate a slightly different token as the first or last in a sentence, further described in the following example in French: *"C'est en outre ..."*. While our tokenizer would designate "C" as the first token in the sequence, a different tokenizer might designate "C'" or even "C'est". This would lead to a wrongly predicted sentence boundary when compared to the gold annotations, although the prediction was actually correct.

True and predicted labels for each document type are compared using Scikit-Learn to calculate binary F1-Scores. Scores are averaged for subsets: "Laws" encompass Criminal Code, Civil Code, and Constitution; "Judgments" include various court decisions.

We trained each CRF model once and the BiLSTM-CRF and transformer models 5 times with random seeds, reporting the mean performance including standard deviation. If not specified differently, reported values are binary F1-scores.

[9]https://github.com/jidasheng/bi-lstm-crf
[10]https://huggingface.co/distilbert-base-multilingual-cased
[11]For efficiency, we used a smaller model; a bigger model is advisable for future work.
[12]https://pytorch.org/
[13]https://huggingface.co/docs/accelerate/index
[14]This led to some wrongly predicted sentence boundaries, however this only occurred a few times and is therefore insignificant to the overall score.
[15]Unfortunately, to the best of our knowledge, so far there do not exist multilingually pretrained efficient transformer models.

Table 3: Mean (±std) F1 Score of baseline and multilingual models on all languages and the Portuguese zero-shot experiment. Best scores are in bold.

Language	French		Spanish		Italian		English	German		Portuguese (Zero-shot)	
Type / Model	Judg.	Laws	Judg.	Laws	Judg.	Laws	Judg.	Judg.	Laws	Judg.	Laws
CoreNLP	74.7	76.7	71.4	89.0	79.8	75.6	81.7	69.0	64.0	-	-
NLTK	72.5	75.8	70.2	89.2	72.3	66.3	77.2	72.3	73.8	64.9	57.0
NLTK-train	82.9	75.8	72.1	81.6	84.8	77.5	84.9	74.2	73.5	71.7	64.3
Spacy	86.6	67.2	60.0	70.3	73.9	73.7	79.7	87.5	67.0	59.0	77.7
Stanza	81.9	81.0	83.2	90.2	85.7	87.4	92.3	72.6	64.7	88.6	73.4
CRF	97.8	98.1	94.8	98.9	97.3	97.7	95.1	95.2	91.6	90.2	78.6
BiLSTM-CRF	97.6±0.3	**98.5**±0.2	97.3±0.1	**99.3**±0.2	97.8±0.1	**99.2**±0.1	95.4±0.3	**97.2**±0.2	97.5±0.5	93.0±0.6	73.2±3.3
Transformer	**98.3**±0.1	98.1±0.2	**97.8**±0.1	99.0±0.0	**98.3**±0.1	99.1±0.1	**97.0**±0.1	92.9±0.2	**97.6**±0.1	**93.6**±0.3	**91.3**±1.1

5 RESULTS

5.1 Baseline Models

The performance of baseline models in Section 4 on each language in our dataset is summarized in the upper section of Table 3.

The results for the baseline models are clearly lower than the reported scores for user-generated content by Read et al. [26], supporting the hypothesis that the performance of out-of-the-box models is subpar on legal data for all tested languages. The difference in performance could be explained in one part by the special sentence structures presented in Section 3.1, while the challenging nature of legal text accounts for another part.

Of interest is the gap between NLTK and NLTK-train in most languages, as training NLTK improves its ability to recognize and correctly predict abbreviations. This showcases that abbreviations are one part of the challenging nature of legal texts. To note here is that Spacy uses a slightly different notion of a sentence compared to the other models: Usually, when two sentences are separated by a newline character, the newline character would not be part of any sentence span, however Spacy would include it in the span of the second sentence. This leads to a false prediction, even though Spacy correctly recognized that there are two sentences. Therefore, the scores Spacy achieves are lower than expected.

5.2 Monolingual Models

We report the performance of our trained monolingual models in Table 4. Each model was trained and tested on the same language.

We observe that each model's performance, when applied to their training subset, reaches high nineties for almost all languages, significantly improving over the baseline models from Section 5.1 and comparable to reported SBD system performance on English news articles [26]. Our models also perform similarly to the reported performance of CRFs and CNNs on English [27, 29], as well as CRFs and NNs on German datasets [10].

Comparing the performance of the models when trained on one subset and evaluated on the other unseen set, i.e. a zero-shot experiment, the transformer model outperforms CRF and BiLSTM-CRF on most languages, dropping down to 81.8% on the Italian dataset, comparable to the best baseline models, when trained on judgements and evaluated on laws. Unsurprisingly, the models' performance in the zero-shot experiment is almost always lower than the performance on the subset they were trained on. This gap can be explained by the large difference of writing and formatting styles between judgements and laws, with the transformer model being the best at generalizing knowledge between the two subsets. We further hypothesize that it was easier for the models to generalize their knowledge to different domains, when being trained on judgements, than when being trained on laws, resulting in higher scores on unseen data. One factor here might be that legal text in judgements contain a higher variety of different sentence structures, while laws usually reuse the same structures.

The CRF and BiLSTM-CRF model showcase especially poor performance on the Spanish dataset when trained on laws and evaluated on judgements, with scores down to 43.4% and 54.3%. We hypothesize that both models possess a worse ability to generalize to different domains compared to transformer models.

To conclude, while training on both laws and judgments together not always produces the absolute best performance, it is most robust and does not result in performance degradation.

5.3 Multilingual Models

The performance of our multilingual models trained on laws and judgements is reported in the lower section of Table 3. Each multilingual model was trained on all languages except Portuguese.

The multilingual models clearly outperform the baseline models by a large margin, with F1-scores up to 99.2%. Both the BiLSTM-CRF and transformer models perform very well, with transformers performing slightly better on judgements and BiLSTM-CRFs on laws. The CRF model is close behind the other two, mostly reaching scores in the higher nineties. Comparing the performance of the multilingual models to the monolingual models, showcases that there is no loss of performance when training on a much larger dataset, with multilingual models performing comparably or in case of the transformer and BiLSTM-CRF model even better than the monolingual models on each respective language.

5.4 Zero-shot Experiment on Portuguese Data

We conducted a more challenging experiment, evaluating multilingual models on Portuguese data, comparing them to the baseline. Figure 1 provides an overview, while Table 3 details the differences in judgements and laws against the baseline.

Table 4: Mean (±std) F1 Score of monolingual models on their respective language. Best scores are in bold.

Model	Language / Type / Trained on	French Judg.	French Laws	Spanish Judg.	Spanish Laws	Italian Judg.	Italian Laws	English Judg.	German Judg.	German Laws
CRF	Judg.	97.9	73.2	97.0	98.3	**98.5**	95.6	96.8	97.8	76.5
	Laws	78.5	**98.8**	54.3	**99.6**	88.6	**99.6**	-	75.8	**97.7**
	Laws + Judg.	97.8	**98.8**	97.0	99.5	98.3	99.5	-	97.2	97.2
BiLSTM-CRF	Judg.	97.3±0.3	56.7±3.0	94.7±0.5	92.1±0.9	95.9±0.3	71.3±2.2	**97.3**±0.4	97.0±0.3	76.9±0.4
	Laws	66.1±4.2	97.9±0.2	43.4±6.8	98.7±0.2	74.1±1.2	98.4±0.2	-	71.9±2.5	97.3±0.3
	Laws + Judg.	97.0±0.4	98.1±0.1	95.6±0.5	98.9±0.4	96.2±0.2	98.2±0.1	-	97.2±0.2	97.6±0.2
Transformer	Judg.	98.2±0.1	84.7±1.2	96.9±0.2	96.9±0.4	97.8±0.2	81.8±0.9	96.5±0.1	98.0±0.2	87.2±0.4
	Laws	92.4±0.5	97.6±0.4	89.5±0.6	97.1±3.7	89.4±0.7	98.8±0.5	-	89.4±0.5	97.4±0.1
	Laws + Judg.	**98.4**±0.1	98.2±0.2	**97.3**±0.1	99.0±0.1	97.1±0.3	99.1±0.1	-	**98.3**±0.1	97.5±0.2

Table 5: Mean F1 Score of monolingual and multilingual models on unseen Portuguese data

Model / Type / Model Language	CRF Judg.	CRF Laws	BiLSTM-CRF Judg.	BiLSTM-CRF Laws	Transformer Judg.	Transformer Laws
French	79.3	75.4	25.5	51.7	82.5	87.1
Spanish	**91.5**	79.4	80.3	**73.5**	88.0	**94.0**
Italian	81.8	**83.3**	12.6	64.8	70.0	73.7
English	90.6	72.1	80.6	62.4	87.6	89.9
German	59.0	25.2	43.6	30.3	79.9	71.1
Multilingual	90.2	78.6	**93.0**	73.2	**93.6**	91.3

For judgements performance is adequate with F1-scores between 90.2% and 93.6%, comparable to user-generated content [26], and outperforming most baselines. However, for laws, only the transformer model scores in the lower nineties, while CRF and BiLSTM-CRF drop to 78.6% and 73.2%, respectively, similar to our usual baseline values. The transformer model's large-scale multilingual pretraining likely makes it more robust to distribution shifts, leading to better cross-lingual transfer to unseen languages than CRFs or BiLSTM-CRFs.

The difficulty of the writing and formatting style in Portuguese law texts could explain the difference between laws and judgements, indicated by lower than usual Portuguese baseline performance. BiLSTM-CRF's reduced performance could also result from the lack of Portuguese word embeddings used in training, as we only extracted embeddings from our training data. To improve BiLSTM-CRF models, future research could explore adding Portuguese word embeddings or using larger, multilingual embedding vocabularies during training. To improve transformer models, fine-tuning larger pre-trained models like XLM-RoBERTa [5] on the SBD task could be a potential avenue as they improve significantly in cross-lingual transfer compared to mBERT [6] or DistilBERT [28] models.

When evaluating the effectiveness of monolingual and multilingual models, trained on the entire monolingual dataset, on previously unseen Portuguese data (Table 5), we observe that the multilingual models outperform corresponding monolingual models in most languages, with Spanish being a notable exception. We hypothesize that the disparity in performance is due to close linguistic ties between Spanish and Portuguese, which enabled the Spanish monolingual models to excel in cross-lingual transfer. However, on other languages linguistically less close to Spanish, the multilingual model is expected to perform better than the monolingual ones.

5.5 Inference Time

Table 6 reports the inference times of our multilingual models trained on laws and judgments. We measured inference time three times on both a GPU (NVIDIA GeForce RTX 3060 TI) and a CPU (Intel Core i5-8600K CPU @ 3.60GHz), and show the average. We did not report standard deviation since there were no significant outliers. Notably, the transformer model saw significant improvements in inference time on a GPU. However, CRF does not benefit from GPU evaluation as it uses sequential operations.

Table 6: Mean inference time in minutes (min), seconds (s), milliseconds (ms) for each multilingual model to predict the entire dataset of ~130000 sentences and one sentence, measured on a GPU and CPU

Model	full dataset (~130000 sentences) CPU	full dataset (~130000 sentences) GPU	One sentence CPU	One sentence GPU
CRF	11 min 57 sec	-	~5.37 ms	-
BiLSTM-CRF	10 min 6 sec	9 min 23 sec	~4.54 ms	~4.21 ms
Transformer	34 min 26 sec	9 min 18 sec	~15.47 ms	~4.18 ms

Considering the results presented in Sections 5.2, 5.3 and 5.4, inference times and ease of use, a recommendation for the multilingual transformer model can be made for most cases, as long as a GPU is available for inference. For language specific tasks or tasks requiring longer input texts, we recommend the CRF models for the respective language, although they have a longer setup time compared to the BiLSTM-CRF and transformer model.

5.6 Error Analysis

We inspected random samples – two thirds of the Portuguese dataset (8 judgements, 20 laws) – predicted by the multilingual transformer model for the zero-shot experiment on Portuguese texts. We selected the multilingual transformer following our recommendation in Section 5.5, and the Portuguese dataset because the model already performed very well on the other datasets.

Standard sentence boundaries are rarely missed and the model performs adequately in that regard; yet, we identified a few sources of common mistakes. We discuss examples with |T| and |P| indicating true and predicted sentence boundaries, respectively. Many errors stem from citations and parentheses as shown in the example below:

- (Bittar, Carlos Alberto. *|P|* Direito de autor. *|P|* Rio de Janeiro: Forense Universitária, 2001, p. 143) *|T| |P|*

In this example, we have a citation sentence with periods being wrongly predicted as sentence boundaries inside the citation.

Another source of errors are datafields and headlines, since there is often little indication e.g., a sentence-terminating character, for the model to recognize it as such:

(1) RELATOR: MINISTRO SIDNEI BENETI *|T|*
(2) ACÓRDÃO *|T|*

The model failed to predict a sentence boundary at the end of both sequences. The errors showcased in the examples above mainly stem from our particularly defined sentence structures (Section 3.1.1) as well as the challenging nature of the legal SBD task.

Another set of errors were caused by the different formatting styles and words used in the Portuguese language, unknown to the model, such as:

(1) A Turma, por unanimidade, deu provimento ao recurso especial, nos termos do voto do(a) Sr(a). *|P|* Ministro(a) Relator(a). *|T| |P|*
(2) Exmos. *|P|* Desembargadores MAURÍCIO PESSOA (Presidente), CLAUDIO GODOY E GRAVA BRAZIL. *|T| |P|*

In (1), we have the abbreviation "Sr(a)", which the model did not recognise as such, thus marking the period as a sentence boundary. A similar mistake is shown in (2), with the abbreviation "Exmos".

5.7 Limitations

Due to the language skills of our annotator, we only annotated data from two language groups (Germanic and Italic). Therefore, our languages have high lexical overlap, making cross-lingual transfer comparatively easy. Future work may investigate legal text from additional diverse language groups to build systems even more robust towards language distribution shifts.

The annotator is a native German speaker, with intermediate French language skills. Due to the similarity of Italian, Spanish, and Portuguese to French, and because the SBD task is largely structural, the annotations were possible. However, having the annotations performed by a native speaker in the respective languages may further increase annotation quality. On the other hand, having one annotator (as done in our case) annotate the entire dataset, enables more consistency across languages.

Because of financial limitations, we performed the annotations using only one annotator. Having a second annotator validate the annotations may further increase annotation quality.

Augmenting the qualitative error analysis from Section 5.6 quantitatively may provide more concrete and actionable evidence for improving the systems further. To achieve this, a more detailed annotation of the sentence type would be helpful, so statistics over the sentences can be computed to get quantitative results of the sentence types performing worst.

6 CONCLUSION AND FUTURE WORK

6.1 Answers to the Research Questions

RQ1: *What is the performance of existing SBD systems on legal data in French, Spanish, Italian, English, and German?*
Existing SBD systems are subpar in all tested languages, lower than reported scores by Read et al. [26] on user-generated content, indicating that SBD is not solved in the legal domain.

RQ2: *To what extent can we improve upon this performance by training mono- and multilingual models based on CRF, BiLSTM-CRF and transformers?*
The monolingual models achieved state-of-the-art F1-scores in the high nineties for all tested languages, comparable to reported scores on news articles [26]. The multilingual models performed similarly to monolingual models, demonstrating the potential of training with larger datasets. The transformer model exhibited superior cross-domain transfer compared to CRF and BiLSTM-CRF models.

RQ3: *What is the performance of the multilingual models on unseen Portuguese legal text, i.e., a zero-shot experiment?*
The transformer models performs adequately on the judgements and laws subsets, reaching F1-scores in the lower nineties, demonstrating the best cross-lingual transfer, while the CRF and BiLSTM-CRF models perform decently around 90% on judgements, but drop down to baseline values on the laws, most likely requiring additional optimization.

6.2 Conclusion

In this work, we curated and publicly released a diverse legal dataset with over 130'000 annotated sentences in 6 languages, enabling further research in the legal domain. Using this dataset, we showed that existing SBD methods perform poorly on multilingual legal data, at most reaching F1-scores in the low nineties. We trained and evaluated mono- and multilingual CRF, BiLSTM-CRF and transformer models, achieving binary F1-scores in the higher nineties on our dataset, demonstrating state-of-the art performance. For a more challenging task, we tested our multilingual models in a zero-shot experiment on unseen Portuguese data, with the transformer model reaching scores in the lower nineties, outperforming the baseline trained on Portuguese texts as well as the CRF and BiLSTM-CRF models by a large margin. We publicly release these models and the code for further use and research in the community.

6.3 Future Work

Further improvement for all models might be achieved by preprocessing the input text more, e.g., replacing newlines with spaces, special characters with more widely used equivalent characters e.g., double quotes (“) with single quotes ("). Furthermore, thorough hyperparameter optimization tailored to the specific dataset could improve multilingual CRF and BiLSTM-CRF models. Finally, transformer models may benefit from legal-oriented models [4, 13, 22], larger pre-trained models like BERT [6], or models designed for cross-lingual transfer tasks, like XLM-RoBERTa [5].

Augmenting the dataset with legal texts from multiple languages and documents from various sources like privacy policies and terms of service may improve multilingual models' performance, particularly in the zero-shot scenario. An interesting impact on the model performance could be observed if the sentence spans were labeled with their sentence structure type such as "Citation" (Section 3.1.1) during training instead of being assigned a single label.

An investigation into whether the positive cross-lingual transfer observed in their study also applies to languages from a different family, such as Hungarian. This assumption is based on the common origin of the languages studied, as mentioned in Section 5.

REFERENCES

[1] Dennis Aumiller, Ashish Chouhan, and Michael Gertz. 2022. EUR-Lex-Sum: A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. http://arxiv.org/abs/2210.13448 arXiv:2210.13448 [cs].

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. arXiv:2004.05150 [cs]

[3] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv:2109.00904 [cs]* (Sept. 2021). http://arxiv.org/abs/2109.00904 arXiv: 2109.00904.

[4] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559 [cs]* (Oct. 2020). http://arxiv.org/abs/2010.02559 arXiv: 2010.02559.

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. https://doi.org/10.48550/arXiv.1911.02116 arXiv:1911.02116 [cs]

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Jinhua Du, Yan Huang, and Karo Moilanen. 2019. AIG Investments.AI at the FinSBD Task: Sentence Boundary Detection through Sequence Labelling and BERT Fine-tuning. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. Macao, China, 81–87.

[8] Dan Gillick. 2009. Sentence Boundary Detection and the Problem with the U.S.. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*. Association for Computational Linguistics, Boulder, Colorado, 241–244.

[9] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 42–47.

[10] Ingo Glaser, Sebastian Moser, and Florian Matthes. 2021. Sentence Boundary Detection in German Legal Documents. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. 812–821. https://doi.org/10.5220/0010246308120821

[11] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (May 2012).

[12] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. (2020). https://doi.org/10.5281

[13] Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber. 2022. LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension. https://doi.org/10.48550/arXiv.2212.08204 arXiv:2212.08204 [cs].

[14] Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. 2023. Natural Language Processing in the Legal Domain.

[15] Tibor Kiss and Jan Strunk. 2006. Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics* 32, 4 (Dec. 2006), 485–525. https://doi.org/10.1162/coli.2006.32.4.485

[16] Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand, 79–86.

[17] Chun-Wei Lin, Yinan Shao, Ji Zhang, and Unil Yun. 2020. Enhanced Sequence Labeling Based on Latent Variable Conditional Random Fields. *Neurocomputing* 403 (May 2020). https://doi.org/10.1016/j.neucom.2020.04.102

[18] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Baltimore, Maryland, 55–60. https://doi.org/10.3115/v1/P14-5010

[19] Ditty Mathew and Chinnappa Guggilla. 2019. AI_Blues at FinSBD Shared Task: CRF-based Sentence Boundary Detection in PDF Noisy Text in the Financial Domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*. Macao, China, 130–136.

[20] Denis Newman-Griffis, Chaitanya Shivade, Eric Fosler-Lussier, and Albert Lai. 2016. A Quantitative and Qualitative Evaluation of Sentence Boundary Detection for the Clinical Domain. *AMIA Joint Summits on Translational Science proceedings. AMIA Summit on Translational Science* 2016 (July 2016), 88–97.

[21] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 19–35. https://aclanthology.org/2021.nllp-1.3

[22] Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? https://doi.org/10.48550/arXiv.2211.17135 arXiv:2211.17135 [cs].

[23] Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. https://doi.org/10.48550/arXiv.2301.13126 arXiv:2301.13126 [cs].

[24] Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online only, 32–46. https://aclanthology.org/2022.aacl-main.3

[25] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

[26] Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jrgen Solberg. 2012. Sentence Boundary Detection: A Long Solved Problem?. In *Proceedings of COLING 2012: Posters*. Mumbai, India, 985–994.

[27] George Sanchez. 2019. Sentence Boundary Detection in Legal Text. In *Proceedings of the Natural Legal Language Processing Workshop 2019*. Association for Computational Linguistics, Minneapolis, Minnesota, 31–38. https://doi.org/10.18653/v1/W19-2204

[28] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. *ArXiv* abs/1910.01108 (2019).

[29] Jaromir Savelka, Vern R Walker, Matthias Grabmair, and Kevin D Ashley. 2017. Sentence Boundary Detection in Adjudicatory Decisions in the United States. *TAL* 58, 21 (2017).

[30] Jaromir Savelka, Hannes Westermann, Karim Benyekhlef, Charlotte S. Alexander, Jayla C. Grant, David Restrepo Amariles, Rajaa El Hamdani, Sébastien Meeùs, Aurore Troussel, Michał Araszkiewicz, Kevin D. Ashley, Alexandra Ashley, Karl Branting, Mattia Falduti, Matthias Grabmair, Jakub Harašta, Tereza Novotná, Elizabeth Tippett, and Shiwanni Johnson. 2021. Lex Rosetta: Transfer of Predictive Models across Languages, Jurisdictions, and Legal Domains. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. ACM, São Paulo Brazil, 129–138. https://doi.org/10.1145/3462757.3466149

[31] Stefan Schweter and Sajawel Ahmed. 2019. Deep-EOS: General-Purpose Neural Networks for Sentence Boundary Detection. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS)*. 5.

[32] Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. Fast WordPiece Tokenization. (2020), 2089–2103. https://doi.org/10.48550/ARXIV.2012.15524

[33] Jorg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)* (2012), 2214–2218.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762 arXiv:1706.03762 [cs]

LEGALBENCH: A COLLABORATIVELY BUILT BENCHMARK FOR MEASURING LEGAL REASONING IN LARGE LANGUAGE MODELS

Neel Guha[‡], Julian Nyarko[∗1], Daniel E. Ho[∗1], Christopher Ré[∗1], Adam Chilton[2], Aditya Narayana[3], Alex Chohlas-Wood[1], Austin Peters[1], Brandon Waldon[1], Daniel N. Rockmore[4], Diego Zambrano[1], Dmitry Talisman[3], Enam Hoque[5], Faiz Surani[1], Frank Fagan[6], Galit Sarfaty[7], Gregory M. Dickinson[8], Haggai Porat[9], Jason Hegland[1], Jessica Wu[1], Joe Nudell[1], Joel Niklaus[1], John Nay[10], Jonathan H. Choi[11], Kevin Tobia[12], Margaret Hagan[13], Megan Ma[10], Michael Livermore[14], Nikon Rasumov-Rahe[3], Nils Holzenberger[15], Noam Kolt[7], Peter Henderson[1], Sean Rehaag[16], Sharad Goel[17], Shang Gao[20], Spencer Williams[18], Sunny Gandhi[19], Tom Zur[9], Varun Iyer , and Zehua Li[1]

[1]Stanford University, [2]University of Chicago, [3]Maxime Tools, [4]Dartmouth College, [5]LawBeta, [6]South Texas College of Law Houston, [7]University of Toronto, [8]St. Thomas University Benjamin L. Crump College of Law, [9]Harvard Law School, [10]Stanford Center for Legal Informatics - CodeX, [11]University of Southern California, [12]Georgetown University Law Center, [13]Stanford Law School, [14]University of Virginia, [15]Télécom Paris, Institut Polytechnique de Paris, [16]Osgoode Hall Law School, York University, [17]Harvard Kennedy School, [18]Golden Gate University School of Law, [19]Luddy School of Informatics - Indiana University Bloomington, [20]Casetext

August 23, 2023

ABSTRACT

The advent of large language models (LLMs) and their adoption by the legal community has given rise to the question: what types of legal reasoning can LLMs perform? To enable greater study of this question, we present LEGALBENCH: a collaboratively constructed legal reasoning benchmark consisting of 162 tasks covering six different types of legal reasoning. LEGALBENCH was built through an interdisciplinary process, in which we collected tasks designed and hand-crafted by legal professionals. Because these subject matter experts took a leading role in construction, tasks either measure legal reasoning capabilities that are practically useful, or measure reasoning skills that lawyers find interesting. To enable cross-disciplinary conversations about LLMs in the law, we additionally show how popular legal frameworks for describing legal reasoning—which distinguish between its many forms—correspond to LEGALBENCH tasks, thus giving lawyers and LLM developers a common vocabulary. This paper describes LEGALBENCH, presents an empirical evaluation of 20 open-source and commercial LLMs, and illustrates the types of research explorations LEGALBENCH enables.

Contents

1 Introduction

Advances in large language models (LLMs) are leading American lawyers and administrators to reexamine the practice of law [53, 63, 160, 57].[2] Proponents have argued that LLMs could alter how lawyers approach tasks ranging from brief writing to corporate compliance [160]. By making legal services more accessible, they could eventually help alleviate the United States' long standing access-to-justice crisis [36, 134]. This perspective is informed by the observation that LLMs possess special properties which, it is argued, make them more suited for legal tasks. The models' capacity to learn new tasks from limited labeled data would reduce the manual data annotation costs that ordinarily burden the development of legal language models [13]. Their apparent proficiency at sophisticated reasoning tasks would also make them ideal for the rigor of law, which requires parsing obtuse texts with heavy jargon, and inferential processes which combine different modalities of reasoning [157].

This excitement, however, is tempered by the fact that legal applications often involve significant risk [48]. Existing work has shown that LLMs are capable of generating content that is offensive, misleading, and factually incorrect [10, 80]. Such behaviors—if replicated in legal applications [114]—could result in substantial harms [146], with much of the potential burden imposed on traditionally marginalized and under-resourced populations [127, 138]. The safety implications thus create a pressing need to develop infrastructure and processes for benchmarking LLMs in legal contexts.

However, significant challenges face practitioners seeking to assess whether LLMs can perform legal reasoning. The first challenge is the limited ecosystem of legal benchmarks [157]. The majority of existing benchmarks, for example, focus on tasks which models learn by finetuning or training on task-specific data [21]. These benchmarks do not measure the aspects of LLMs which generate excitement for law—namely, their ability to perform many different tasks using only few-shot prompts. Relatedly, benchmarking efforts have focused on professional certification exams like the Uniform Bar Exam [71], but these are not always representative of the actual use-cases for LLMs. The second challenge is the incongruity between the ways in which existing benchmarks and lawyers frame "legal reasoning." Existing benchmarks coarsely generalize all tasks involving legal data or laws as measuring "legal reasoning." In contrast, lawyers recognize that legal reasoning is a broad umbrella term encompassing many distinct types of reasoning [47]. Different legal tasks require different skills and bodies of knowledge. Because existing legal benchmarks fail to draw these distinctions, it is difficult for legal professionals to contextualize the performance of modern LLMs within their own understanding of legal competency. In short: legal benchmarks do not use the same vocabulary or conceptual frameworks as the legal profession.

In light of these limitations, we believe that rigorously evaluating the legal reasoning capabilities of LLMs will require the legal community to take a more proactive role in the process of benchmarking. To that end, we present LEGALBENCH: the first steps towards constructing an interdisciplinary collaborative legal reasoning benchmark for the English language.[3] Over the past year, the authors of this paper—drawing from their diverse legal and computer science backgrounds—came together to assemble 162 tasks (from 36 different data sources), each of which measures a specific type of legal reasoning. LEGALBENCH is thus, to the best of our knowledge, the first *open-source legal benchmarking effort*. We believe that this style of benchmark construction—where domain experts take an active and participatory role in the crafting of evaluation tasks—illustrates one approach to interdisciplinary collaboration in LLM research. Importantly, we believe it also shows that legal professionals have an essential role to play in the assessment and development of LLMs for law.

As a research project, we highlight three components of LEGALBENCH:

[2]In using "LLMs", we are referring to language models which evince in-context learning capabilities (also referred to as "foundation models" [13]). This behavior has traditionally been observed in models with at least a billion parameters.

[3]`https://github.com/HazyResearch/legalbench/`

1. LEGALBENCH was constructed from a mix of existing legal datasets (restructured for the few-shot LLM paradigm), and hand-crafted datasets created and contributed by legal professionals (included as authors on this work). The legal professionals involved in this collaboration were asked to contribute datasets that they believed to either measure an interesting legal reasoning skill, or to capture a practically useful application for LLMs in the law. High performance on LEGALBENCH tasks thus provides useful information, allowing lawyers to validate their assessment of an LLM's legal competency, or identify an LLM that could be used in their workflow.
2. LEGALBENCH tasks are organized into an extensive typology which describes the types of legal reasoning required to perform the task. Because this typology is drawn from frameworks familiar to the legal community, it enables legal professionals to meaningfully engage in discussions of LLM performance, using a terminology and conceptual framework familiar to them [47, 124].
3. Finally, LEGALBENCH is intended as a platform to support further research. For AI researchers who lack legal expertise, LEGALBENCH comes with significant support for understanding how to prompt and evaluate different tasks. And as more of the legal community begins to engage with the potential impact and role of LLMs, we hope to grow LEGALBENCH by continuing to solict and incorporate tasks from legal professionals.[4]

In this paper, we make the following contributions:

1. First, we present a typology for organizing and describing legal tasks in terms of the types of reasoning they require. This typology is drawn from frameworks lawyers use to describe legal reasoning [124].
2. Second, we provide an overview of the tasks in LEGALBENCH, describing the process by which they were constructed, important dimensions of heterogeneity, and limitations. A full description of each task is provided in the Appendix.
3. Finally, we use LEGALBENCH to evaluate 20 LLMs from 11 different families, across a range of size points. We make observations regarding the performance of different models and present an initial study into different prompt-engineering strategies. Ultimately, these results are intended to highlight different directions of future work that LEGALBENCH may enable.

We hope that this benchmark will be interesting to a diverse set of communities. Practitioners may use these tasks to determine whether and where LLMs can be integrated into existing workflows to improve outcomes for clients. Legal academics may benefit from observing the types of annotation that LLMs are capable of [159], and different forms of empirical scholarly work they may enable. Computer scientists may benefit from studying the performance of these models in a domain like law, where distinct lexical properties and unique tasks may surface new insights.

Before we progress further, we note that the purpose of this work isn't to evaluate whether computational systems *should* replace lawyers and legal officers, or to understand the positive and negative impacts of that replacement [48, 128, 4]. Rather, our goal is to construct artifacts that enable the relevant stakeholders and affected communities to better understand, *empirically*, the capacity for LLMs to perform different types of legal tasks. Given the proliferation of computational legal tools, we believe that answering this question is vital for ensuring their safe and ethical usage.

2 Related work

2.1 Legal reasoning benchmarks

Understanding the extent to which NLP models can perform tasks or skills traditionally associated with lawyers—or be useful in legal analysis—has been the focus of significant work [6, 110, 126, 84, 86, 104, 88, 22]. Researchers have approached this question in a variety of ways [72]. First, prior work has identified manually arduous tasks currently performed by lawyers—like forms of document review [62, 142] or case summarization [120, 121, 90, 69]—and developed benchmarks to assess the performance of current state-of-the-art techniques. Here, research has focused on the aspects of legal text which are often challenging for NLP methods, like the length of documents or the presence of jargon [21, 111, 87, 41, 79]. A second line of work has focused on developing tasks to evaluate forms of inferential reasoning common to law [21]. This includes, for instance, tasks which require a model to identify the best supporting statement for an argument [80, 157], or perform statutory reasoning [65]. Other work has focused on creating datasets for pretraining models [60, 129], non-English/multilingual tasks [95, 96, 66, 151, 52, 94, 70, 20, 23, 103, 18], legal judgement prediction [89, 38, 17, 158], legal role labeling [85], and different forms of retrieval [68].

[4]Cognizant of LEGALBENCH's current skew towards American law, we hope that additional contributions incorporate tasks from other jurisdictions.

Importantly, the majority of previous benchmarking efforts have focused on language models which learn by supervised training or finetuning (e.g., BERT variants [45]), and researchers have consequently studied questions related to the role of domain specific datasets [157, 19, 20]. More recently, researchers have begun to ask whether *large* language models (LLMs) like GPT-3/4 can perform legal reasoning [73, 153, 67, 12, 29, 31, 154], citing to evidence of these models' capacity to perform sophisticated reasoning tasks in domains like math or programming [145, 24]. Unlike BERT-based models, LLMs are evaluated on their ability to learn tasks *in-context*, primarily through prompting. While a few works have experimented with LLMs on existing benchmarks [16, 12], most evaluations focus on standardized tests or other exam equivalents [71, 93, 30]. Studies have explored the role of prompt-engineering [154, 153, 75], potential applications [93, 29, 147, 117, 116], questions regarding human-LLM interaction [31, 63], and comparisons to older finetuned-models [91].

LEGALBENCH builds on prior work in several ways. First, LEGALBENCH enhances opportunities to study legal reasoning in LLMs, by making available 162 evaluation tasks. LEGALBENCH systematizes and standardizes these tasks for LLM evaluation, specifying potential prompts, in-context demonstrations, and metrics. Second, LEGALBENCH presents a framework for organizing and comparing tasks, allowing researchers to identify trends in performance across groupings of tasks. This enables researchers, for instance, to distinguish between task types for which current LLMs are highly performant, and task types for which further work is needed.

A notable consequence of focusing on few-shot LLMs is that LEGALBENCH can contribute a much more diverse set of legal reasoning tasks. Traditional NLP methods require a large training set and a smaller evaluation set. The cost of legal annotations means that constructing benchmarks has required extraordinary financial investment [62, 120] or a "natural" source of existing annotations [157, 21]. Because the few-shot prompting regime requires only a few labeled demonstrations, creating large training sets isn't necessary, and the effort they otherwise would have consumed can be allocated towards developing new tasks.

2.2 Connections to other LLM benchmarking efforts

We highlight connections to two broader research efforts. First, we draw inspiration from existing efforts within NLP and machine learning to define fine-grained measures of performance, which allow researchers to discuss model capabilities with precision and specificity. Examples include the diagnostic set of the GLUE Benchmark [141], the "reasoning patterns" studied in [100], the task organization used in HELM [80], and the BigBench effort [123]. Fine-grained measurements are valuable because they allow researchers to identify how particular modifications to model architectures or training regimes affect performance. They hold particular value for the field of legal NLP, in which researchers continue to debate how best to specialize language models to the domain [60, 157, 56]

We additionally draw inspiration from other large-scale collaborative efforts in AI, including the BigBench project [123], and studies in medicine [28]. In particular, we believe that LEGALBENCH illustrates a new model of open-source and interdisciplinary collaboration between the legal and AI communities. To the extent that LLMs gain adoption for legal tasks, legal professionals will be primarily charged with supervising them and selecting application use-cases. Involving the legal community in the design and construction of evaluation tasks allows for the construction of benchmarks which are more responsive to their interests and information needs.

3 The LEGALBENCH typology

LEGALBENCH identifies six types of legal reasoning that LLMs can be evaluated for: (1) issue-spotting, (2) rule-recall, (3) rule-application, (4) rule-conclusion, (5) interpretation, and (6) rhetorical-understanding. We first justify the selection of these types by providing background on how the legal profession frames "legal reasoning," and the connections to our typology. We then illustrate how task datasets may be used to evaluate LLMs for each type, using examples from LEGALBENCH.

Though this framework draws heavily on American legal thought, we find it can be easily extended to characterize LEGALBENCH tasks that implicate non-American bodies of law. We also note that our types are non-exhaustive, and in future work hope to consider additions to these types.

3.1 Frameworks for legal reasoning

IRAC American legal scholars often describe "legal reasoning" as the process of determining the legal conditions that arise from a set of events or occurrences, with reference to both prior cases and codified laws [47]. A common framework for executing this type of legal reasoning is the **I**ssue, **R**ule, **A**pplication and **C**onclusion (**IRAC**) framework [148, 124]. In this framework, legal reasoning decomposes into four sequential steps.

First, lawyers identify the legal issue in a given set of facts (**issue-spotting**). An issue is often either (1) a specific unanswered legal question posed by the facts, or (2) an area of law implicated in the facts. Depending on the setting, a lawyer may be told the issue, or be required to *infer* a possible issue.

Second, lawyers identify the relevant legal rules for this issue (**rule-recall**). A rule is a statement of law which dictates the conditions that are necessary (or sufficient) for some legal outcome to be achieved. In the United States, rules can come from a variety of sources: the Constitution, federal and state statutes, regulations, and court opinions (case law). Importantly, rules often differ between jurisdictions. Hence, the relevant rule in California might be different than the relevant rule in New York.

Third, lawyers apply these rules to the facts at hand (**rule-application**). Application, or the analysis of rule applicability, consists of identifying those facts which are most relevant to the rule, and determining how those facts influence the outcome under the rule. Application can also involve referencing prior cases involving similar rules (i.e. *precedent*), and using the similarities or differences to those cases to determine the outcome of the current dispute.

Finally, lawyers reach a conclusion with regards to their application of law to facts, and determine what the legal outcome of those facts are (**rule-conclusion**).

Example We illustrate this framework with a simple example. Suppose that BusinessMart—a large manufacturing corporation—is being sued by Amy in federal court on diversity jurisdiction.[5] BusinessMart sells the majority of its goods in Texas, has its headquarters (where its CEO and board members sit and work) in California, and maintains a factory in Florida. A court is trying to determine—for the purposes of diversity jurisdiction—where BusinessMart's "principal place of business is."

- Issue-spotting: Here, a narrow issue is offered—where is BusinessMart's principal place of business?
- Rule-recall: A lawyer would recognize that the most relevant rule here comes from the case *Hertz Corp. v. Friend*,[6] in which the Supreme Court determined "that the phrase 'principal place of business' refers to the place where the corporation's high level officers direct, control, and coordinate the corporation's activities."
- Rule-application: Applying this rule to the facts above yields two observations. First, a corporation's CEO and board members are examples of high level officers referred to in *Hertz* that control and conduct a company. Second, the place where BusinessMart's high level officers control the company is California, as that is where the CEO and board sit and work.
- Rule-conclusion: Based on the chain of inference spelled out in the application stage, a lawyer would thus conclude that California is BusinessMart's principal place of business.

The extent to which the outcome of the application and conclusion steps follow each other is dictated by the level of ambiguity in the fact patterns. When the law on a particular question is clear and there is little ambiguity in the facts (as the case in the above example), then the application and conclusion steps point towards the same outcome. Sometimes however, the facts may be unclear or contested, and reasonable minds may differ as the conclusion step. For now, LEGALBENCH focuses entirely on the former setting (unambiguous answers), and all tasks are considered to have objectively "correct" answers.

Other types of reasoning Though IRAC is the most formal framework for legal reasoning, lawyers recognize a variety of skills which are useful to practice of law [47, 77]. For instance, lawyers are often required to exercise interpretive skills, in order to identify the rights, obligations, or limitations of certain legal language (e.g., what a contractual clause may or may not enable). They must also exhibit rhetorical skills, and understand the types of arguments that are made. Though these tasks require the knowledge base and skill set of lawyers, they, arguably, do not always fit neatly within the IRAC framework. Hence, we consider these to be distinct from the examples offered in the previous section.

3.2 Evaluating legal reasoning in large language models

LEGALBENCH identifies six categories of legal reasoning. For each category, we describe how a LLM task may evaluate the typified legal reasoning, using examples from LEGALBENCH.

Issue-spotting LEGALBENCH evaluates issue-spotting through tasks in which an LLM must determine if a set of facts raise a particular set of legal questions, implicate an area of the law, or are relevant to a specific party. Issue tasks evaluate a LLM's ability to reason over the legal implications of different activities, events, and occurrences.

[5] Diversity jurisdiction gives federal courts the ability to hear cases between parties that are "citizens" of different states.

[6] Hertz Corp. v. Friend, 559 U.S. 77 (2010).

An example of an issue-spotting task is the `learned_hands_benefits` task, which requires an LLM to determine (Yes/No) whether a post on a public legal aid forum raises issues related to welfare law (i.e., public benefits or social services). The box below shows how a LLM might be prompted for this task.

Issue-spotting example: `learned_hands_benefits`

Does the post discuss public benefits and social services that people can get from the government, like for food, disability, old age, housing, medical help, unemployment, child care, or other social needs?

Post: "I am currently receiving support from social services, idk why, this is just how my life turned out. They have asked for all of my bank information for the past 12 months. I don't know what this means. Why would they want that?"
Answer: Yes

Rule-recall LEGALBENCH evaluates rule-recall through tasks which require the LLM to generate the correct legal rule on an issue in a jurisdiction (e.g., the rule for hearsay in US federal court). A rule task can be an open-ended generation task—in which the LLM must generate the text of the rule for a jurisdiction—or a classification task—in which the LLM must determine whether the rule exists in that jurisdiction. Anchoring to jurisdiction is important, as legal rules differ across different jurisdictions. Rule tasks are particularly useful for measuring *hallucinations* [81]. An example of a rule-recall task is `rule_qa`, a question-answer task where questions include asking the model to state the formulations for different legal rules, identify where laws are codified, and general questions about doctrine.

Rule-recall example: `rule_qa`

Question: What are the four requirements for class certification under the Federal Rules of Civil Procedure?"
Answer: Numerosity, commonality, typicality, adequacy

Rule-conclusion LEGALBENCH evaluates rule-conclusion through tasks which require an LLM to determine the legal outcome of a set of facts under a specified rule. LLMs are evaluated purely on whether their predicted outcome is correct. For example, the `ucc_v_common_law` task asks a LLM to determine whether a contract is governed by the Uniform Commercial Code (UCC) or the common law of contracts. The LLM is always provided with the relevant rule, via the prompt (see below).

Conclusion example: `ucc_v_common_law`

The UCC (through Article 2) governs the sale of goods, which are defined as moveable tangible things (cars, apples, books, etc.), whereas the common law governs contracts for real estate and services. For the following contracts, determine if they are governed by the UCC or by common law.

Contract: Alice and Bob enter into a contract for Alice to sell her bike to Bob for $50. Is this contract governed by the UCC or the common law?
Governed by: UCC

Rule-application LEGALBENCH evaluates rule-application through the same tasks used to measure rule-conclusion. When evaluating rule-application however, we prompt the LLM to provide an explanation of how the rule applies to a set of facts, and evaluate the quality of the generated explanation along two dimensions: (1) whether the explanation is *correct*, and (2) whether it contains *analysis*. Each metric captures a different dimension upon which a particular rule-application may be good.

Correctness corresponds to the criteria that explanations should not contain errors. We focus on five types of errors: misstatements of the legal rule, misstatements of the fact pattern, incorrectly asserting the legal outcome, logic errors, and arithmetic errors. Analysis corresponds to the criteria that explanations should contain inferences from the facts that are relevant under the rule, and illustrate how a conclusion is reached. Consider, for example, an explanation which restates the rule, the fact pattern, and the predicted legal outcome. If the predicted legal outcome is correct, than the explanation in its entirety would be correct, because it contains no error. However, as prior works have noted [71, 30], examples like this are conclusory, and often unsatisfactory in the context of legal work.

To standardize evaluation and enable future work, we have released an "answer guide" for each task used for rule-application, which contains the inferences required for each sample, and describes common modes of errors. All evaluations in LEGALBENCH for rule-application have been performed with respect to this answer-guide.

Table 1 presents an examples of how three generations (corresponding to the Alice/Bob example above) would be evaluated under the above metrics. The first generation is incorrect, because it misstates the rule. The second generation is correct because it contains no falsehoods, but performs no analysis because it does not articulate inferences. The third generation is both correct and contains analysis, because it has no errors, and explicitly mentions an essential inference (e.g., that a bike is a "good").

Interpretation LEGALBENCH evaluates interpretation through tasks which require the LLM to parse and understand a legal text. Interpretive tasks provide the LLM with a text, and ask the LLM to either extract a relevant piece of information, answer a

Incorrect	Correct, but no analysis	Correct and contains analysis
The contract is for Alice to sell her bike to Bob. The contract is governed by the common law, because all goods are governed by the common law.	The contract is for Alice to sell her bike to Bob. The contract is governed by the UCC, because the UCC governs all goods.	The contract is for Alice to sell her bike to Bob. The contract is governed by the UCC, because a bike is a good and all goods are governed by the UCC.

Table 1: An example of how different generations are evaluated for correctness and analysis.

question, or categorize the text by some property. Interpretive tasks are among the most studied and practically relevant tasks in LEGALBENCH, and many have been taken from actual use-cases. An example of an interpretive task is `cuad_audit_right`, which asks the LLM to determine if a contractual clause contains an "audit right." An example is shown below:

Interpretation example: `cuad_audit_right`

Does the clause give a party the right to audit the books, records, or physical locations of the counterparty to ensure compliance with the contract?

Clause: "We shall have the right at all times to access the information system and to retrieve, analyze, download and use all software, data and files stored or used on the information system."
Answer: Yes

Rhetorical-understanding LEGALBENCH evaluates rhetorical-understanding through tasks which require an LLM to reason about legal argumentation and analysis. In these tasks, an LLM is provided with a legal argument (usually excerpted from a judicial opinion), and asked to determine whether it performs a certain function or has a certain property. An example is the `definition_classification` task, in which an LLM must determine if a sentence from a judicial opinion provides a definition of a term.

Rhetorical-understanding example: `definition_classification`

Does the sentence define a term?

Sentence:"To animadvert carried the broader implication of "turn[ing] the attention officially or judicially, tak[ing] legal cognizance of anything deserving of chastisement or censure; hence, to proceed by way of punishment or censure." 1 Oxford English Dictionary 474 (2d ed.1989)."
Answer: Yes

We emphasize one aspect of LEGALBENCH: IRAC in this work is used as an organizing principle for grouping tasks. On a law exam, a student would be expected to generate an answer which structurally resembles IRAC, where each step builds on the inferences of the previous step [71, 30]. LEGALBENCH tasks, in contrast, each evaluate a single type of legal reasoning. Hence, a task like `learned_hands_benefits` can only be used to evaluate issue-spotting, and not rule-recall. In future work we hope to add tasks which evaluate multiple steps jointly.

4 LEGALBENCH tasks

Appendix F discusses each task in detail, providing a description of the reasoning that each task evaluates, how task data was constructed, task examples, and evaluation protocols. This section provides an overview of LEGALBENCH.

4.1 Construction process

Task sources LEGALBENCH tasks are drawn from three sources. The first source of tasks are existing available datasets and corpora. Most of these were originally released for non-LLM evaluation settings. In creating tasks for LEGALBENCH from these sources, we often significantly reformatted data and restructured the prediction objective. For instance, the original CUAD dataset [62] contains annotations on long-documents and is intended for evaluating extraction with span-prediction models. We restructure this corpora to generate a binary classification task for each type of contractual clause. While the original corpus emphasized the long-document aspects of contracts, our restructured tasks emphasize whether LLMs can identify the distinguishing features of different types of clauses. The second source of tasks are datasets that were previously constructed by legal professionals but never released. This primarily includes datasets hand-coded by legal scholars as part of prior empirical legal projects (e.g., [27]). The last category of tasks are those that were developed specifically for LEGALBENCH, by the authors of this paper. Overall, tasks are drawn from 36 distinct corpora.

Collaborative component In August 2022, we published a call for tasks, describing the goals of the project and its structure [59]. We publicized the project through mailing lists and legal computational conferences. Submitted tasks were vetted for legal correctness and task validity. Task contributors are drawn from diverse professional backgrounds within the law (e.g., academics, practitioners, computational legal researchers) and constitute the authors of this paper.

Infrastructure LEGALBENCH comes with support designed to enable non-law AI researchers to use and study LEGALBENCH tasks. First, each LEGALBENCH task is accompanied by extensive documentation describing how the task is performed, its legal significance, and the construction procedure. The objective of this documentation is to provide AI researchers with a working understanding of the mechanical processes behind each task, for the purposes of better understanding LLM performance. Second, each task is accompanied by a "base" prompt, which contains task instructions and demonstrations. The base prompt is provided to promote replicability and standardization. We anticipate that future research efforts building off of LEGALBENCH will identify higher performing prompts/prompt formats. We intended to update the LEGALBENCH GitHub repository with these prompts as they are discovered.

Limitations We note several limitations of the current LEGALBENCH tasks (additional limitations are noted in Appendix B). First, when this project began, most LLM context-windows were constrained to a few pages of text. As a result, the initial round of LEGALBENCH tasks does not involve longer documents. We hope to include such tasks in future work, particularly as recent technical developments have resulted in significantly longer context windows [42, 54, 43, 109]. Second, LEGALBENCH's tasks focus on legal reasoning questions with objectively correct answers. LEGALBENCH is thus not helpful for evaluating legal reasoning involving degrees of correctness or tasks where "reasonable minds may differ." Third, LEGALBENCH only considers English language tasks, is skewed towards certain jurisdictions (American law), and certain areas of the law (contracts). Thus, the current iteration of the benchmark limits inferences regarding how LLMs may generalize to legal tasks involving other jurisdictions. As we continue to solicit and incorporate contributions to LEGALBENCH, we hope to add tasks addressing these limitations. Finally, LEGALBENCH evaluates IRAC abilities independently, while law exams and other legal work requires lawyers to generate outputs which follow IRAC in a multi-hop matter (i.e., each aspect is applied to the same fact pattern).

4.2 Dimensions of variation

Task structure All LEGALBENCH tasks contain at least 50 samples, with an average task size of 563 samples (Appendix D.4). These tasks are comparable in size to those used in benchmarking efforts like BigBench [130], HELM [80] or RAFT [1]. LEGALBENCH tasks also span different formats: multiple-choice questions (35 tasks), open-generation (7 tasks), binary classification (112 tasks), and multi-class/multi-label classification (8 tasks).

Reasoning types and legal domains LEGALBENCH provides tasks for each of the reasoning categories discussed above: rule-recall (5 tasks), issue-spotting (16 tasks), rule-application (16 tasks), rule-conclusion (16 tasks), interpretation (119 tasks), and rhetorical-understanding (10 tasks). Tasks are predominantly drawn from areas of law implicating civil matters, including contracts (58 tasks), civil procedure (8 tasks), evidence law (1 task), and corporate law (58 tasks). The skew towards interpretation tasks and tasks from contract law can be explained by the ubiquity of legal documents from these areas (e.g., contracts, terms-of-service agreements, disclosures, and etc.) and their immediate commercial implications [62, 76].

Language variation Legal language is highly heterogeneous, varying in sentence structure, vocabulary, and rhetorical style across different legal areas and document types [60]. This poses a distinct challenge for LLMs, which are extremely sensitive to structure of input text and the vocabulary used [80]. LEGALBENCH tasks are drawn from a diverse set of legal language types, thus enabling researchers to study performance variation across different categories of legal text. Specifically, LEGALBENCH encompasses tasks with language drawn from plain English (32 tasks), legal opinions (11 tasks), merger agreements (34 tasks), contracts (55 tasks), statutory text (3), and other sources.

4.3 Tasks

We offer a brief summary of the tasks present in each reasoning category.

Issue-spotting There are 17 issue-spotting tasks. 16 tasks are derived from the "Learned Hands" Dataset (Section F.14). Each of these tasks is a binary classification task, in which the LLM must determine if a post from /r/legaladvice implicates a particular domain of law (e.g., immigration). The last task is the `corporate_lobbying` task (Section F.7), which requires determining if a legislative bill has legal implications for a described company.

Rule-recall There are 5 rule-recall tasks. Two tasks require an LLM to either generate the citation for a particular legal quote, or identify if a candidate citation is correct (Section F.3). The remaining three tasks are:

- `rule_qa`, in which the LLM must generate the text of different legal tests and identify where they're codified (Section F.25).
- `international_citizenship_questions`, in which the LLM must answer yes/no questions about citizenship requirements in different countries (Section F.13).

- `nys_judicial_ethics`, in which the LLM must answer yes/no questions corresponding to different ethical rules under the guidance provided by the New York State Advisory Committee on Judicial ethics (Section F.17).

Rule-application and rule-conclusion There are 12 tasks used for both rule-application and rule-conclusion.

- Six tasks evaluate an LLM's ability to apply the diversity jurisdiction test to information about plaintiff and defendant citizenships and the amount-in-controversy for different claims (Section F.9). This requires both arithmetic and logical reasoning. The simplest (`diversity_1`) involves one plaintiff, one defendant, and one legal claim. The most complex (`diversity_6`) involves two plaintiffs, two defendants, and two claims against each defendant.
- `abecrombie` evaluates an LLM's ability to apply the *Abercrombie* test to classify how distinctive a product/service name is for a particular product/service (Section F.1).
- `hearsay` evaluates an LLM's ability to identify—given a particular piece of evidence and an issue being litigated—whether the evidence would count as hearsay for that issue (i.e., an out-of-court statement introduced to prove the truth of the matter asserted) (Section F.11).
- `personal_jurisdiction` evaluates an LLM's ability to identify when a court in a particular forum may excercise personal jurisdiction over a defendant, given basic facts about the defendant's place of domicile, their interactions with the state, and the claims brought against them by plaintiffs (Section F.21).
- `successor_liability` evaluates an LLM's ability to identify the potential successor liability exceptions present in fact patterns describing a sale of assets from one company to another (Section F.29).
- `telemarketing_sales_rule` evaluates an LLM's ability to identify whether the representations made by a company covered under the Telemarketing Sales Rule violate either 16 C.F.R. § 310.3(a)(1) and 16 C.F.R. § 310.3(a)(2), which outline a series of specific telemarketing sales practices defined as "deceptive" (Section F.31).
- `ucc_v_common_law` evaluates an LLM's ability to determine whether a particular contract is covered by the Uniform Commercial Code (UCC) or the common law, given information about the contract (Section F.33).

Interpretation There are 118 interpretation tasks.

- `consumer_contracts_qa`, which evaluates an LLM's ability to determine the rights/obligations imposed by terms of service clauses from popular websites (Section F.5).
- `contract_qa`, which evaluates an LLM's ability to identify different types of contractual provisions.
- 14 tasks designed from the ContractNLI dataset [74]. Each task evaluates an LLM's ability to identify whether a candidate contract excerpt adheres to a task-specific assertion (Section F.6).
- 38 binary-classification tasks designed from the CUAD dataset [62]. Each task evaluates an LLM's ability to identify whether a candidate contractual clause is of a certain type (Section F.4.1).
- `insurance_policy_interpretation`, which evaluates an LLM's ability to determine whether a particular claim is covered by an insurance policy (Section F.12).
- `jcrew_blocker`, which evaluates an LLM's ability to identify whether a particular loan clause is a J.Crew Blocker provision (Section F.4.2).
- 34 tasks from the MAUD dataset [142], which evaluates an LLM's ability to answer multiple-choice questions about the content of excerpts from merger-agreements (Section F.16). Each task corresponds to a different question.
- 9 tasks from the OPP-115 dataset [149], each of which evaluates an LLM's ability to determine whether a privacy policy clause discusses a particular issue (Section F.18). Each task is a binary classification task corresponding to a different issue.
- `privacy_policy_entailment` [161], which evaluates an LLM's ability to answer entailment questions from privacy policies (Section F.22).
- `privacy_policy_qa` [112], which evaluates an LLM's ability to determine if a clause from a privacy policy contains the answer to a particular question (Section F.23).
- 2 tasks designed from the SARA dataset [65], which evaluate an LLM's ability to interpret and apply sections of the tax-code (Section F.26).
- 10 tasks which evaluate an LLM's ability to identify when a supply chain disclosure discusses or describes a particular type of information (Section F.30). Each task corresponds to a different disclosure objective.
- `unfair_tos` [82], which evaluates an LLM's ability to classify clauses from terms of service agreements into one of muliple categories (Section F.4.3).

Rhetorical-understanding There are 10 tasks which evaluate rhetorical-understanding.

- `canada_tax_court_outcomes` evaluates an LLM's ability to identify the outcome of a tax court decision, based on the text of the decision (Section 16).
- 2 tasks evaluate an LLM's ability to (1) identify sentences from US Supreme Court opinions which define a term, and (2) extract that term (Section F.8).
- `function_of_decision_section` evaluates an LLM's ability to identify the function that an excerpt of a legal opinion has (e.g., statement of rule) (Section F.10).
- `legal_reasoning_causality` evaluates an LLM's ability to identify when an excerpt of a court's opinion relies on statistical evidence (Section F.15).
- `oral_argument_question_purpose` evaluates an LLM's ability to identify the purpose that a particular question (from Supreme Court oral arguments) plays (Section F.19).
- `overruling` [157] evaluates an LLM's ability to identify when a sentence from a judicial opinion overrules a previous case (Section F.20).
- `scalr` evaluates an LLM's ability to assess which holding statement (amongst several options) best answers a provided legal question.
- 2 tasks evaluate an LLM's ability to identify whether excerpts of judicial reasoning rely on certain textualist tools (Section F.32). Each task corresponds to a different tool.

5 Results

We use LEGALBENCH to conduct a three-part study.

- In the first part (Section 5.2), we conduct a sweeping evaluation of 20 LLMs from 11 different families, at four different size points. We use this study to make initial observations on performance differences across families, the role of model size, and the gap between open-source and commercial LLMs.
- In the second part (Section 5.3), we show how LEGALBENCH can be used to conduct in-depth evaluations of models. To illustrate, we use LEGALBENCH to highlight similarities and differences in the performance of three popular commercial models: GPT-4, GPT-3.5, and Claude-1.
- In the final part (Section 5.4), we show how LEGALBENCH can support the development of law-specific LLM methods. We focus on prompting, and conduct a series of experiments that begin to surface tradeoffs and challenges with regards to guiding LLMs towards certain tasks.

Ultimately, our study here serves to illustrate the types of analyses that LEGALBENCH enables, and highlight potential directions for future work.

5.1 Setup

5.1.1 Models

Commercial models We study three commercial API-access models. From the OpenAI GPT family, we study GPT-3.5 [14] (text-davinci-003) and GPT-4 [98]. Results from these models were retrieved between May and August of 2023. From the Anthropic family, we study Claude-1 (v1.3) [3]. Results from this model were retrieved in July of 2023. These models are believed to be large (hundreds of billions of parameters), though exact details on their architecture and training process are unknown. It is thus possible that some LEGALBENCH tasks leaked into pretraining data. Details on the extent to which different LEGALBENCH tasks have been previously made available online can be found in Appendix D.

Open-source models We study 17 open-source models at three different size points: 3B parameters, 7B parameters, and 13B parameters. All inference was performed on two-GPU GCP 40GB A100s, using the Manifest library [99]. HuggingFace links for each model are provided in Appendix G.

- From Together, we study three models: Incite-Instruct-7B, Incite-Base-7B, and Incite-Instruct-3B [35, 135].
- From Meta's OPT family, we study three models: OPT-2.7B, OPT-6.7B, and OPT-13B [156].
- From TII's Falcon family, we study Falcon-7B-Instruct [2, 105].
- From MosaicML's MPT family, we study MPT-7B-8k-Instruct [131].
- From LMSYS' Vicuna family, we study Vicuna-7B-16k and Vicuna-13B-16k [26].
- From Google's FLAN-T5 family, we study Flan-T5-XL (3B parameters) and Flan-T5-XXL (11B parameters) [32].
- From Meta's LLama-2 family, we study LLaMA-2-7B, and LLaMA-2-13B [136].

- From the Wizard family, we study WizardLM-13B [152].
- From the BigScience BLOOM family, we study BLOOM-3b and BLOOM-7B [118].

Future work Our selected LLMs represent only a sample of the models available. For instance, we do not evaluate LLMs larger than 13B parameters, which have been observed to perform well [9]. Studied LLMs are also "general domain," in that we don't find evidence that any were specifically customized to perform well on legal text.[7] In future work we hope to expand our evaluation to a broader set of LLMs.

5.1.2 Prompts

We designed a prompt for each task by manually writing instructions for the task, and selecting between zero and eight samples from the available train split to use as in-context demonstration. The number of samples selected depended on the availability of data and the sequence length of samples (Appendix G.2). For instance, the inputs to the Supply Chain Disclosure tasks are disclosure statements between 1-2 pages long, making the inclusion of multiple demonstrations infeasible. For application evaluation, we augmented the prompt with an instruction for the LLM to explain its reasoning.

We used the same prompts across all LLMs with one exception. In contrast to the OpenAI and open-source LLMs, Anthropic recommends specific prompting formats when using Claude.[8] This includes surrounding in-context samples with `<example>//example>` tags, and adding instructions specifying the output space. We observed that failing to adhere to these guidelines led Claude to generate text which made extracting a prediction challenging. Therefore, when prompting Claude, we added example-tags to the in-context demonstrations and instructions specifying the prediction space (e.g., "Reply with either: generic, descriptive, suggestive, arbitrary, fanciful").

LLM outputs were generated using next-token generation at a temperature of 0.0. For classification/extraction tasks, we terminated at a new-line token. For `rule_qa` and all application tasks except `diversity_jurisdiction_6` we generated 150 tokens. For `diversity_jurisdiction_6` we generated 300 tokens.

We believe there is significant scope for improving and refining prompts on LegalBench. Hence, our results here provide a lower-bound on performance, as better prompts may elicit higher scores. Our prompts correspond to what we believe would be reasonable, based on experience with prompt engineering in other settings, and the guidance provided by model developers. We make all prompts available as a starting point for future work on LegalBench.

5.1.3 Evaluation

Classification tasks are evaluated using "exact-match" (following HELM [80]). Because some tasks contain significant label imbalances, we use balanced-accuracy as a metric. For extraction tasks, we perform normalization on generated outputs to account for differences in tense/casing/punctuation. A few tasks (e.g., `successor_liability` and `ssla_individual_defendants`) requires the LLM to produce multiple classes or extracted terms per instance. For these, we evaluate using F1. Appendix E provides more details.

Rule-application tasks were evaluated manually by a law-trained individual, who analyzed LLM responses for both correctness and analysis.[9] This type of manual evaluation is consistent with previous works evaluating LLM generations in the legal domain [30, 71]. As rule-application requires LLMs to generate "explanations" detailing legal reasoning—a capability primarily exhibited by larger models—we only evaluated GPT-4, GPT-3.5, and Claude-1. `rule_qa` was also manually evaluated by a law-trained individual. Appendix E provides more details on our approach to manual grading. All manual evaluation was performed with reference to a grading guide, which we additionally make available.

5.2 Performance trends

Table 2 provides the average task performance for all 20 models in five reasoning categories (issue-spotting, rule-recall, rule-conclusion, interpretation, and rhetorical-understanding). The first block of rows corresponds to large commercial models, the second block corresponds to models in the 11B-13B range, the third block corresponds to models in the 6B-7B range, and the final block corresponds to models in the 2B-3B range. Table 3 provides the average task performance for the three large models on rule-application. Appendix G provides full results for each model on each task.

Overall, we find significant variation in performance across tasks, suggesting that LegalBench captures a diverse spectrum of difficulty (Appendix G). These results emphasize that assessments of LLM capabilities for legal applications must be made on a task-by-task basis, and informed by the nuances of specific tasks. While certain types of tasks appear beyond the scope of current-day LLMs, others seem more within reach. In this section, we offer preliminary observations on performance trends across model size, family, and reasoning categories.

[7] We note that as of July 2023, we were unable to identify public law-specific English large language models to evaluate.

[8] `https://docs.anthropic.com/claude/docs/introduction-to-prompt-design`

[9] For the six diversity jurisdiction tasks, we sampled 30 instances from each task. For all other rule-application tasks, we manually evaluated the entirety of the dataset.

LLM	Issue	Rule	Conclusion	Interpretation	Rhetorical
GPT-4	82.9	59.2	89.9	75.2	79.4
GPT-3.5	60.9	46.3	78.0	72.6	66.7
Claude-1	58.1	57.7	79.5	67.4	68.9
Flan-T5-XXL	66.0	36.0	63.3	64.4	70.7
LLaMA-2-13B	50.2	37.7	59.3	50.9	54.9
OPT-13B	52.9	28.4	45.0	45.1	43.2
Vicuna-13B-16k	34.3	29.4	34.9	40.0	30.1
WizardLM-13B	24.1	38.0	62.6	50.9	59.8
BLOOM-7B	50.6	24.1	47.2	42.8	40.7
Falcon-7B-Instruct	51.3	25.0	52.9	46.3	44.2
Incite-7B-Base	50.1	36.2	47.0	46.6	40.9
Incite-7B-Instruct	54.9	35.6	52.9	54.5	45.1
LLaMA-2-7B	50.2	33.7	55.9	47.7	47.7
MPT-7B-8k-Instruct	54.3	25.9	48.9	42.1	44.3
OPT-6.7B	52.4	23.1	46.3	48.9	42.2
Vicuna-7B-16k	3.9	14.0	35.6	28.1	14.0
BLOOM-3B	47.4	20.6	45.0	45.0	36.4
Flan-T5-XL	56.8	31.7	52.1	51.4	67.4
Incite-3B-Instruct	51.1	26.9	47.4	49.6	40.2
OPT-2.7B	53.7	22.2	46.0	44.4	39.8

Table 2: Average performance for each LLM over the different LEGALBENCH categories. The first block of rows corresponds to large commercial models, the second block corresponds to models in the 11B-13B range, the third block corresponds to models in the 6B-7B range, and the final block corresponds to models in the 2B-3B range. The columns correspond to (in order): issue-spotting, rule-recall, rule-conclusion, interpretation, and rhetorical-understanding. For each class of models (large, 13B, 7B, and 3B), the best performing model in each category of reasoning is underlined.

LLM	Correctness	Analysis
GPT-4	82.2	79.7
GPT-3.5	58.5	44.2
Claude-v1	61.4	59.0

Table 3: Average performance for the large LLMs on rule-application tasks.

Parameter count Within LLM families, we observe that larger models usually outperform smaller models. For instance, Flan-T5-XXL (11B parameters) outperforms Flan-T5-XL (3B parameters) on average across all five reasoning categories, and LLaMA-2-13B outperforms LLaMA-2-7B on average across four reasoning categories. Notably, the margin of the gap varies across LLM families and reasoning categories. For instance, on rule-recall, the 7B Incite-Instruct model outperforms the 3B Incite-Instruct model by almost 10pts, while the 6.7B OPT model outperforms the 2.7B OPT model by less than 1pt. We additionally note that the largest LLM (GPT-4) outperforms virtually all other models.

Variation across families Even for LLMs of the same size, we find considerable differences in performance. For instance, we observe significant gaps in performance between Flan-T5-XXL (11B parameters) and Vicuna-13B-16k (13B parameters), across all reasoning categories. This suggests, unsurprisingly, that the choice of pretraining data, regime of instruction-tuning, and architecture play an important role in determining performance, and that certain configurations may be better aligned for LEGALBENCH tasks. Interestingly, we observe that such choices may affect which types of reasoning categories LLMs appear to perform well at. For instance, we observe that WizardLM-13B performs worse than all peers on issue-spotting tasks, best on rule-recall tasks, and nearly matches the performance of the best-performing peer on rule-conclusion tasks. Comparing Incite-7B-Instruct to Incite-7B-Base also provides insight into the effect of instruction-tuning across different categories, at one size point (7B parameters). We observe that instruction-tuning improves performance on four categories (issue-spotting, rule-conclusion, interpretation, and rhetorical-understanding), and worsens performance on rule-recall.

We additionally find that family-specific trends appear to hold across different size points. For instance, the Flan-T5 models outperform all others at both the 3B and 13B scale, while the Vicuna models appear to underperform competitors at both the 7B and 13B scale. We attribute the Vicuna models' low performance to their frequency tendency to generate poorly-formed outputs, which did not map to the expected verbalizer tokens (e.g., blank spaces, random characters, etc.).[10] This could possibly be attributed to the type of data used to fine the model (e.g., user-conversation), although more in-depth experimentation is necessary.

The gap between open-source and commercial models Finally, we find evidence that open-source models are capable of performance that matches or exceeds certain commercial models. For instance, Flan-T5-XXL outperforms GPT-3.5 and Claude-1 on two categories (issue-spotting and rhetorical-understanding), despite the relative gap in parameter count. Notably, the gap between closed and open-source models is largest for the rule-conclusion category. Amongst LEGALBENCH tasks, rule-conclusion tasks most like the other types of multi-step/common-sense reasoning tasks where commercial LLMs have been found to perform well.

5.3 Comparing GPT models

This section provides a more in-depth study of performance, focusing on the three commercial models (GPT-4, GPT-3.5, and Claude-1). The purpose of this section is to illustrate how LEGALBENCH enables fine-grained analysis of LLM performance. In particular we highlight how LEGALBENCH can provide more rigorous empirical support for anecdotal observations arising out of the legal community's use of these models, and explain performance differences between models.

5.3.1 Issue-spotting

We first consider average model performance across all issue-spotting tasks. We observe that GPT-4 outperforms GPT-3.5 and Claude-1 (both at $p < 0.001$).[11] In absolute terms, issue tasks present the largest gap in performance between GPT-4 and other closed-API models, with an absolute margin of 20+ points. GPT-3.5 and Claude-1, in contrast, appear to match each other in performance, separated by an average gap of only 2 points. We additionally find that the open-source models perform poorly here. On 9 tasks, Incite-Base collapses to predicting a single class for all samples.

We note one limitation to our results: because 16/17 of our issue-spotting tasks are drawn from one source (Learned Hands data), average issue performance is skewed by properties of the Learned Hands data distribution (i.e.,, user-generated questions). For instance, though GPT-3.5 outperforms Claude-1 on 12/16 Learned Hands tasks, Claude-1 outperforms GPT-3.5 on the one non-Learned Hands task (`corporate_lobbying`). Despite the skew, we observe that these tasks appear to vary in difficulty. While GPT-4's balanced-accuracy on `learned_hands_torts` is only 70.6%, on three tasks—`learned_hands_immigration`, `learned_hands_traffic`, and `learned_hands_estate`—it scores $> 95\%$.

5.3.2 Rule-recall

We first consider average model performance across all rule-recall tasks. While GPT-4 outperforms GPT-3.5 ($p < 0.05$), we surprisingly find that Claude-1 also outperforms GPT-3.5 ($p < 0.05$), and appears almost on par with GPT-4. Moreover, Claude-1 outperforms GPT-4 on three tasks: `rule_qa`, `international_citizenship_questions`, and `nys_judicial_ethics`. This is the only task category where Claude-1 provides performance comparable to GPT-4. Because little is known regarding the architecture and training processes for these models however, it is difficult to explain why this is the case.

Because rules/laws can be analogized to law-specific "facts," rule-recall tasks are similar to general domain LLM tasks designed to measure "hallucination." There, an extensive literature has documented the propensity for LLMs to both generate factually incorrect information, and answer fact-based questions incorrectly [80, 106]. Our results align with the primary findings of that literature. For example, we observe that the small open source models perform considerably worse than the larger models, consistent with the observation that model size plays an important role in fact-retention. Overall, performance on the rule-recall tasks also lend additional empirical support to more anecdotal reports—from the legal community—regarding how LLMs often misstate the law or cases [114].

5.3.3 Rule-application

Application tasks evaluate whether LLMs can explain how a legal rule applies to a set of facts, and verbalize the necessary inferences. With respect to correctness, we observe that GPT-4 outperforms both GPT-3.5 ($p < 0.01$) and Claude-1 ($p < 0.01$). Across LLMs, we find that variation in performance across tasks is consistent with subjective impressions of task difficulty. For instance, performance on `diversity_jurisdiction_1` (an easy task requiring a model to determine if an amount is greater than $75k and if the plaintiff and defendant are from different states) is much higher than performance on `successor_liability` (a harder task requiring a model to identify multiple successor liability exceptions in a fact pattern describing a complex transaction).

We observe that LLM generations may be incorrect in many different ways. On the Diversity Tasks, LLMs sometimes perform incorrect arithmetic operations or mathematical comparisons (i.e., stating that $50,000 is greater than $75,000). On `telemarketing_sales_rule` in contrast, LLMs will cite to an incorrect portion of the rule. For instance, a generation may explain that certain conduct by a telemarketer runs afoul of the rule because the telemarketer failed to make a mandatory disclosure (16 CFR § 310.3(a)(1)), but cite to the portion of the rule prohibiting misrepresentations (16 CFR § 310.3(a)(2)). Examples of other types of incorrect generations can be found in Table 4.

[10] In further experimentation, we found that writing prompts using the "### Human:" and "'### Assistant:"' templates did not appear to help.

[11] Statistical significance is computed using a paired t-test over the tasks in the category.

Fact Pattern (Task)	**LLM generation**	**Why is it incorrect?**
Sophia is from Nebraska. Ava is from Washington. Sophia sues Ava for legal malpractice for $34,000. (`diversity_1`)	"No. The parties are not diverse because Sophia and Ava are both from different states. The AiC is too small because it is less than 75k. Because the AiC requirement is not met there is no jurisdiction." **(GPT-3.5)**	While GPT-3.5 correctly identifies that there is no jurisdiction because the AiC is less than $75k (and thus too small), it fails to recognize that the parties (Sophia and Ava) are diverse because they are from different states.
The mark "Trim" for nail clippers. (`abercrombie`)	"The mark trim for nail clippers would be considered a suggestive mark. It suggests a characteristic or quality of the product which is that the nail clippers are used to trim ones nails. It implies this purpose and function without directly describing it so trim for nail clippers would be classified as a suggestive trademark." **(Claude-1)**	Claude-1's assertion that the mark "implies this purpose and function without directly describing it" misses the fact that "Trim" directly describes, rather than implies, what nail clippers are used for (i.e., the *trim* nails). Thus, the mark is actually descriptive, and not suggestive.
Ana is a lawyer who resides in Texas. While visiting Louisiana, she meets David, who runs a bike shop. She buys a bike from him and uses it to bike back to Texas. Right after she crosses the border, the bike seat explodes, injuring Ana. Ana sues David in Louisiana. (`personal_jurisdiction`)	"Yes. There is personal jurisdiction over David in Louisiana. Explanation. 1. David is domiciled in Louisiana so there is general jurisdiction over him in Louisiana. 2. David has sufficient minimum contacts with Louisiana such that he has availed himself of the privileges of conducting business in Louisiana. He operates a bike shop in Louisiana so he purposefully avails himself of doing business in the state.. 3. Ana's claim arises out of Davids contacts with Louisiana she bought the defective bike from David's bike shop in Louisiana so the claim is directly related to David's business contacts in the state. 4. Requiring David to litigate in Louisiana would not offend traditional notions of fair play." **(Claude-1)**	The LLM states that David is domiciled in Louisiana, when the facts make no mention of this.

Table 4: Examples of generations for rule-application tasks which are *incorrect*. The last column provides analysis for why the generation is incorrect.

With respect to analysis, we observe that GPT-4 again outperforms both GPT-3.5 ($p < 0.01$) and Claude-1 ($p < 0.01$). Explanations which failed to exhibit analysis can be grouped into several categories. First, some generations will contain just a prediction as to the legal outcome, without an explanation (even when the LLM has been prompted to generate one). The same prompt–applied to other samples in the dataset–will elicit explanations containing analysis. Second, we observe a tendency for LLMs to sometimes generate explanations which merely restate the facts and legal rule, without actually offering an explanation for how the outcome is reached. Examples of such instances are provided in the table below.

5.3.4 Rule-conclusion

Rule-conclusion evaluates on the same tasks as rule-application, but only requires the LLM to generate a prediction as to the outcome, and not an explanation. We observe that GPT-4 once again outperforms GPT-3.5 ($p < 0.01$) and Claude-1 ($p < 0.01$). Claude-1 and GPT-3.5 appear approximately level on performance.

The rule-conclusion tasks offer a heuristic for characterizing the types of legal inferences LLMs are capable of or struggle with. In particular, several of these tasks organize samples into *slices*, where the samples contained within a slice all represent a similar type of fact pattern, and thus interact with the legal rule in a comparable way. For instance, the `hearsay` task contains a slice corresponding to "non-verbal hearsay." This slice contains fact patterns where an individual communicates something non-verbally (e.g., pointing), thus qualifying their conduct as a "statement" under the hearsay rule. In order to make accurate predictions on this slice, an LLM must recognize that (1) the hearsay rule applies to non-verbal communicative conduct, and (2) the non-verbal conduct in these fact patterns is communicative.

Though slices are small—and thus not intended for rigorous statistical analysis—they provide some intuition as to the source of GPT-4's improvement over GPT-3.5, and the overall areas of strength and weakness for both models. On the `hearsay` task for instance (Table 6), the difference between GPT-4 and GPT-3.5 appears primarily attributable to improvements over the slices corresponding to non-verbal hearsay and statements made in court. In looking across slices moreover, it's clear that some are comfortably within the realm of model capabilities (e.g., non-assertive conduct), while others (e.g., not introduced to prove the truth of the matter asserted) still pose a considerable challenge.

Fact Pattern (task)	**LLM generation**	**What analysis is missing?**
Martha makes special pinwheel hats in her home in Salem, Oregon. She only markets her hats on the west coast (California, Oregon, and Washington). Alex, a citizen of Maine, purchases one of the hats while visiting Salem. When Alex returns to Maine, the pinwheel comes off and seriously cuts Alex. Alex sues Martha for damages in Washington. (`personal_jurisdiction`)	“No. There is not personal jurisdiction.” **(Claude-1)**	Claude-1 does not provide any explanation for *why* there is no personal jurisdiction, so the explanation does not contain analysis.
On the issue of whether Henry negligently hit Kelly with his car, Kelly's statement, "a witness told me after the crash that Henry was on his cell phone while driving. (`hearsay`)	“Yes, there is hearsay in this situation. The reasoning is as follows:. . 1. The statement in question is an out-of-court statement, as it was made by a witness who is not testifying in court.. 2. The statement is being introduced to prove the truth of the matter asserted, which is that Henry was on his cell phone while driving.. 3. Since the statement is an out-of-court statement being introduced to prove the truth of the matter asserted, it is considered hearsay.” **(GPT-4)**	GPT-4 does not explain how the truthfulness of the out-of-court statement (whether or not Henry was on his phone) relates to the purpose of its introduction (proving/disproving that Henry negligently drove).

Table 5: Examples of generations for rule-application tasks which do not contain analysis. The last column explains why the generation is deficient.

Slice	**Slice description**	**GPT-3.5**	**GPT-4**
Non-assertive conduct ($n = 19$)	The fact pattern describes conduct which is non-communicative and therefore not hearsay.	100%	100%
Statement made in-court ($n = 14$)	The fact pattern describes a statement that was made in court and therefore not hearsay.	57%	93%
Standard hearsay ($n = 29$)	The fact pattern describes traditional hearsay (out-of-court statement introduced to prove the truth of the matter asserted).	97%	97%
Non-verbal hearsay ($n = 12$)	The fact pattern describes non-verbal communicative conduct that qualifies as hearsay.	33%	75%
Not introduced to prove truth ($n = 20$)	The fact pattern describes a statement *not* introduced to prove the truth of the matter asserted, which is therefore not hearsay.	25%	45%

Table 6: Comparison between GPT-3.5 and GPT-4 on `hearsay` slices. Accuracy is reported for each slice.

Another example is provided by the `abercrombie` task, in which an LLM must determine the relationship between a product and a potential trademark name, by classifying the product-name pair into one of five categories recognized by courts: generic, descriptive, suggestive, arbitrary, and fanciful. Loosely, these categories measure how *distinctive* a product name is for a product, with generic being the least distinctive, and fanciful being the most distinctive. Just as with `hearsay`, comparing LLM performance on each of these categories provides insight into the relative areas of improvement (Table 7). Here, GPT-4's improved overall performance appears most attributable to performance on marks which are suggestive or arbitrary. However, GPT-4 still makes a number of errors for both categories. Interestingly, performance on descriptive marks is consistent between both models.

5.3.5 Interpretation

On the interpretation tasks, we find that on average GPT-4 outperforms GPT-3.5 ($p < 0.01$), ans GPT-3.5 outperforms Claude-1 ($p < 0.01$). Here, the larger API-models are highly performant on tasks which involve binary classification over short clauses. Averaged across the 38 CUAD tasks (contract clauses), for instance, GPT-4, GPT-3.5, and Claude-1 all have a balanced-accuracy $\geq 88\%$. And on `proa` (statutory clauses), both GPT-4 and GPT-3.5 have a balanced-accuracy $\geq 90\%$. Notably, performance degrades on tasks which contain longer text sequences or involve multi-class classification. On the Supply Chain Disclosure tasks for instance—in which LLMs must classify disclosures which are 1-2 pages in length—the average balanced-accuracy of the large commercial models ranges between 74-75%. And on the MAUD tasks—which require answering multiple choice questions about merger deals—the average balanced-accuracy of GPT-4 drops to 47.8% accuracy.

Mark	**Mark description**	**GPT-3.5**	**GPT-4**
Generic ($n = 19$)	The name connotes the basic nature of the product/service.	94%	100%
Descriptive ($n = 19$)	The name identifies a characteristic or quality of the product/service.	73%	72%
Suggestive ($n = 20$)	The name suggests, rather than describes, a characteristic of the product/service.	38%	70%
Arbitrary ($n = 18$)	The name is a real world but has no relation to the product/service.	41%	82%
Fanciful ($n = 19$)	The name is a made-up word.	84%	100%

Table 7: Comparison between GPT-3.5 and GPT-4 on `abercrombie` categories. Accuracy is reported for each slice.

5.3.6 Rhetorical-analysis

On average across all rhetorical-understanding tasks, we find that GPT-4 outperforms both GPT-3.5 ($p \leq 0.05$) and Claude-1 ($p \leq 0.05$). We note several results. First, on `definition_extraction`—which requires a LLM to extract the term defined by a sentence taken from a Supreme Court opinion—Incite-Base almost equals GPT-4 in performance (80.6% accuracy to 81.8%). Second, nearly all evaluated models struggle on two tasks requiring LLMs to label the legal "roles" played by either a question or excerpt from an opinion (`function_of_decision_section` and `oral_argument_question_purpose`). Notably, both tasks require the LLM to classify text into one of six or more categories

5.4 Prompt engineering strategies

Finally, we illustrate—through a series of micro-studies—how LegalBench can be used to explore different aspects of prompt-engineering for LLMs in legal settings. We focus on three questions:

1. Can LLMs rely on their latent knowledge of a rule for rule-conclusion tasks?
2. Does simplifying task descriptions to plain language affect performance?
3. Are LLMs sensitive to the choice of in-context demonstrations?

Reliance on latent knowledge When prompting for general-domain tasks like sentiment or topic classification, prompt-engineers will often rely on the LLM's latent knowledge of the task [5]. In topic classification for instance, a prompt may use the instructions to label whether a news article is about "sports," without offering a detailed description of what "sports" refers to or encompasses. Such a description is not necessary, because general-domain terms like "sports" appear frequently in LLM training corpora, and LLMs can learn from these occurrences what general-domain terms mean. Prompting for legal tasks, however, may require a different strategy. Because legal terms occur less frequently in general domain training corpora, legal prompting may require practitioners to provide additional background information. For example, a general domain LLM may not know what the requirements for diversity jurisdiction are, because diversity jurisdiction is not as commonly discussed in pretraining corpora.

We explore this question through a study of rule-conclusion tasks. For a selection of these tasks, we evaluate GPT-3.5 with two zero-shot prompts: a reference-based prompt and a description-based prompt. In the reference prompt, the task instructions merely state the rule to be applied, i.e., "Determine if the following fact patterns give rise to diversity jurisdiction." In the description-based prompt, the instructions provide an explicit description of the rule, i.e., "Diversity jurisdiction exists when there is (1) complete diversity between plaintiffs and defendants, and (2) the amount-in-controversy (AiC) is greater than $75k." By comparing performance between the reference and description prompt, we can measure whether providing a description of the rule in the prompt provides additional performance boost over the LLM's latent knowledge of the rule.

Figure 1 provides a comparison for the different prompts. Interestingly, we find considerable variation across tasks. On tasks like `abercrombie`, `ucc_v_common_law`, `diversity_2`, and `diversity_4`, description prompts appear to offer significant increase in performance. On the other tasks, performance is approximately the same (or even worse). We identify two possible explanations for diverging results across tasks. First, on certain tasks, subsets of fact-patterns are too challenging for LLMs like GPT-3.5, and description-based prompts do not provide sufficient guidance for LLMs to reason through those fact patterns. Second, legal rules may be described to varying extents within pretraining corpora. Hence, tasks where we observe performance improvements from description-based prompting may correspond to rules which occur less frequently in pretraining data.

Plain language descriptions of tasks Next, we examine the extent to which domain specialization in the language of the prompts affects performance. Like experts in other specialized domains, lawyers have developed their own language (i.e., "legalese"), which forms the basis for most legal writing and communication. It is unclear whether—in interacting with large language models through prompting—lawyers should continue to rely on formalistic legal language, or instead use simpler plain language. While most large language models are "general domain" and thus less specialized to legalese, formalistic legal language is more precise, and may thus induce more accurate behavior from the model.

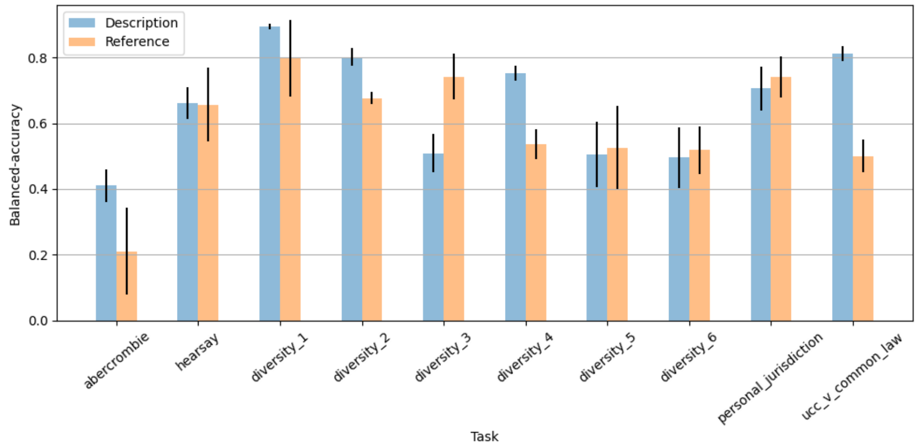


Figure 1: We compare performance of prompts which describe the legal rule to be applied ("description") against prompts which reference the legal rule to be applied ("reference"). Error bars measure standard error, computed using a bootstrap with 1000 resamples.

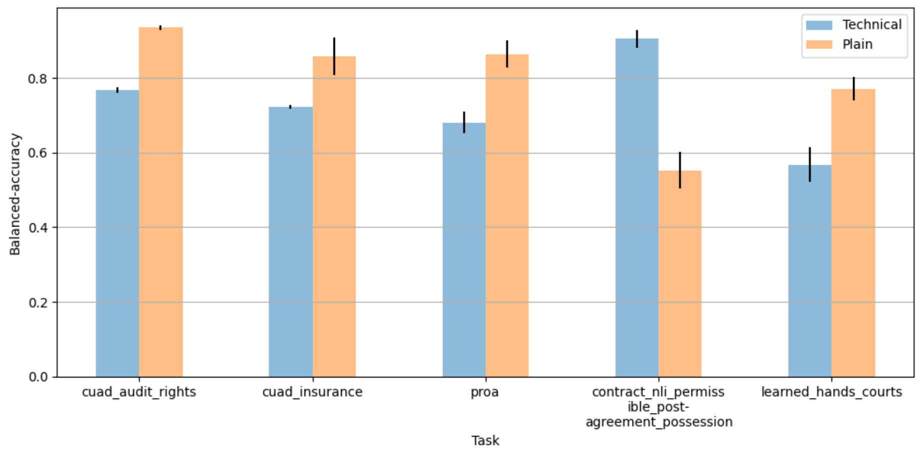


Figure 2: We compare performance of prompts which describe the task in plain language to prompts which describe the task in technical legal language (for GPT-3.5). Error bars measure standard error, computed using a bootstrap with 1000 resamples.

We explore this question by comparing "plain language" and "technical language" prompts. For a subset of LEGALBENCH tasks, we have access to the formal language provided to law-trained annotators when creating task data. By comparing the performance of a prompt which uses this language—to one which uses a plain-language version—we can measure how the technicality of language affects results.

We conduct preliminary experiments on a sample of five LEGALBENCH tasks (Figure 2).[12] On four of the five tasks, we find that the plain-language prompt significantly outperforms the technical language prompt, by up to 21 points (balanced-accuracy). Interestingly, on `contract_nli_permissible_post-agreement_possession`, we find the opposite phenomenon holds: the plain language prompt is substantially *worse* than the technical prompt.

Sensitivity to in-context demonstrations Finally, we investigate the influence of the in-context demonstrations used in prompts. Prior work in general domain LLMs have observed that few-shot performance is highly sensitive to the choice of demonstrations [58, 125, 143]. We evaluate whether LLMs are similarly sensitive for legal tasks, focusing on a subset of 8 binary classification tasks. For each task we merge the train and evaluation split into a single dataset, and randomly sample four in-context samples to include in the prompt (two from each class), five different times. We evaluated GPT-3.5 and Incite-Instruct-7B with each of the five generated prompts, and plot the the balanced-accuracy of each prompt in Figure 3.

[12]Prompts are made available in the LEGALBENCH repository.

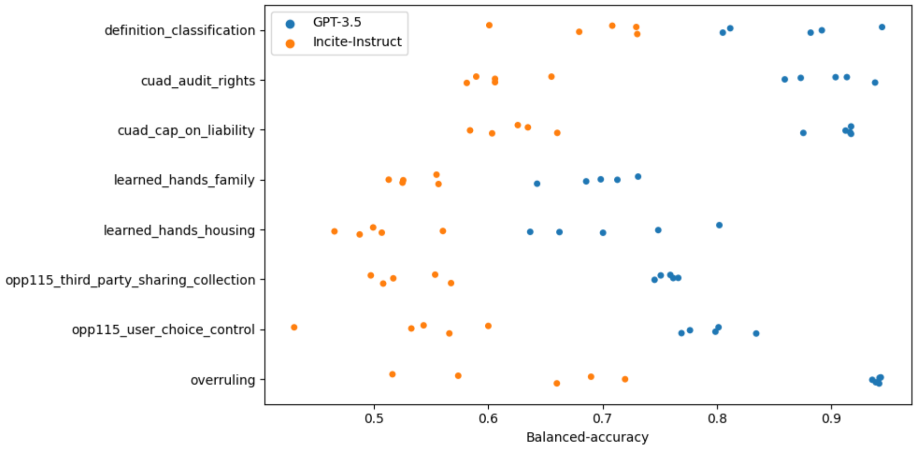


Figure 3: We evaluate GPT-3.5 and Incite-Instruct on five prompts constructed by randomly selecting different samples to use as in-context demonstrations (maintaining class balance in the prompt). In the figure above, each point corresponds to a different prompt.

Consistent with findings on general-domain tasks, we observe that LLMs on legal tasks are also highly sensitive to the choice of in-context samples. Notably, this appears to be the case for both GPT-3.5 and Incite-Instruct. Under a permutation test, we find significant differences ($p < 0.01$) between the best and worst performing prompt for Incite-Instruct (on all tasks), and for GPT-3.5 (on all tasks except `opp115_third_party_sharing_collection` and `overruling`).[13] For many tasks, the magnitude of difference is substantial. On `overruling` for instance, the best Incite-Instruct prompt improves upon the worst prompt by over 20 points (balanced-accuracy). Overall, these results suggest that future work is needed to understand how different demonstrations influence performance.

6 Conclusion

Our work here describes LEGALBENCH: a collaboratively constructed benchmark of 162 tasks for measuring the legal reasoning capabilities of LLMs. In future work, we hope to expand this project, by continuing to solicit and collect interesting and useful tasks from the legal community.

References

[1] Neel Alex, Eli Lifland, Lewis Tunstall, Abhishek Thakur, Pegah Maham, C Jess Riedel, Emmie Hine, Carolyn Ashurst, Paul Sedille, Alexis Carlier, et al. Raft: A real-world few-shot text classification benchmark. *arXiv preprint arXiv:2109.14076*, 2021.

[2] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. Falcon-40B: an open large language model with state-of-the-art performance. 2023.

[3] Anthropic. Introducing claude. `https://www.anthropic.com/index/introducing-claude`, 2023.

[4] Yonathan A Arbel and Samuel Becher. How smart are smart readers? llms and the future of the no-reading problem. *LLMs and the Future of the No-Reading Problem (June 25, 2023)*, 2023.

[5] Simran Arora, Avanika Narayan, Mayee F Chen, Laurel J Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*, 2022.

[6] Kevin D Ashley. *Artificial intelligence and legal analytics: new tools for law practice in the digital age*. Cambridge University Press, 2017.

[7] Ian Ayres and Alan Schwartz. The no-reading problem in consumer contract law. *Stan. L. Rev.*, 66:545, 2014.

[8] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. Does anyone read the fine print? consumer attention to standard-form contracts. *The Journal of Legal Studies*, 43(1):1–35, 2014.

[9] Edward Beeching, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.

[13] We conduct the permutation test with 1000 resamples.

[10] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[11] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Shelter check: Proactively finding tax minimization strategies via ai. *Tax Notes Federal, Dec*, 12, 2022.

[12] Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. Can gpt-3 perform statutory reasoning? *arXiv preprint arXiv:2302.06100*, 2023.

[13] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[15] Bryan, Cave, Leighton, and Paisner. 2023 state-by-state ai legislation snapshot. https://www.bclplaw.com/en-US/events-insights-news/2023-state-by-state-artificial-intelligence-legislation-snapshot.html, 2023.

[16] Ilias Chalkidis. Chatgpt may pass the bar exam soon, but has a long way to go for the lexglue benchmark. *arXiv preprint arXiv:2304.12202*, 2023.

[17] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural legal judgment prediction in english. *arXiv preprint arXiv:1906.02059*, 2019.

[18] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multieurlex–a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2109.00904*, 2021.

[19] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*, 2020.

[20] Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. Lexfiles and legallama: Facilitating english multinational legal language model development. *arXiv preprint arXiv:2305.07507*, 2023.

[21] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330. Association for Computational Linguistics, 2022.

[22] Ilias Chalkidis and Dimitrios Kampas. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, 27(2):171–198, 2019.

[23] Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Felix Schwemer, and Anders Søgaard. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. *arXiv preprint arXiv:2203.07228*, 2022.

[24] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[25] Edward K Cheng, Ehud Guttel, and Yuval Procaccia. Unenforceable waivers. *Vanderbilt Law Review, Forthcoming (2023)*, 2022.

[26] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[27] Adam S Chilton and Galit A Sarfaty. The limitations of supply chain disclosure regimes. *Stan. J. Int'l L.*, 53:1, 2017.

[28] Travers Ching, Daniel S Himmelstein, Brett K Beaulieu-Jones, Alexandr A Kalinin, Brian T Do, Gregory P Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M Hoffman, et al. Opportunities and Obstacles for Deep Learning in Biology and Medicine. *Journal of The Royal Society Interface*, 15(141):20170387, 2018.

[29] Jonathan H. Choi. How to use large language models for empirical legal research. *Journal of Institutional and Theoretical Economics*, 2023.

[30] Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Available at SSRN*, 2023.

[31] Jonathan H. Choi and Daniel Schwarcz. Ai assistance in legal analysis: An empirical study. *Available at SSRN 4539836*, 2023.

[32] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

[33] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

[34] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[35] Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, April 2023.

[36] Legal Services Corporation. The Justice Gap: Measuring the Unmet Civil Legal Needs of Low-Income Americans, 2017.

[37] Legal Services Corporation. Eviction laws database: Local dataset. https://www.lsc.gov/initiatives/effect-state-local-laws-evictions/lsc-eviction-laws-database, 2021.

[38] Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *arXiv preprint arXiv:2204.04859*, 2022.

[39] Faraz Dadgostari, Mauricio Guim, Peter A Beling, Michael A Livermore, and Daniel N Rockmore. Modeling law search as prediction. *Artificial Intelligence and Law*, 29:3–34, 2021.

[40] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*, pages 177–190. Springer, 2006.

[41] Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. Revisiting transformer-based models for long document classification. *arXiv preprint arXiv:2204.06683*, 2022.

[42] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

[43] Tri Dao, Daniel Y Fu, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

[44] Yasmin Dawood. Campaign finance and american democracy. *Annual Review of Political Science*, 18:329–348, 2015.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[46] Gregory M Dickinson. A computational analysis of oral argument in the supreme court. *Cornell JL & Pub. Pol'y*, 28:449, 2018.

[47] Phoebe C Ellsworth. Legal reasoning. In K. J. Holyoak and R. G. Morrison Jr., editors, *The Cambridge Handbook of Thinking and Reasoning*, pages 685–704. Cambridge University Press, New York, 2005.

[48] David Freeman Engstrom and Jonah B Gelbach. Legal Tech, Civil Procedure, and the Future of Adversarialism. *University of Pennsylvania Law Review*, 169:1001, 2020.

[49] Epiq. Pandemics and force majeure: How can ai help you? https://www.jdsupra.com/legalnews/pandemics-and-force-majeure-how-can-ai-90757/, 2020.

[50] Frank Fagan. From policy confusion to doctrinal clarity: successor liability from the perspective of big data. *Va. L. & Bus. Rev.*, 9:391, 2014.

[51] Sean Farhang. The litigation state. In *The Litigation State*. Princeton University Press, 2010.

[52] Yi Feng, Chuanyi Li, and Vincent Ng. Legal judgment prediction: A survey of the state of the art. *IJCAI. ijcai. org*, pages 5461–9, 2022.

[53] Jens Frankenreiter and Julian Nyarko. Natural language processing in legal tech. *Legal Tech and the Future of Civil Justice (David Engstrom ed.)*, 2022.

[54] Daniel Y Fu, Elliot L Epstein, Eric Nguyen, Armin W Thomas, Michael Zhang, Tri Dao, Atri Rudra, and Christopher Ré. Simple hardware-efficient long convolutions for sequence modeling. *arXiv preprint arXiv:2302.06646*, 2023.

[55] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.

[56] Saibo Geng, Rémi Lebret, and Karl Aberer. Legal transformer models may not always help. *arXiv preprint arXiv:2109.06862*, 2021.

[57] Kurt Glaze, Daniel E Ho, Gerald K Ray, and Christine Tsang. Artificial Intelligence for Adjudication: The Social Security Administration and AI Governance. In Justin Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew Young , and Baobao Zhang, editors, *The Oxford Handbook of AI Governance*, pages 685–704. Oxford University Press, 2022.

[58] Neel Guha, Mayee F Chen, Kush Bhatia, Azalia Mirhoseini, Frederic Sala, and Christopher Ré. Embroid: Unsupervised prediction smoothing can improve few-shot classification. *arXiv preprint arXiv:2307.11031*, 2023.

[59] Neel Guha, Daniel E Ho, Julian Nyarko, and Christopher Ré. Legalbench: Prototyping a collaborative benchmark for legal reasoning. *arXiv preprint arXiv:2209.06120*, 2022.

[60] Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset, 2022.

[61] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[62] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[63] David Hoffman and Yonathan Arbel. Generative interpretation. *Available at SSRN 4526219*, 2023.

[64] David A Hoffman. Defeating the empire of forms. *Available at SSRN 4334425*, 2023.

[65] Nils Holzenberger, Andrew Blair-Stanek, and Benjamin Van Durme. A dataset for statutory reasoning in tax law entailment and question answering. *arXiv preprint arXiv:2005.05257*, 2020.

[66] Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. A multi-task benchmark for korean legal language understanding and judgement prediction. *Advances in Neural Information Processing Systems*, 35:32537–32551, 2022.

[67] Cong Jiang and Xiaolei Yang. Legal syllogism prompting: Teaching large language models for legal judgment prediction. *arXiv preprint arXiv:2307.08321*, 2023.

[68] Abhinav Joshi, Akshat Sharma, Sai Kiran Tanikella, and Ashutosh Modi. U-creat: Unsupervised case retrieval using events extraction. *arXiv preprint arXiv:2307.05260*, 2023.

[69] Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula. Text summarization from legal documents: a survey. *Artificial Intelligence Review*, 51:371–402, 2019.

[70] Arnav Kapoor, Mudit Dhawan, Anmol Goel, TH Arjun, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. Hldc: Hindi legal documents corpus. *arXiv preprint arXiv:2204.00806*, 2022.

[71] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Available at SSRN 4389233*, 2023.

[72] Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*, 2023.

[73] Noam Kolt. Predicting consumer contracts. *Berkeley Technology Law Journal*, 37, 2022.

[74] Yuta Koreeda and Christopher D Manning. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*, 2021.

[75] Aditya Kuppa, Nikon Rasumov-Rahe, and Marc Voses. Chain of reference prompting helps llm to think like a lawyer. *Generative AI + Law Workshop*, 2023.

[76] Kwok-Yan Lam, Victor CW Cheng, and Zee Kin Yeong. Applying large language models for enhancing contract drafting. *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workspace (LegalAIIA 2023)*, 2023.

[77] Grant Lamond. Precedent and Analogy in Legal Reasoning. *Stanford Encyclopedia of Philosophy*, 2006.

[78] Sarah B Lawsky. A logic for statutes. *Fla. Tax Rev.*, 21:60, 2017.

[79] Zehua Li, Neel Guha, and Julian Nyarko. Don't use a cannon to kill a fly: An efficient cascading pipeline for long documents. *International Conference on AI and Law*, 2023.

[80] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[81] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.

[82] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Giovanni Sartor, and Paolo Torroni. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139, 2019.

[83] William V Luneburg and Thomas M Susman. The lobbying manual: a complete guide to federal lobbying law and practice. American Bar Association, 2009.

[84] Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. Learning to predict charges for criminal cases with legal basis. *arXiv preprint arXiv:1707.09168*, 2017.

[85] Vijit Malik, Rishabh Sanjay, Shouvik Kumar Guha, Angshuman Hazarika, Shubham Nigam, Arnab Bhattacharya, and Ashutosh Modi. Semantic segmentation of legal documents via rhetorical roles. *arXiv preprint arXiv:2112.01836*, 2021.

[86] Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripa Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. *arXiv preprint arXiv:2105.13562*, 2021.

[87] Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. *arXiv preprint arXiv:2211.00974*, 2022.

[88] Stelios Maroudas, Sotiris Legkas, Prodromos Malakasiotis, and Ilias Chalkidis. Legal-tech open diaries: Lesson learned on how to develop and deploy light-weight models in the era of humongous language models. *arXiv preprint arXiv:2210.13086*, 2022.

[89] Masha Medvedeva, Martijn Wieling, and Michel Vols. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212, 2023.

[90] Kaiz Merchant and Yash Pande. Nlp based latent semantic analysis for legal text summarization. In *2018 international conference on advances in computing, communications and informatics (ICACCI)*, pages 1803–1807. IEEE, 2018.

[91] Guilherme Moraes Rosa, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Roberto Lotufo, and Rodrigo Nogueira. Billions of parameters are worth more than in-domain training data: A case study in the legal case entailment task. *arXiv e-prints*, pages arXiv–2205, 2022.

[92] John J Nay. Predicting and understanding law-making with word vectors and an ensemble model. *PloS one*, 12(5):e0176999, 2017.

[93] John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. Large language models as tax attorneys: A case study in legal capabilities emergence, 2023.

[94] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*, 2021.

[95] Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*, 2023.

[96] Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*, 2023.

[97] Jonathan A Obar and Anne Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, 23(1):128–147, 2020.

[98] OpenAI. Gpt-4 technical report, 2023.

[99] Laurel Orr. Manifest. `https://github.com/HazyResearch/manifest`, 2022.

[100] Laurel Orr, Megan Leszczynski, Simran Arora, Sen Wu, Neel Guha, Xiao Ling, and Christopher Re. Bootleg: Chasing the tail with self-supervised named entity disambiguation. *arXiv preprint arXiv:2010.10363*, 2020.

[101] Anja Oskamp and Marc Lauritsen. Ai in law practice? so far, not much. *AI & L.*, 10:227, 2002.

[102] Adam Pah, David Schwartz, Sarath Sanga, Charlotte Alexander, Kristian Hammond, Luis Amaral, SCALES OKN Consortium, et al. The promise of ai in an open justice system. *AI Magazine*, 43(1):69–74, 2022.

[103] Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. Multi-granular legal topic classification on greek legislation. *arXiv preprint arXiv:2109.15298*, 2021.

[104] Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11139–11146, 2022.

[105] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

[106] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

[107] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070, 2021.

[108] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

[109] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*, 2023.

[110] Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. Overview and discussion of the competition on legal information extraction/entailment (coliee) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133, 2022.

[111] Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E Ho, and Joel Niklaus. Scale: Scaling up the complexity for advanced language model evaluation. *arXiv preprint arXiv:2306.09237*, 2023.

[112] Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*, 2019.

[113] Danilo Ribeiro, Shen Wang, Xiaofei Ma, Henry Zhu, Rui Dong, Deguang Kong, Juliette Burger, Anjelica Ramos, William Wang, Zhiheng Huang, et al. Street: A multi-task structured reasoning and explanation benchmark. *arXiv preprint arXiv:2302.06729*, 2023.

[114] James Romoser. No, ruth bader ginsburg did not dissent in obergefell — and other things chatgpt gets wrong about the supreme court. *SCOTUSblog*, 2023.

[115] Erik F Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*, 2003.

[116] Jaromir Savelka. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. *arXiv preprint arXiv:2305.04417*, 2023.

[117] Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv preprint arXiv:2306.13906*, 2023.

[118] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[119] Robert E Scott, Stephen J Choi, and Mitu Gulati. Contractual landmines. *Available at SSRN*, 2022.

[120] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-lexsum: Real-world summaries of civil rights lawsuits at multiple granularities. *arXiv preprint arXiv:2206.10883*, 2022.

[121] Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, and Saptarshi Ghosh. Legal case document summarization: Extractive and abstractive methods and their evaluation. *arXiv preprint arXiv:2210.07544*, 2022.

[122] Cecilia Silver. Breaking news: Drafting client alerts to prepare for practice. *Perspectives: Teaching Legal Research and Writing*, 27, 2019.

[123] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv preprint arXiv:2206.04615*, 2022.

[124] Norman Otto Stockmeyer. Legal Reasoning? It's All About IRAC, Mar 2021.

[125] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*, 2022.

[126] Harry Surden. Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35:19–22, 2019.

[127] Harry Surden. The ethics of artificial intelligence in law: Basic questions. *Forthcoming chapter in Oxford Handbook of Ethics of AI*, pages 19–29, 2020.

[128] Harry Surden. Values embedded in legal artificial intelligence. *IEEE Technology and Society Magazine*, 41(1):66–74, 2022.

[129] Mirac Suzgun, Luke Melas-Kyriazi, Suproteem K Sarkar, Scott Duke Kominers, and Stuart M Shieber. The harvard uspto patent dataset: A large-scale, well-structured, and multi-purpose corpus of patent applications. *arXiv preprint arXiv:2207.04043*, 2022.

[130] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[131] MosaicML NLP Team. Introducing mpt-30b: Raising the bar for open-source foundation models, 2023. Accessed: 2023-06-22.

[132] Wex Definitions Team. ejusdem generis. `https://www.law.cornell.edu/wex/ejusdem_generis`, 2022.

[133] Wex Definitions Team. textualism. https://www.law.cornell.edu/wex/textualism, 2022.

[134] Joel Tito. How ai can improve access to justice, 2017.

[135] Together. Releasing 3b and 7b redpajama-incite family of models including base, instruction-tuned & chat models. `https://www.together.xyz/blog/redpajama-models-v1`, 2023.

[136] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[137] Maarten Peter VINK, Luuk Van der Baaren, Rainer Bauböck, Iseult Honohan, and Bronwen Manby. Globalcit citizenship law dataset. 2021.

[138] Eugene Volokh. Chatgpt coming to court, by way of self-represented litigants. *The Volokh Conspiracy*, 2023.

[139] Brandon Waldon, Madigan Brodsky, Megan Ma, and Judith Degen. Predicting consensus in legal document interpretation. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, to appear.

[140] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

[141] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[142] Steven H. Wang, Antoine Scardigli, Leonard Tang, Wei Chen, Dimitry Levkin, Anya Chen, Spencer Ball, Thomas Woodside, Oliver Zhang, and Dan Hendrycks. Maud: An expert-annotated legal nlp dataset for merger agreement understanding, 2023.

[143] Xinyi Wang, Wanrong Zhu, and William Yang Wang. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*, 2023.

[144] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019.

[145] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022.

[146] Benjamin Weiser. Here's what happens when your lawyer uses chatgpt. *New York Times*, 2023.

[147] Hannes Westermann, Jaromir Savelka, and Karim Benyekhlef. Llmediator: Gpt-4 assisted online dispute resolution. *arXiv preprint arXiv:2307.16732*, 2023.

[148] Wikipedia. Irac. https://en.wikipedia.org/wiki/IRAC.

[149] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1330–1340, 2016.

[150] Lawrence Wrightsman. *Oral arguments before the Supreme Court: An empirical approach*. Oxford University Press, 2008.

[151] Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, et al. Cail2018: A large-scale legal dataset for judgment prediction. *arXiv preprint arXiv:1807.02478*, 2018.

[152] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions, 2023.

[153] Fangyi Yu, Lee Quartey, and Frank Schilder. Legal prompting: Teaching a language model to think like a lawyer. *arXiv preprint arXiv:2212.01326*, 2022.

[154] Fangyi Yu, Lee Quartey, and Frank Schilder. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, 2023.

[155] Diego Zambrano, Neel Guha, Austin Peters, and Jeffrey Xia. Private enforcement in the states. *University of Pennsylvania Law Review, forthcoming*, 2023.

[156] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[157] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 159–168, 2021.

[158] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. Legal judgment prediction via topological learning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3540–3549, 2018.

[159] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*, 2023.

[160] Lee B. Ziffer. The robots are coming: Ai large language models and the legal profession. *American Bar Association*, 2023.

[161] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel R Reidenberg, N Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. *Proc. Priv. Enhancing Tech.*, 2019:66, 2019.

A Acknowledgements

We are grateful to the following individuals and groups for feedback on this project: Alex Chao, Amit Haim, Arjun Desai, Armin Thomas, Avanika Narayan, Ben Spector, Brandon Yang, Eric Nguyen, Gautam Machiraju, Javed Qadrud-Din, Jian Zhang, Jonathan Zittrain, Karan Goel, Khaled Saab, Joshua Arp, Krista Opsahl-Ong, Laurel Orr, Lisa Ouellette, Lucia Zheng, Martin Gajek, Mayee Chen, Michael Zhang, Mike Wornow, Pablo Arredondo, Percy Liang, Rishi Bommasani, Roland Vogl, Sabri Eyuboglu, Sarah Hooper, Sergio Servantez, Simran Arora, Tengyu Ma, Tony Kim, Tri Dao and Vishnu Sarukkai. We presented and recieved feedback on earlier versions of this project at various forums, including: the Center for Research on Foundation Models, the Stanford Regulation and Governance Lab, the New York LLM x Law Hackathon (June 2023), the 2023 Stanford Data Science Conference, the 2023 Stanford CodeX Conference, and the Stanford Generative AI and Foundation Models Workshop. We are grateful to the organizers and attendees of these events for engaging with our work.

We gratefully acknowledge the support of NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), 2204926 (Computational Statutory Reasoning), and 1937301 (RTML); US DEVCOM ARL under No. W911NF-21-2-0251 (Interactive Human-AI Teaming); ONR under No. N000141712266 (Unifying Weak Supervision); ONR N00014-20-1-2480: Understanding and Applying Non-Euclidean Geometry in Machine Learning; N000142012275 (NEPTUNE); NXP, Xilinx, LETI-CEA, Intel, IBM, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, Google Cloud, Salesforce, Total, the Center for Research on Foundation Models, the HAI-GCP Cloud Credits for Research program, the Stanford Data Science Initiative (SDSI), and members of the Stanford DAWN project: Facebook, Google, and VMWare. We thank Casetext for assistance with evaluating GPT-4. PH is supported by an Open Philanthropy AI Fellowship. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of NIH, ONR, or the U.S. Government.

B Limitations and social impact

Limitations We note several limitations of our work. Legal applications—and what constitutes "legal reasoning"—is broad. Thus, LEGALBENCH will necessarily be an incomplete effort, and important tasks/document types/reasoning types are not included. To enumerate a few examples:

- LEGALBENCH does not include tasks over *long* documents. Long documents are significant for legal practice, as writings like contracts, corporate filings, statutory codes, and judicial opinions can be hundreds of pages long [79].
- The legal reasoning dimensions identified in LEGALBENCH constitute a subset of the possible legal reasoning abilities for which we wish to evaluate LLMs. An example of a reasoning ability which is not currently evaluated in LEGALBENCH would be analogical reasoning grounded in case law.
- LEGALBENCH tasks are skewed towards certain legal domains (e.g., contracts and civil procedure) and others are unrepresented.
- LEGALBENCH tasks skew towards US Federal law, and thus may not be representative for studies of other jurisdictions, or tasks involving international law.
- LEGALBENCH does not enable evaluation for multilingual, or non-English, legal tasks.
- LEGALBENCH does not evaluate more subjective legal tasks, or tasks which contain more ambiguity. These tasks are common to the legal field.

We hope to work on these limitations as part of future work. In particular, we would like to expand LEGALBENCH to include other jurisdictions and a broader cross-section of legal domains.

Nothing in LEGALBENCH should be construed as legal advice.

Social impact A potential negative social impact of our work would be if others either (1) construed our work as unequivocally endorsing automation in the legal industry, or (2) used performance on LEGALBENCH as the sole justification for AI deployments. We therefore take efforts to mitigate these impacts, noting the following.

As we state in Section 1, the purpose of our work is not to determine whether large language models are capable of replacing legal professionals, the types of legal work that should/can be automated, or the broader implications of new technology on the practice of law. Rather, our focus is on developing technical artifacts which better enable stakeholders and affected parties to answer these questions themselves. Rigorous evaluation is essential to the safe and ethical usage of AI. LEGALBENCH, as a benchmark, is intended to *improve* the ability for stakeholders to conduct evaluations. We additionally note that LEGALBENCH, as a tool for research, is not a substitute for more in-depth and context-specific evaluation efforts. The deployment of any AI application in the law must be accompanied by evaluation on in-domain data, and assessments for ethical and legal compliance.

We finally note that potential negative impact will depend significantly on the task studied and the broader social context. The consequences of mistakes in using LLMs to annotate datasets, for instance, has significantly different consequences from the cost of mistakes when LLMs are used to answer legal aid questions.

C Datasheet

Following recent work, we provide a datasheet [55] below. The datasheet below provides general answers to each of the questions, while Appendix F provides more in-depth details for each individual task. In addition, a number of LEGALBENCH tasks have been adapted from previously released datasets, and the datasheets accompanying their publication provide further details.

C.1 Motivation

For what purpose was the data set created? Was there a specific task in mind? If so, please specify the result type (e.g. unit) to be expected.

LEGALBENCH was created to evaluate LLMs on legal tasks and better understand their legal reasoning capabilities. Recent advances in language modeling techniques have led to the emergence of "large" language models, and spurred interest within the legal community. This has led to two questions:

- What technical adaptations are necessary to enable LLMs to perform legal tasks? Legal tasks often involve longer text sequences, jargon, and multi-step reasoning, making them more difficult than traditional NLP tasks.
- For which legal tasks can current LLMs be trusted to perform safely and reliably?

LEGALBENCH encompasses many different tasks. The specification for each task and the expected output can be found in the full task descriptions (Section F).

Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g. company, institution, organization)?

LEGALBENCH consists of novel datasets (which were created by the authors of this paper), and transformed/adapted datasets (which were originally released as part of prior research). In Section F we discuss the origins of each dataset.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

LEGALBENCH and its contributors have been generously funded by a range of entities that include the institutional affiliations provided for each author, governmental grants, and other sources.

Any other comments?

None.

C.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

All LEGALBENCH tasks consist of instances which are text. These include: sentences, paragraphs, and documents. Some instances are drawn from real world sources of text (e.g., actual contracts, corporate disclosures, judicial opinions, or complaints). Other instances were synthetically crafted. Section F provides details for each task.

How many instances are there in total (of each type, if appropriate)?

Section D provides details for each task.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Nearly every LEGALBENCH task corresponds to a sample of a population, or entirely synthetic data. Section F contains a more detailed description for each dataset. We highlight several broader explanations for the difficulty in acquiring complete or representative data which generalizes across tasks:

- As prior work on legal benchmarks has noted [60, 120], not all legal documents are published or reported. Hence, many are only accessible through special request, or only available in paper. The lack of easily available representative data is a noted challenge in many justice systems [60, 102].
- Acquiring legal annotations is exceedingly expensive. The CUAD project, for instance, estimated that a modestly sized dataset of 500 contracts (relative to the standards of NLP) had an estimated cost of $2 million US dollars [62]. As a result, it is often possible to only annotate a small sample of data, even when a larger population is available.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.

Instances in LEGALBENCH largely correspond to unprocessed text. Section F contains a more detailed description for each dataset.

Is there a label or target associated with each instance? If so, please provide a description.

Yes. Labels correspond to: classes, extracted entities, and open-ended generation. Section F contains a more detailed description of the labels/targets for each dataset.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

For reused/adapted datasets, we refer readers to the original data sheets which document redactions/missing data. Newly contributed tasks should not be missing information.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Not applicable.

Are there recommended data splits (e.g.,training, development/validation,testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Yes. Tasks are split into train and test splits. Train splits consist of a small random sample of the original dataset (i.e., between 2-8 instances). We select small training splits in order to capture the true few-shot setting [107], in which a practitioner only has access to a handful of labeled instances. This design choice is also reflected in the structure of the RAFT benchmark [1].

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

A significant amount of legal data is the product of scanning and OCR. Hence, this data often contains artifacts of these processes, which appear as errant or missing characters.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

LEGALBENCH is self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.

No. All LEGALBENCH data is derived from public sources or was generated by authors. There is no confidential information in our dataset.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.

No.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

LEGALBENCH data relates to people to the extent that LEGALBENCH contains tasks which contain language drawn from judicial cases involving individuals, or posts by individuals to legal forums (i.e., the Learned Hands Tasks).

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

As LEGALBENCH is drawn entirely from public datasets—which themselves may contain additional information—it is possible to identify the original documents that LEGALBENCH data was drawn from.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The Learned Hands tasks correspond to posts on public forums. In these posts individuals discuss legal questions, and sometimes disclose information that would meet the above definition of "sensitive."

Any other comments?

We note that the data distributions from which some LEGALBENCH tasks were drawn—like judicial cases or legal forums—have been used by prior work published in the NeurIPS Datasets and Benchmarks Track [60, 120]. These works offer additional information.

C.3 Collection process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Data underlying LEGALBENCH tasks were collected using different processes, and Section F contains a detailed discussion for each task.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

Please refer to Section F for background on each task.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

Please see the discussion in the Composition section above.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Section F contains a detailed discussion for each task.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Section F contains a detailed discussion for each task.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Where applicable, Section F provides information relevant to each task.

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

The dataset relates to people insofar as it draws text from documents which relate to people, or people created.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

Section F contains a detailed discussion for each task.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

No. Following other works which incorporate data from public judicial sources [60, 120], we note that judicial filings are public, and the individuals implicated in those proceedings are aware of the public nature.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Individuals whose names and circumstances appear in the original datasets did not separately consent to be a part of LEGALBENCH. Again, we note that these documents are generally public, and already accessible to a wide range of parties, through many different judicial data services.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

No.

Any other comments?

None.

C.4 Preprocessing, cleaning, labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No.

C.5 Use

Has the dataset been used for any tasks already? If so, please provide a description.

We have used the constructed datasets to evaluate several LLMs.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

LEGALBENCH is available at `https://github.com/HazyResearch/legalbench/`.

What (other) tasks could the dataset be used for?

We envision this dataset could be used for the following:

- Evaluation of LLMs.
- Finetuning LLMs, either on task data directly, or self-instruct style generations derived from task data.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

We emphasize that LEGALBENCH—like all generalized benchmarks—can offer only a preliminary understanding of LLM performance. LEGALBENCH tasks do not generalize to all legal reasoning tasks or all types of legal documents. We thus emphasize that practitioners seeking to deploy LLMs within their own applications should perform their own data collection and validation specific to their use case.

Are there tasks for which the dataset should not be used? If so, please provide a description.

These datasets should not be used to predict the legality of real world events, the outcome of lawsuits, or as legal advice.

C.6 Distribution

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

Table 8 provides the license that applies to each individual LEGALBENCH task.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

Yes. Tasks which consist of adapted/transformed data are released under the same license as the original dataset. Table 8 provides these licenses, and Section F provides a reference to the original dataset for transformed tasks.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

C.7 Maintenance

Who will be supporting/hosting/maintaining the dataset?

Neel Guha will be supporting this dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

Neel Guha can be reached at nguha@cs.stanford.edu. He will be available to answer any questions.

Is there an erratum? If so, please provide a link or other access point.

We have currently not found any, but will make them available on the website.

Willthe dataset be updated (e.g.,to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes. There will be two types of updates to LEGALBENCH:

- First, we will update LEGALBENCH to reflect new contributions from the legal community.
- Second, we will update LEGALBENCH to reflect identified errors in the data.

We will strive to make and publicize updates as soon as errors are identified and new tasks are contributed. Neel Guha will be in charge of managing these updates.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told thattheir data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

N/A.

Will older versions ofthe dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes. We will make older versions available on request by email.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users?If so, please provide a description.

Yes. We encourage members of the legal community to contribute new tasks. We are in the process of formalizing procedures for reviewing, validating, and incorporating submissions.

We additionally note that many of the LEGALBENCH tasks are available under permissive licenses, and other researchers may thus modify them.

D Task overview

D.1 Licenses

LEGALBENCH tasks are subject to different licenses, due to the choices of dataset contributors, or the license under which the original data was released. Table 8 summarizes the licenses. The authors bear all responsibility in case of violation of rights, and confirm the dataset licenses.

License	Tasks
Creative Commons Attribution 4.0	Abercrombie, CUAD Tasks, Citation Prediction Tasks, Contract NLI Tasks, Contract QA, Corporate Lobbying, Diversity Tasks, Function of Decision Section, Hearsay, Insurance Policy Interpretation, International Citizenship Questions, J.Crew Blocker, Legal Reasoning Causality, MAUD Tasks, Oral Argument Question Purpose, Overruling, Personal Jurisdiction, Private Right of Action, Rule QA, SCALR, Securities Complaint Extraction Tasks, Successor Liability, Supply Chain Disclosure Tasks, Telemarketing Sales Rule, UCC v. Common Law, Unfair Terms of Service
Attribution-NonCommercial 4.0 International	Canada Tax Court Outcomes, Consumer Contracts QA, Textualism Tools
Attribution-ShareAlike 4.0 International	Definition Tasks
Attribution-NonCommercial-ShareAlike 4.0 International	Learned Hands Tasks
MIT	New York State Judicial Ethics, Privacy Policy QA, SARA Tasks
Creative Commons Attribution-NonCommercial License	OPP-115 Tasks
Attribution-NonCommercial 3.0 Unported	Privacy Policy Entailment

Table 8: Licenses

D.2 Public availability status

Given that many commercially available LLMs are trained on the "entirety of the web"—and little is known as to how they are trained—there are concerns that many benchmarks have inadvertantly become part of the training data for these models. Therefore, this section identifies and organizes LEGALBENCH tasks into three categories:

- Previously published tasks, which were derived from datasets that were initially published as part of other works and available on the web for download.
- Original but available tasks, which are original creations of the LEGALBENCH project but previously made available online.
- Original and unavailable tasks, which are original creations of the LEGALBENCH project but have not been released online.

Table 9 summarizes the availability status of each of the tasks.

D.3 Reasoning type

Table 10 organizes tasks by the LEGALBENCH reasoning type they can be used to assess. Table 11 similarily organizes tasks according to reasoning-types recognized in the NLP literature. For each reasoning type, we provide examples of general-domain NLP benchmarks which are similar. For more information on the types of reasoning required for each task, please see the individual task descriptions provided in Appendix F.

D.4 Task statistics

Table 13 provides statistics for the LEGALBENCH tasks. For each task, we list the number of samples and the average length (in words) of each input. LEGALBENCH encompasses tasks ranging from short (a single sentence) to longer inputs (two pages of text)

Publication status	**Tasks**
Previously published tasks	CUAD Tasks, Contract NLI Tasks, MAUD Tasks, OPP-115 Tasks, Overruling, Privacy Policy Entailment, Privacy Policy QA, SARA Tasks, Unfair Terms of Service
Original but available tasks	Abercrombie, Diversity Tasks, Hearsay, International Citizenship Questions, Learned Hands Tasks, New York State Judicial Ethics, Personal Jurisdiction, Private Right of Action, Rule QA
Original and unavailable tasks	Canada Tax Court Outcomes, Citation Prediction Tasks, Consumer Contracts QA, Contract QA, Corporate Lobbying, Definition Tasks, Function of Decision Section, Insurance Policy Interpretation, J.Crew Blocker, Legal Reasoning Causality, Oral Argument Question Purpose, SCALR, Securities Complaint Extraction Tasks, Successor Liability, Supply Chain Disclosure Tasks, Telemarketing Sales Rule, Textualism Tools, UCC v. Common Law

Table 9: Task publication status

LEGALBENCH reasoning type	**Tasks**
Issue-spotting	Corporate Lobbying, Learned Hands Tasks
Rule-recall	Citation Prediction Tasks, International Citizenship Questions, New York State Judicial Ethics, Rule QA
Rule-application	Abercrombie, Diversity Tasks, Hearsay, Personal Jurisdiction, Successor Liability, Telemarketing Sales Rule, UCC v. Common Law
Rule-conclusion	Abercrombie, Diversity Tasks, Hearsay, Personal Jurisdiction, Successor Liability, Telemarketing Sales Rule, UCC v. Common Law
Interpretation	CUAD Tasks, Consumer Contracts QA, Contract NLI Tasks, Contract QA, Insurance Policy Interpretation, J.Crew Blocker, MAUD Tasks, OPP-115 Tasks, Privacy Policy Entailment, Privacy Policy QA, Private Right of Action, SARA Tasks, Securities Complaint Extraction Tasks, Supply Chain Disclosure Tasks, Unfair Terms of Service
Rhetorical-understanding	Canada Tax Court Outcomes, Definition Tasks, Function of Decision Section, Legal Reasoning Causality, Oral Argument Question Purpose, Overruling, SCALR, Textualism Tools

Table 10: Tasks by LEGALBENCH reasoning type.

(Figure 4). The average LEGALBENCH task contains between 500-600 instances. All tasks consist of at least 50 instances. A more detailed breakdown is available in Table 12.

Table 13: Task Statistics

Task	**Number of Samples**	**Mean Sample Length (Words)**
abercrombie	100	7.1
canada_tax_court_outcomes	250	99.2

Table 13 – continued from previous page

Task	Number of Samples	Mean Sample Length (Words)
citation_prediction_classification	110	35.9
citation_prediction_open	55	32.4
consumer_contracts_qa	400	486.8
contract_nli_confidentiality_of_agreement	90	73.9
contract_nli_explicit_identification	117	75.7
contract_nli_inclusion_of_verbally_conveyed_information	147	78.1
contract_nli_limited_use	216	63.8
contract_nli_no_licensing	170	65.6
contract_nli_notice_on_compelled_disclosure	150	77.9
contract_nli_permissible_acquirement_of_similar_information	186	66.6
contract_nli_permissible_copy	95	60.2
contract_nli_permissible_development_of_similar_information	144	61.0
contract_nli_permissible_post-agreement_possession	119	81.7
contract_nli_return_of_confidential_information	74	74.2
contract_nli_sharing_with_employees	178	83.9
contract_nli_sharing_with_third-parties	188	78.9
contract_nli_survival_of_obligations	165	64.6
contract_qa	88	48.3
corporate_lobbying	500	878.1
cuad_affiliate_license-licensee	204	75.9
cuad_affiliate_license-licensor	94	99.7
cuad_anti-assignment	1178	54.6
cuad_audit_rights	1222	53.6
cuad_cap_on_liability	1252	59.8
cuad_change_of_control	422	62.8
cuad_competitive_restriction_exception	226	67.1
cuad_covenant_not_to_sue	314	64.6
cuad_effective_date	242	44.5
cuad_exclusivity	768	58.0
cuad_expiration_date	882	49.9
cuad_governing_law	882	46.9
cuad_insurance	1036	56.8
cuad_ip_ownership_assignment	582	65.3
cuad_irrevocable_or_perpetual_license	286	72.1
cuad_joint_ip_ownership	198	59.1
cuad_license_grant	1402	63.5
cuad_liquidated_damages	226	57.2
cuad_minimum_commitment	778	57.9
cuad_most_favored_nation	70	68.0
cuad_no-solicit_of_customers	90	61.2
cuad_no-solicit_of_employees	148	66.6

Table 13 – continued from previous page

Task	Number of Samples	Mean Sample Length (Words)
cuad_non-compete	448	60.6
cuad_non-disparagement	106	63.8
cuad_non-transferable_license	548	61.5
cuad_notice_period_to_terminate_renewal	228	57.0
cuad_post-termination_services	814	67.3
cuad_price_restrictions	52	53.6
cuad_renewal_term	392	55.2
cuad_revenue-profit_sharing	780	59.8
cuad_rofr-rofo-rofn	696	64.0
cuad_source_code_escrow	124	64.4
cuad_termination_for_convenience	436	52.0
cuad_third_party_beneficiary	74	42.4
cuad_uncapped_liability	300	70.1
cuad_unlimited-all-you-can-eat-license	54	56.8
cuad_volume_restriction	328	49.0
cuad_warranty_duration	326	55.9
definition_classification	1345	41.0
definition_extraction	695	53.9
diversity_1	306	16.2
diversity_2	306	23.4
diversity_3	306	21.1
diversity_4	306	23.5
diversity_5	306	28.3
diversity_6	306	47.9
function_of_decision_section	374	87.4
hearsay	100	25.3
insurance_policy_interpretation	138	87.2
international_citizenship_questions	9310	33.2
intra_rule_distinguishing	60	33.8
jcrew_blocker	60	167.2
learned_hands_benefits	72	253.2
learned_hands_business	180	225.3
learned_hands_consumer	620	246.3
learned_hands_courts	198	225.6
learned_hands_crime	694	233.7
learned_hands_divorce	156	240.2
learned_hands_domestic_violence	180	262.3
learned_hands_education	62	265.7
learned_hands_employment	716	242.0
learned_hands_estates	184	230.9
learned_hands_family	2271	258.4

Table 13 – continued from previous page

Task	**Number of Samples**	**Mean Sample Length (Words)**
learned_hands_health	232	286.8
learned_hands_housing	4500	254.8
learned_hands_immigration	140	232.8
learned_hands_torts	438	272.2
learned_hands_traffic	562	229.6
legal_reasoning_causality	59	245.7
maud_"ability_to_consummate"_concept_is_subject_to_mae_carveouts	70	688.6
maud_"financial_point_of_view"_is_the_sole_consideration	113	307.5
maud_accuracy_of_fundamental_target_r&ws:_bringdown_standard	176	143.7
maud_accuracy_of_target_"general"_r&w:_bringdown_timing_answer	182	142.5
maud_accuracy_of_target_capitalization_r&w_(outstanding_shares):_bringdown_standard_answer	182	142.0
maud_additional_matching_rights_period_for_modifications_(cor)	159	314.3
maud_application_of_buyer_consent_requirement_(negative_interim_covenant)	181	85.6
maud_buyer_consent_requirement_(ordinary_course)	182	121.3
maud_change_in_law:__subject_to_"disproportionate_impact"_modifier	100	702.6
maud_changes_in_gaap_or_other_accounting_principles:__subject_to_"disproportionate_impact"_modifier	99	703.1
maud_cor_permitted_in_response_to_intervening_event	101	305.6
maud_cor_permitted_with_board_fiduciary_determination_only	101	303.1
maud_cor_standard_(intervening_event)	85	326.1
maud_cor_standard_(superior_offer)	101	308.0
maud_definition_contains_knowledge_requirement_-_answer	148	243.4
maud_definition_includes_asset_deals	147	314.7
maud_definition_includes_stock_deals	149	313.6
maud_fiduciary_exception:__board_determination_standard	180	246.8
maud_fiduciary_exception:_board_determination_trigger_(no_shop)	180	245.1
maud_fls_(mae)_standard	78	705.1
maud_general_economic_and_financial_conditions:_subject_to_"disproportionate_impact"_modifier	99	704.6
maud_includes_"consistent_with_past_practice"	182	122.7
maud_initial_matching_rights_period_(cor)	159	313.4
maud_initial_matching_rights_period_(ftr)	133	336.4
maud_intervening_event_-_required_to_occur_after_signing_-_answer	148	242.2
maud_knowledge_definition	168	334.8
maud_liability_standard_for_no-shop_breach_by_target_non-d&o_representatives	157	38.1

Table 13 – continued from previous page

Task	Number of Samples	Mean Sample Length (Words)
maud_ordinary_course_efforts_standard	182	122.7
maud_pandemic_or_other_public_health_event:__subject_to_ "disproportionate_impact"_modifier	99	707.5
maud_pandemic_or_other_public_health_event:_specific_ reference_to_pandemic-related_governmental_responses_or_ measures	99	707.5
maud_relational_language_(mae)_applies_to	91	705.5
maud_specific_performance	179	96.8
maud_tail_period_length	180	95.1
maud_type_of_consideration	173	128.2
nys_judicial_ethics	300	25.7
opp115_data_retention	96	31.5
opp115_data_security	1342	38.6
opp115_do_not_track	118	37.1
opp115_first_party_collection_use	2094	32.6
opp115_international_and_specific_audiences	988	52.1
opp115_policy_change	439	33.2
opp115_third_party_sharing_collection	1598	35.3
opp115_user_access,_edit_and_deletion	470	35.1
opp115_user_choice_control	1554	33.5
oral_argument_question_purpose	319	50.2
overruling	2400	27.5
personal_jurisdiction	54	67.8
privacy_policy_entailment	4343	111.9
privacy_policy_qa	10931	41.1
proa	100	42.6
rule_qa	50	11.7
scalr	571	275.4
ssla_company_defendants	1231	310.0
ssla_individual_defendants	1015	313.7
ssla_plaintiff	1036	308.4
sara_entailment	276	148.0
sara_numeric	100	12222.1
successor_liability	50	71.5
supply_chain_disclosure_best_practice_accountability	387	510.0
supply_chain_disclosure_best_practice_audits	387	508.3
supply_chain_disclosure_best_practice_certification	386	508.4
supply_chain_disclosure_best_practice_training	387	508.2
supply_chain_disclosure_best_practice_verification	387	507.0
supply_chain_disclosure_disclosed_accountability	386	510.4
supply_chain_disclosure_disclosed_audits	387	508.0
supply_chain_disclosure_disclosed_certification	386	509.9

Table 13 – continued from previous page

Task	**Number of Samples**	**Mean Sample Length (Words)**
supply_chain_disclosure_disclosed_training	387	506.9
supply_chain_disclosure_disclosed_verification	387	507.6
telemarketing_sales_rule	51	58.4
textualism_tool_dictionaries	111	151.3
textualism_tool_plain	169	160.9
ucc_v_common_law	100	20.9
unfair_tos	3822	34.3

NLP reasoning type	Tasks
Knowledge (e.g., MMLU [61], WikiFact in HELM [80, 108])	Citation Prediction Tasks, International Citizenship Questions, New York State Judicial Ethics, Rule QA
Linguistic inference (e.g., CoLA [144])	Canada Tax Court Outcomes, Definition Tasks, Function of Decision Section, Legal Reasoning Causality, Oral Argument Question Purpose, Overruling, Textualism Tools
Topic classification (e.g., RAFT [1])	Corporate Lobbying, Learned Hands Tasks, CUAD Tasks, J.Crew Blocker, OPP-115 Tasks, Private Right of Action, Unfair Terms of Service
Entailment (e.g., RTE [140])	Contract NLI Tasks, Privacy Policy Entailment, Privacy Policy QA, Insurance Policy Interpretation, SARA Tasks
Arithmetic (e.g., GSM8K [34])	Diversity Tasks, SARA Tasks
Multi-step reasoning (e.g., STREET [113])	Abercrombie, Hearsay, Personal Jurisdiction, Successor Liability, Telemarketing Sales Rule, UCC v. Common Law
Document-based QA (e.g., BoolQ [33])	Consumer Contracts QA, Contract QA, MAUD Tasks, Supply Chain Disclosure Tasks
Named entity recognition (e.g., CoNLL-2003 [115])	Securities Complaint Extraction Tasks
Casual reasoning (e.g., CoPA [140])	SCALR

Table 11: Tasks by NLP reasoning type. For each reasoning type, we provide examples of general-domain NLP benchmarks which are similar. For more information on the types of reasoning required for each task, please see the individual task descriptions provided in Appendix F.

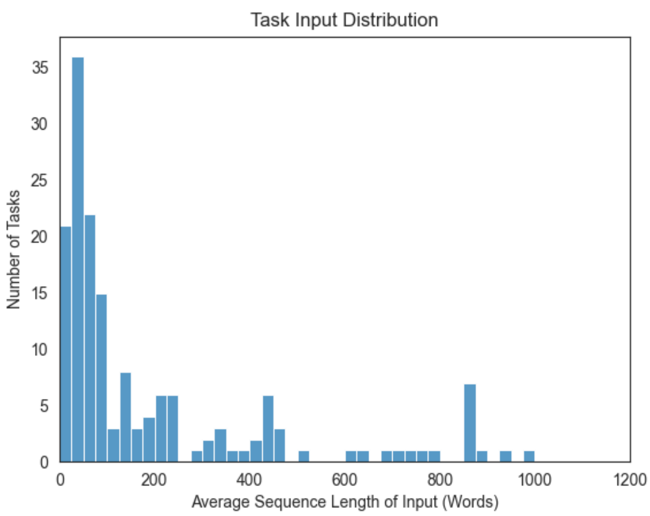


Figure 4: LEGALBENCH task sizes and input text lengths.

Size range (samples)	**Number of tasks**
50-100	28
100-500	97
500-2000	29
2000+	8

Table 12: Number of LEGALBENCH tasks at different dataset sizes.

E Evaluation

This section describes metrics and evaluation protocols.

Rule-application To evaluate an LLM's performance on a rule-application task, a law-trained expert manually validated each generation. We computed two metrics. The first metric—*correctness*—corresponds to the proportion of generations which do not misstate the fact pattern, the legal outcome, the rule, or contain a logical error. Logical errors include arithmetic mistakes, or assertions which are plainly wrong (e.g., that an apple is not a tangible object).

- The LLM would incorrectly assert the legal outcome. For instance, an LLM would assert diversity jurisdiction existed, when it actually did not.
- The LLM would incorrectly assert an intermediate conclusion. For instance, an LLM would assert that a rental agreement for a boat was a contract for a service, rather than a moveable and tangible good. In this case, the LLM would fail to realize that a boat is a moveable or tangible good.
- The LLM would hallucinate a piece of information not explicitly stated in the fact pattern.
- The LLM would misstate the content of a rule. For instance, the LLM would assert that subprovision of a statute barred one type of conduct, when in fact, that conduct was barred by a different subprovision.

The second metric—*anlaysis*—corresponds to the proportion of generations which contain the necessary inferences to reach the correct legal conclusion from the provided fact-pattern. Thus, it is insufficient for a LLM explanation to merely state that a piece of evidence is hearsay is insufficient: an explanation must reference the qualities of the evidence which make it hearsay. We introduced this measurement after discovering that for some tasks, LLMs often generate explanations which—though correct——largely restate the rule being applied, without any reference to the underlying facts. Explanations which are incorrect are automatically deemed to be insufficient on analysis. We compute an overall analysis score for a LLM on a task by measuring the proportion of samples for which the explanation contains sufficient analysis.

To standardize evaluation and enable future work, we have released an "answer guide" for each task used for rule-application, which contains the inferences required for each sample, and describes common modes of errors. All evaluations in LEGALBENCH for rule-application have been performed with respect to this answer-guide.

Generation tasks LEGALBENCH contains the following generation-tasks, which are evaluated as follows:

- `rule_qa` is a question-answer task in which a LLM must generate a response to an open-ended question. A law-trained individual evaluated the generations against an answer-key for the task, which is available for download from the website.
- The Securities Complaint Extraction tasks require an LLM to extract the names of different parties from excerpts of securities class-action complaints. Because some samples require the extraction of multiple entities, we evaluate using F1 score.
- `definition_extraction` requires the LLM to identify the term that is being defined in a sentence from a Supreme Court opinion. For a small number of sentences, any one of multiple terms may constitute the correct answer. We therefore evaluate performance on this task using accuracy, and by counting the fraction of sentences for which the LLM identified a permissible term. To account for edge-cases involving word tense, we compare stemmed versions of the answers to stemmed versions of the generation.
- `sara_numeric` requires an LLM to generate an estimate of the amount of tax that is owed. We compute performance here using an accuracy metric, which treats a prediction as accurate if it is within 10% of the true amount.
- `citation_open` requires an LLM to predict the name of the case that should be cited for a particular sentence of judicial text. We evaluate by checking if the LLM generation contains the correct case name.
- `successor_liability` requires an LLM to identify the multiple possible successor liability exceptions to a fact pattern. We evaluate using F1.

Classification tasks We evaluate all classification tasks in LEGALBENCH using exact-match on class-balanced-accuracy. We do this because a number of LEGALBENCH tasks are class-imbalanced.

F Task descriptions

This section provides a detailed description for each family of tasks.

F.1 Abercrombie

In LEGALBENCH, the Abercrombie task is denoted as `abercrombie`.

Background A particular mark (e.g., a name for a product or service) is only eligible for trademark protection if it is considered to be distinctive. In assessing whether a mark is distinctive, lawyers and judges follow the framework set out in the case *Abercrombie & Fitch Co. v. Hunting World, Inc.*,[14] which enumerates five categories of distinctiveness. These categories characterize the relationship between the dictionary definition of the term used in the mark, and the service or product it is being attached to. They are:

- Generic: A name is generic with respect to a product or service if it connotes the basic nature of the product/service, rather than more individualized characteristics of the product. For example, the mark "Salt" for packaged sodium chloride would be generic under Abercrombie because "salt" is the common name for sodium chloride. It is also common to think of generic marks as merely referring to the class of goods for which a particular product is a species.
- Descriptive: A name is descriptive if it identifies a characteristic or quality of an article or service, such as color, odor, function, dimensions, or ingredients. For example, the name "Sharp" for a television would be descriptive, because it describes a plausible characteristic of television (i.e., their sharp image quality).
- Suggestive: A name is suggestive if it suggests, rather than describes, some particular characteristic of the goods or services to which it applies. An essential aspect of suggestive names is that it requires the consumer to exercise the imagination in order to draw a conclusion as to the nature of the goods and services. For example, the name "Greyhound" would be suggestive for a bus service, because greyhounds are considered to be fast, and "fast" is an adjective that could be used to describe a bus service.
- Arbitrary: A name is arbitrary if it is a "real" word but seemingly "arbitrary" with respect to the product or service. For example, the mark "Apple" for a software company is arbitrary, because apples are unrelated to software.
- Fanciful: A name is fanciful if it is entirely made up, and not found in the English dictionary. For example, "Lanmbe" is a fanciful mark, because it is a made-up word.

The Abercrombie spectrum is commonly taught as part of Intellectual Property courses in law school, and students are expected to understand how to determine the Abercrombie classification for a particular product/mark combination.

Performing the Abercrombie task requires reasoning about the literal meaning of a word and the degree of its connection to a particular product/service. It requires having some understanding of the types of words that could plausibly be used to describe a particular good/service, and the extent to which those words relate to a particular mark. It also requires reasoning as to whether a particular word is a real English word.

Task The Abercrombie task requires an LLM to determine–given a candidate mark and a description of a product/service–which of the five Abercrombie categories above apply.

Facts	Abercrombie Classification
The mark "Whirlpool" for an oven.	arbitrary
The mark "Compact" for wallets.	descriptive
The mark "Imprion" for a line of sports drinks.	fanciful
The mark "Car" for a line of automobiles.	generic
The mark "Quick Green" for grass seed.	suggestive

Table 14: Task examples.

Construction process We manually create a dataset to evaluate a model's ability to classify a mark's distinctiveness (into one of the above 5 categories) with respect to a product. In writing samples, we draw inspiration from similar exercises available in legal textbooks and practice study guides. Hence, the samples provided have a definite answer, and are not subject to ambiguity. There is an expectation that a law student learning intellectual property would be able to answer these questions to a high degree of accuracy.

We create approximately 20 samples for each category of distinctiveness, and randomly select a single sample from each category to constitute the train set. The remaining 19 samples (for each category) are assigned to the test set (for a total of 95 samples).

[14] *Abercrombie & Fitch Co. v. Hunting World*, 537 F.2d 4 (2nd Cir. 1976).

Class	**Number of samples**
generic	19
descriptive	19
suggestive	20
arbitrary	18
fanciful	19

Table 15: Test set class distribution.

Significance and value Given how easy this task is for lawyers with a basic training in intellectual property law, it is unlikely that LLMs will be called on to perform this task in the actual practice of law, or that the ability for LLMs to perform this task would alter the way in which lawyers approach IP practice. Instead, the Abercrombie task is significant as a measurement of reasoning ability. Because it is "simplistic" by the standards of human lawyers, it provides a useful objective measure of reasoning progress for LLMs.

F.2 Canada Tax Court Outcomes

In LEGALBENCH, the Canada Tax Court Outcomes task is also denoted as `canada_tax_court_outcomes`.

Background The Tax Court of Canada hears appeals of government decisions related to taxation.[15] The Court's decisions, which are written in natural language, are published on the Court's website, in both French and English.[16] Decisions typically include a section at the beginning summarizing the outcome of the appeal, followed by sections describing the factual background and various procedural steps, a section identifying the issues under consideration, sections with legal analysis, and a concluding section. While this is the standard format, judges are free to use other formats if they prefer. Decision length varies depending on the complexity of the litigation, with some decisions being only a few hundred words, and others being many thousands of words.

Appeals in Tax Court of Canada cases are brought by individuals or organizations who ask the Court to overturn a government taxation decision. Outcomes of appeals are generally binary: appeals are either granted, in which case the government taxation decision is overturned in whole or in part, or appeals are denied in which case the government taxation decision is upheld. Occasionally published decisions will not involve the outcome of an appeal, including where the decision is about a procedural step (e.g. the admissibility of particular evidence).

The `canada_tax_court_outcomes` task involves identifying whether an excerpt from a Tax Court of Canada decision includes the outcome of the appeal and, if so, what the outcome is. While the task is straightforward, one challenge is that the model must distinguish between outcomes of the appeal as a whole and outcomes of particular aspects of the appeal. Another challenge is that where the excerpt does not include the outcome, the model must avoid predicting the outcome – even if the model might plausibly correctly infer the likely outcome from the excerpt provided.

Task The Canada Tax Court Outcomes task requires an LLM to classify whether an excerpt from a given decision includes the outcome of the appeal, and if so whether the appeal was allowed or dismissed. Some excerpts do not include an outcome of the appeal, in which case the model should return 'other'. Where the excerpt includes the outcome and the appeal is allowed in whole or in part, the model should return 'allowed'. Where the excerpt includes an outcome, and the appeal is dismissed the model should return 'dismissed'. The model should disregard outcomes that are not about the ultimate outcome of the appeal, such as costs awards (i.e. orders requiring a party to pay the other party's legal costs).

Construction process We obtained the full text of English-language versions of decisions from 2001 to 2022 by scraping the Tax Court of Canada website.[17] We then cleaned and parsed the text to extract excerpts that are most likely to contain the outcome of the appeal. For example, many decisions contain a brief introductory section describing the outcome of the appeal using a specific header, and if the decision contained a section with such a header, we excerpted only that section. Where our parsing code could not identify such a section, we excerpted the first and last 2,500 characters, because outcomes are generally described at either the beginning or end of decisions. After initially attempting outcome classification on these excerpts using OpenAI's ChatGPT, we selected a quasi-random sample of 250 excerpts (quasi random because we selected these manually, we over- sampled excerpts where the outcome is 'other', and we chose some excerpts that were challenging due to factors such as length or unusual format). We manually reviewed outcomes for these excerpts, correcting some that had been miscategorized.

Two random cases from each class are selected for the training split, while the remainder are used as the test set.

Significance and value Legal scholars frequently gather data about outcomes in large numbers of legal decisions in order to examine patterns in judicial decision-making. For example, a legal scholar may be interested in comparing outcomes in similar processes across jurisdictions or they might examine whether a legislative change resulted in different outcomes over time. Lawyers and legal information technology companies may also be interested in gathering data on outcomes for the purposes of judicial analytics or to predict future outcomes.

Gathering such data is typically straightforward. It is, for example, a common task assigned to first year law student research assistants who can frequently achieve close to 100% accuracy on such tasks with only minimal training. However, because the data is often useful only when gathered on large numbers of decisions, this type of data gathering using human research assistants can be cost prohibitive. If LLMs can obtain high accuracy on these tasks, substantial savings could be achieved – which would increase the ability of researchers to pursue new projects.

[15] *Tax Court of Canada Act*, RSC, 1985, c T-2, online: `https://laws-lois.justice.gc.ca/eng/acts/t-2/index.html`, s 12.

[16] Tax Court of Canada, "Find a Decision", online: `https://decision.tcc-cci.gc.ca/tcc-cci/en/nav.do`.

[17] Ibid. As per the terms of service of the website, we are required to note that the text of the scraped decisions are not the official versions (official versions can be obtained from the website), and that the reproduction of these cases has not been produced in affiliation with or with the endorsement of the Government of Canada.

Excerpt	Outcome
The appeal is allowed in part and the assessment is referred back to the Minister of National Revenue for reconsideration and reassessment to reflect a 25% reduction of the tax owed by the appellant and adjustments to the interest and penalties, as agreed to by the respondent. Costs are to be determined after hearing both parties. In all other respects, the assessment is confirmed. Signed this 23rd day of February 2012. "Franois Angers" Angers J.	allowed
IN ACCORDANCE with the Reasons for Judgment attached, the appeal from the decision of the Respondent in relation to the income of the Appellant for the purposes of determining his entitlement to the Guaranteed Income Supplement under the Old Age Security Act for the payment period from July 1, 2014 to June 30, 2015 is dismissed, without costs. Signed at Ottawa, Canada, this 24th day of October 2017. R.S. Bocock Bocock J.	dismissed
(These Reasons for Judgment are issued in substitution for the Reasons for Judgment signed on January 22, 2002) Lamarre, J. [1] These are appeals under the informal procedure against assessments made by the Minister of National Revenue ("Minister") under the Income Tax Act ("Act") for the 1995, 1996, 1997, 1998 and 1999 taxation years. [2] In filing her 1995 income tax return, the appellant claimed a business investment loss of $268,897 with respect to investments in eight mortgages held "in trust" for the appellant and her father Henry Sokolowski by Kiminco Acceptance Co. Ltd. ("Kiminco"), a member of the Glen Coulter group of companies. The eight mortgage investments were made in 1987 and 1988 and are identified as follows in paragraph 13 of the Reply to the Notice of Appeal: Account/Mortgage Number Ultimate Borrower ... For the Appellant: Name: Firm: For the Respondent: Morris Rosenberg Deputy Attorney General of Canada Ottawa, Canada	other

Table 16: Task examples.

Outcome	Number of samples
allowed	101
dismissed	131
other	12

Table 17: Test set class distribution.

F.3 Citation Prediction Tasks

In the LEGALBENCH, the Citation Prediction tasks are also denoted as `citation_prediction_*`.

Background The importance of locating relevant legal materials, or "law search" has long been recognized as an essential aspect of legal practice. This process involves uncovering case law, statutes, and other materials pertinent to legal questions or arguments. As a fundamental aspect of legal reasoning, law search plays a crucial role in bridging the gap between the initial translation of behaviors into legal questions and the subsequent interpretation and application of the relevant law.

Legal professionals are often valued for their ability to find and apply the appropriate law to their clients' situations. Given the intricate nature of the contemporary legal domain, the process of law search has evolved into a complex and nuanced task that demands a comprehensive understanding of the law.

A core component of law search is legal relevance. From a sociological perspective, the relevance of legal documents to a specific legal question is a social fact. This fact is determined by the judgments made by members of the legal community, who must determine which legal materials are applicable to a given question. Relevance relates legal questions to sources of legal authority.

In functional legal communities, law search leads to some degree of convergence over legal materials. Convergence occurs when competent members of a legal community, faced with the same legal question, identify the same sources of relevant legal authority. This process is essential to ensuring that the legal system operates consistently, predictably, and coherently.

As a critical process that connects the translation of behaviors into legal questions and the subsequent interpretation and application of the relevant law, law search is indispensable to legal reasoning.

The Citation Prediction task requires reasoning concerning the relationship between the text of judicial opinions and legal propositions. Successful prediction would entail encoding a notion of legal relevance and would allow a system to determine whether a legal proposition was or was not supported by the extant body of law.

Task The citation task is based on a version of the evaluation approaches used in [39]. There are two Citation Prediction tasks. The first (`citation_prediction_classification`) requires an LLM to predict whether a given sentence (i.e. legal proposition) is or is not supported by a given case. The second (`citation_prediction_open`) requires an LLM to predict a case (by name) that supports a provided sentence.

Construction process We collected a sample of circuit court opinions published after January 1, 2023. To the best of our knowledge, most existing LLMs haven't been trained on any data generated in 2023. For each opinion, we manually collected sentences which were supported by a citation to a judicial opinion, where (1) the sentence contained some quotation from the original case, and (2) the sentence was supported by a single cite. We chose sentences which included quotation fragments and were only supported by a single cite to avoid sentences which could be supported by a broad set of cases. When a sentence is supported by a much larger universe of cases, verifying that an LLM answer is incorrect is difficult. We also recorded the circuit for each opinion that we pulled language from. As a result, we can include the circuit information in the prompt, since circuits prefer citing their previous decisions. We collected 55 sentences using this process. For the citation generation task (`citation_prediction_open`), we ask the LLM to predict the citation given the sentence.

The `citation_prediction_classification` task is then constructed as follows. We use each sentence-citation pair to create two task samples. The first sample corresponds to the sentence and the correct citation (positive label). The second sample corresponds to the sentence and a randomly selected citation from the remainder of the data (negative label). This generates a dataset of 110 sentence-citation pairs, two of which are assigned to the training split.

Input	Citation	Supported?
Exclusions are always strictly construed against the insurer and in favor of the insured.	Nationwide Mut. Ins. Co. v. Cosenza	Yes
The Supreme Court and this court have repeatedly "held that environmental plaintiffs adequately allege injury in fact when they aver that they use the affected area and are persons for whom the aesthetic and recreational values of the area will be lessened by the challenged activity."	United States v. Pearce	No

Table 18: Examples for `citation_prediction_classification`.

Input	Citation
In other words, the DJA "creates a means by which rights and obligations may be adjudicated in cases involving an actual controversy that has not reached the stage at which either party may seek a coercive remedy."	United States v. Doherty
To be “equivalent to a demotion,” the action need not "result in a decrease in pay, title, or grade; it can be a demotion if the new position proves objectively worse—such as being less prestigious or less interesting or providing less room for advancement."	Alvarado v. Tex. Rangers

Table 19: Examples for `citation_prediction_open`.

Significance and value Law search is a core function of legal thinking. In addition, the difficulty of identifying relevant law is a core barrier in the public’s ability usefully access the law. The ability of an LLM to accurately engage in citation prediction would have important practical value in providing access to law, and would also allow the LLM to more reliably support legal statements with relevant authority.

F.4 Clause Classification Tasks

LEGALBENCH includes a number of tasks in which the LLM must determine the "type" or "category" of a provision/clause in a legal document. Specifically:

- The Contract QA task (Section F.4.4), in which the LLM is provided with the name of a common type of contractual clause and a clause, and must determine if the clause is an example of the example type.
- 38 tasks derived from the CUAD dataset (Section F.4.1), where each task is a binary-classification task requiring the LLM to identify whether a clause (from an EDGAR contract) belongs to a certain category (e.g., audit rights clauses) [62].
- The J.Crew blocker task (Section F.4.2), in which the LLM must classify whether a clause (from a loan agreement) is a J-Crew blocker provision.
- The Unfair Terms of Service task (Section F.4.3), in which the LLM must classify a clause (from a terms of service agreement) to one of eight types, where seven of the types denote clauses that would potentially be considered "unfair" under European law [82].

Significance and value Lawyers spend significant time and energy reviewing legal documents (e.g., contracts, leases, etc.). Manual review serves an important purpose, allowing parties to identify potentially problematic terms [119]. Parties will sometimes review agreements that have already been signed, in response to changing world events. For instance, the COVID-19 pandemic led many firms to inspect agreements for the existence of *force majeure* clauses, which ordinarily specify how contractual expectations should be handled in the event of major world crises [49]. Because legal documents are long and require legal training to understand, the process of reviewing is often extremely expensive [62]. This, in turn, presents significant access-to-justice concerns. Because most individuals do not have the financial capacity to consult lawyers prior to entering legal agreements, they are oblivious to when those agreements contain predatory, oppressive, or unconscionable terms. A rich legal scholarship has noted, for instance, the frequency at which legal agreements contain terms that would be invalidated by a court [25].

The clause classification tasks in LEGALBENCH are thus amongst the most *practically useful* tasks in LEGALBENCH, as they capture an actual current-day use case for LLMs. As the complexity of clause classification depends both on the clause category and document type, LEGALBENCH tasks span a range of clause categories and source documents.

F.4.1 CUAD Tasks

We adapt the CUAD dataset for LEGALBENCH [62]. The original dataset consists of 500 contracts, each annotated with up to 41 different clause types. These contracts varied significantly in length, ranging from a few pages to over one-hundred pages. In the original word, [62] studied the ability for BERT-base language models to identify the text spans corresponding to different types of clauses. The principal difficulties were (1) the length of the contract, and (2) the lack of significant training data.

We adapt the CUAD dataset as follows. We select 38 of the 41 clause categories. For each selected category, we construct a dataset consisting of (1) clauses in the CUAD contracts which are assigned to that category, and (2) an equal number of clauses randomly sampled from other categories. This produces a balanced binary classification task for clause category, where the purpose is to identify which clauses belong to the respective category. A table with the selected categories, and their descriptions is found below.

Table 20 lists each task, a "description" of the category corresponding to the task, and an example of a clause which meets the category criteria. In accordance with [62], the description is presented as the question posed to the annotators during data labeling. If a clause yields an affirmative answer with regards to the question, then the label is "Yes". Otherwise the label is "No".

In LEGALBENCH, the CUAD tasks are denoted as `cuad_*`.

Table 20: CUAD Tasks

Task
Task name: cuad_affiliate_license-licensee **Description**: Does the clause describe a license grant to a licensee (incl. sublicensor) and the affiliates of such licensee/sublicensor? **Example**: [***], Valeant hereby grants to Dova a fully paid-up, royalty free, non-transferable, non- exclusive license (with a limited right to sub-license to its Affiliates) to any Valeant Property that appears on, embodied on or contained in the Product materials or Product Labeling solely for use in connection with Dova's promotion or other commercialization of the Product in the Territory.
Task name: cuad_affiliate_license-licensor **Description**: Does the clause describe a license grant by affiliates of the licensor or that includes intellectual property of affiliates of the licensor?

Table 20 – continued from previous page

Task
Example: "Company Licensed Know-How" means all Know-How owned by any Company Entity as of the Effective Date and used or held for use in the Arizona Field as of the Effective Date.<omitted>Subject to the terms and conditions of this Agreement, the Company hereby grants to Seller a perpetual, non- exclusive, royalty-free license in, to and under the Company Licensed Know-How for use in the Arizona Field throughout the world.
Task name: cuad_anti-assignment **Description**: Does the clause require consent or notice of a party if the contract is assigned to a third party? **Example**: Except as otherwise set forth herein, neither party shall transfer, assign or cede any rights or delegate any obligations hereunder, in whole or in part, whether voluntarily or by operation of law, without the prior written consent of the other party, which consent may be withheld at the other party's reasonable business discretion; provided, however, that either party may transfer this Agreement without prior written consent of the other to an Affiliate of such party, or to the surviving party in a merger or consolidation, or to a purchaser of all or substantially all of its assets.
Task name: cuad_audit_rights **Description**: Does the clause give a party the right to audit the books, records, or physical locations of the counterparty to ensure compliance with the contract? **Example**: For avoidance of doubt, all audits under this Section shall be conducted solely by an independent public accountant as described in the foregoing sentence.
Task name: cuad_cap_on_liability **Description**: Does the clause specify a cap on liability upon the breach of a party's obligation? This includes time limitation for the counterparty to bring claims or maximum amount for recovery. **Example**: EXCEPT FOR LIABILITIES UNDER SECTION 7.2 [Indemnity], NEITHER PARTY'S AGGREGATE LIABILITY ARISING OUT OF OR IN CONNECTION WITH THIS AGREEMENT OR THE TRANSACTIONS CONTEMPLATED HEREBY, WHETHER IN CONTRACT, TORT (INCLUDING NEGLIGENCE), WARRANTY OR OTHERWISE, SHALL EXCEED [***].
Task name: cuad_change_of_control **Description**: Does the clause give one party the right to terminate or is consent or notice required of the counterparty if such party undergoes a change of control, such as a merger, stock sale, transfer of all or substantially all of its assets or business, or assignment by operation of law? **Example**: Notwithstanding the foregoing, if any Party to this Agreement (or any of its successors or permitted assigns) (a) shall enter into a consolidation or merger transaction in which such Party is not the surviving entity and the surviving entity acquires or assumes all or substantially all of such Party's assets, (b) shall transfer all or substantially all of such Party's assets to any Person or (c) shall assign this Agreement to such Party's Affiliates, then, in each such case, the assigning Party (or its successors or permitted assigns, as applicable) shall ensure that the assignee or successor- in-interest expressly assumes in writing all of the obligations of the assigning Party under this Agreement, and the assigning Party shall not be required to seek consent, but shall provide written notice and evidence of such assignment, assumption or succession to the non-assigning Party.
Task name: cuad_competitive_restriction_exception **Description**: Does the clause mention exceptions or carveouts to Non-Compete, Exclusivity and No-Solicit of Customers? **Example**: Notwithstanding the foregoing, Excite may make available opportunities on the Excite Site to purchase Music Products from parties other than Sponsor if such Music Products are not available from Sponsor so long as, prior to entering into arrangements to make available opportunities to purchase Music Products from parties other than Sponsor, Excite notifies Sponsor of its interest in the Music Products and gives Sponsor thirty (30) days to make the desired Music Products available through the Sponsor Site.
Task name: cuad_covenant_not_to_sue **Description**: Is a party restricted from contesting the validity of the counterparty's ownership of intellectual property or otherwise bringing a claim against the counterparty for matters unrelated to the contract? **Example**: In connection with any reference to the Trademarks, Distributor shall not in any manner represent that it has an ownership interest in the Trademarks or registration(s) thereof, and Distributor acknowledges that no action by it or on its behalf shall create in Distributor's favor any right, title, or interest in or to the Trademarks.
Task name: cuad_exclusivity **Description**: Does the clause specify exclusive dealing commitment with the counterparty? This includes a commitment to procure all "requirements" from one party of certain technology, goods, or services or a prohibition on licensing or selling technology, goods or services to third parties, or a prohibition on collaborating or working with other parties), whether during the contract or after the contract ends (or both). **Example**: Bosch hereby grants to Client the exclusive rights to sell and distribute the Product, subject to the Territory as set forth below, to certain select companies in the Automotive Industry, each of which shall be approved by Bosch in writing as requested by the Client on a case by case basis.
Task name: cuad_insurance **Description**: Is there a requirement for insurance that must be maintained by one party for the benefit of the counterparty?

Table 20 – continued from previous page

Task
Example: Throughout the entire Term, you must maintain such types of insurance, in such amounts, as we may require.
Task name: cuad_ip_ownership_assignment **Description**: Does intellectual property created by one party become the property of the counterparty, either per the terms of the contract or upon the occurrence of certain events? **Example**: Upon written request of ArTara, University will assign the IND to ArTara.
Task name: cuad_irrevocable_or_perpetual_license **Description**: Does the clause specify a license grant that is irrevocable or perpetual? **Example**: Subject to the terms and conditions of this Agreement, as of the Distribution Date, SpinCo hereby grants to Nuance and the members of the Nuance Group a worldwide, non-exclusive, fully paid-up, perpetual and irrevocable, transferable (subject to ARTICLE VIII), sublicensable (subject to Section 4.01(g)) license to install, access, use, reproduce, perform, display, modify (including the right to create improvements and derivative works), further develop, sell, manufacture, distribute and market products and services based on, using or incorporating the SpinCo Shared Technology Assets within the Nuance Field of Use, together with natural extensions and evolutions thereof.
Task name: cuad_joint_ip_ownership **Description**: Does the clause provide for joint or shared ownership of intellectual property between the parties to the contract? **Example**: If the Domain Name is deemed a combination mark, neither party shall use the Domain Name for any purpose except as expressly provided herein or attempt to register the Domain Name, and the parties will jointly cooperate on any enforcement action of infringement of the Domain Name.
Task name: cuad_license_grant **Description**: Does the clause contain a license granted by one party to its counterparty? **Example**: Neoforma hereby grants VerticalNet a non-exclusive, non-transferable, royalty-free, right and license to link to the Neoforma Sites through a Neoforma Link.
Task name: cuad_liquidated_damages **Description**: Does the clause award either party liquidated damages for breach or a fee upon the termination of a contract (termination fee)? **Example**: You and each of your principals agree that the liquidated damages provision does not give us an adequate remedy at law for any default under, or for the enforcement of, any provision of this Agreement other than the Royalty Fee sections.
Task name: cuad_minimum_commitment **Description**: Does the clause specify a minimum order size or minimum amount or units pertime period that one party must buy from the counterparty? **Example**: If the Quarterly Average Sales Force Size is less than [***] Sales Representatives for an applicable Calendar Quarter, then in calculating the promotion fee due under Section 6.1.1, the Applicable Percentage for such Calendar Quarter shall be reduced to a new percentage equal to [***].
Task name: cuad_most_favored_nation **Description**: Does the clause state that if a third party gets better terms on the licensing or sale of technology/goods/services described in the contract, the buyer of such technology/goods/services under the contract shall be entitled to those better terms? **Example**: Eutectix agrees that in the event any Licensed Products shall be sold (1) to any Affiliate (as defined herein), or (2) to a corporation, firm, or association with which, or individual with whom Eutectix or its stockholders or Affiliates shall have any agreement, understanding, or arrangement (such as, among other things, an option to purchase stock, or an arrangement involving a division of profits or special rebates or allowances) without which agreement, understanding, or arrangement, prices paid by such a corporation, firm, association or individual for the Licensed Products would be higher than the Net Sales Price reported by Eutectix, or if such agreement, understanding, or arrangement results in extending to such corporation, firm, association, or individual lower prices for Licensed Products than those charged to outside concerns buying similar products in similar amounts and under similar conditions, then, and in any such events, the royalties to be paid hereunder in respect of such Licensed Products shall be computed based on an assumed or deemed Net Sales Price equal to those charged to such outside concerns.
Task name: cuad_no-solicit_of_customers **Description**: Does the clause restrict a party from contracting or soliciting customers or partners of the counterparty, whether during the contract or after the contract ends (or both)? **Example**: During the Term of this Agreement, and for a period of one year thereafter, except as expressly provided in this Agreement, PlanetCAD shall not market any services to Customers without the prior written approval of Dassault Systemes.
Task name: cuad_no-solicit_of_employees **Description**: Does the clause restrict a party's soliciting or hiring employees and/or contractors from the counterparty, whether during the contract or after the contract ends (or both)? **Example**: You covenant that during the term of this Agreement, except as otherwise approved in writing by us, you will not, either directly or indirectly, for yourself, or through, on behalf of, or in conjunction with any person, persons, partnership, corporation or company:<omitted>2. Employ or seek to employ any person who is at that time employed by us, our affiliates, or by any other franchisee of ours, or otherwise directly or indirectly induce or seek to induce such person to leave his or her employment thereat.

Table 20 – continued from previous page

Task
Task name: cuad_non-compete **Description**: Does the clause restrict the ability of a party to compete with the counterparty or operate in a certain geography or business or technology sector? **Example**: Agent may not offer or promote competitive products without the consent of Kallo.
Task name: cuad_non-disparagement **Description**: Does the clause require a party not to disparage the counterparty? **Example**: The Company shall not tarnish or bring into disrepute the reputation of or goodwill associated with the Seller Licensed Trademarks or Arizona.
Task name: cuad_non-transferable_license **Description**: Does the clause limit the ability of a party to transfer the license being granted to a third party? **Example**: Subject to the terms and conditions of this Agreement, Licensor hereby grants to Licensee, and Licensee hereby accepts from Licensor, a personal, non-exclusive, royalty-free right and license to use the Licensed Mark solely and exclusively as a component of Licensee's own corporate name and in connection with marketing the investment management, investment consultation and investment advisory services that Investment Advisor may provide to Licensee.
Task name: cuad_post-termination_services **Description**: Does the clause subject a party to obligations after the termination or expiration of a contract, including any post-termination transition, payment, transfer of IP, wind-down, last-buy, or similar commitments? **Example**: Cisco agrees to repurchase all Product in Distributor's inventory within [*****] days following the effective date of termination or expiration.
Task name: cuad_price_restrictions **Description**: Does the clause place a restriction on the ability of a party to raise or reduce prices of technology, goods, or services provided? **Example**: The prices set forth in Section 2.4(a) shall be subject to adjustment annually on the first day of each Product Year beginning in the calendar year 2000 and on the first day of each succeeding Product Year for the remainder of the Term and all renewals of this Agreement in proportion to the increase or decrease in the Consumer Price Index (CPI) as compared to the CPI as it existed on the first day of the Term of this Agreement.
Task name: cuad_revenue-profit_sharing **Description**: Does the clause require a party to share revenue or profit with the counterparty for any technology, goods, or services? **Example**: In consideration for the licenses granted to Corio pursuant to Section 2 (except Section 2.5) of this Agreement, Corio shall pay the revenue sharing fees specified in EXHIBIT B hereto.
Task name: cuad_rofr-rofo-rofn **Description**: Does the clause grant one party a right of first refusal, right of first offer or right of first negotiation to purchase, license, market, or distribute equity interest, technology, assets, products or services? **Example**: If Licensee shall have exercised such right, the closing shall be held at the corporate offices of Licensee on the closing date specified in the Offering Notice or the date that is ninety (90) days after the date of Licensee's notice of its exercise of such right, whichever is later.
Task name: cuad_source_code_escrow **Description**: Does the clause require one party to deposit its source code into escrow with a third party, which can be released to the counterparty upon the occurrence of certain events (bankruptcy, insolvency, etc.)? **Example**: With each delivery of Software to Bank of America hereunder, Supplier shall deliver to Bank of America the Source Code for all Software and for all Updates, Upgrades and new releases of the Software.
Task name: cuad_termination_for_convenience **Description**: Does the clause specify that one party can terminate this contract without cause (solely by giving a notice and allowing a waiting period to expire)? **Example**: Customer may terminate this Agreement during the Term upon at least one (1) years' written notice to M&I, provided that Customer pays M&I an early termination fee ("Termination for Convenience Fee") in an amount equal to REDACTED of the Estimated Remaining Value.
Task name: cuad_third_party_beneficiary **Description**: Does the clause specify that that there a non-contracting party who is a beneficiary to some or all of the clauses in the contract and therefore can enforce its rights against a contracting party? **Example**: Such covenants must be on a form that we provide, which form will, among other things, designate us as a third party beneficiary of such covenants with the independent right to enforce them.
Task name: cuad_uncapped_liability **Description**: Does the clause specify that a party's liability is uncapped upon the breach of its obligation in the contract? This also includes uncap liability for a particular type of breach such as IP infringement or breach of confidentiality obligation

Table 20 – continued from previous page

Task
Example: Subject to the foregoing as wen as Mobimagic's obligations under this Agreement, Mobimagic shall not in any manner be held or be responsible or liable for any unforeseen contingency, claims, liabilities, demands. losses, damages or expenses arising due to absence of storage or retention of any PC Financial data which shall be the sole responsibility of PC Financial .
Task name: cuad_unlimited-all-you-can-eat-license **Description**: Does the clause grant one party an "enterprise," "all you can eat" or unlimited usage license? **Example**: Subject to the terms and conditions of this Agreement, Commerce One hereby grants to Corio a fee-bearing, perpetual and irrevocable, nonexclusive, nontransferable (except in accordance with Section 14.1 of this Agreement), right and license in the Territory to<omitted>(iv) sublicense an unlimited number of Customers to access and use the Software and MarketSite.net Service only through the installation on Corio servers;
Task name: cuad_volume_restriction **Description**: Does the clause specify a fee increase or consent requirement, etc. if one party's use of the product/services exceeds certain threshold? **Example**: Make himself available for four (4) sessions for production of photographs, or radio, television, video or other multi-media programming for use in Bizzingo's advertising or promotional materials, with each such session not exceeding eight (8) hours.
Task name: cuad_effective_date **Description**: Does the clause specify the date upon which the agreement becomes effective? **Example**: This JV Agreement shall become effective on the signing date and shall have a duration of * years, extendable for a further * years, unless notice of non- renewal is sent one year before the natural expiry date.<omitted>2 April 2020
Task name: cuad_expiration_date **Description**: Does the clause specify the date upon which the initial term expires? **Example**: This Agreement shall be effective as of the Effective Date and shall continue in effect through December 31, 2021 and any Renewal Term (the "Term"), unless terminated earlier as set forth herein.
Task name: cuad_governing_law **Description**: Does the clause specify which state/country's law governs the contract? **Example**: This Agreement shall be governed by and construed under the laws of the State of California, excluding conflict of laws provisions and excluding the 1980 United Nations Convention on Contracts for the International Sale of Goods.
Task name: cuad_notice_period_to_terminate_renewal **Description**: Does the clause specify a notice period required to terminate renewal? **Example**: Unless either party gives written notice to terminate this Agreement at least six (6) months prior to the end of said Initial Term, this Agreement shall continue on a year to year basis ("Extended Term(s)") until terminated by either party by giving written notice of termination thereof to the other party at least six (6) months prior to the end of the then current Extended Term.
Task name: cuad_renewal_term **Description**: Does the clause specify a renewal term? **Example**: This Agreement shall commence on the Effective Date and, unless sooner terminated in accordance with its terms, including by Ginkgo pursuant to Section 7.3 (Buy-Down Election) or extended by the mutual written agreement of the Parties, shall continue until the Intended End of Term (such time period, as may be extended pursuant to this Section 13.3.1 (Term - General), the "Term"); provided that, if,<omitted>at the expiration of the Intended End of Term, Ginkgo has paid the Minimum Cumulative Purchase Commitment, but will not have paid to BLI the Full Purchase Target, then the Term of this Agreement shall automatically extend for an additional [***] ([***]) year period from the date of the expiration of the then-Intended End of Term so that, among other things, BLI may potentially receive the benefit of the Full Purchase Target and Ginkgo may receive the continuing benefit of royalty-free licenses.
Task name: cuad_warranty_duration **Description**: Does the clause specify a duration of any warranty against defects or errors in technology, products, or services provided under the contract? **Example**: Airspan warrants that, following repair or replacement, the repaired or replaced Equipment or Software by Airspan shall be free from defects in materials and faulty workmanship and that the Software will conform in all material respects to Airspan's published specifications therefor for ninety (90) days from date of shipment from Airspan to Distributor or until the end of the Initial Warranty Period, whichever is longer.

F.4.2 J.Crew Blocker

In LEGALBENCH, the J.Crew Blocker task is denoted as `jcrew-blocker`.

Background Loan agreements often contain restrictive covenants that place limits on a borrower's activities to protect the lender's interests. One such restrictive covenant that has become popular in recent years is the "J.Crew blocker" provision. This provision was created in response to actions taken by the retailer J.Crew in 2016. J.Crew transferred valuable intellectual property assets out of the collateral pool for its existing loans by moving them into a new unrestricted subsidiary. This subsidiary was then able to use the IP assets as collateral to obtain new financing.

The J.Crew blocker provision aims to prevent this type of activity by prohibiting borrowers from transferring IP assets out of the reach of existing lenders. There are two key components to a J.Crew blocker:

1. A prohibition on transferring IP assets to unrestricted subsidiaries. This prevents the borrower from moving assets outside the scope of lender restrictions.

2. A requirement to obtain lender consent for any IP transfers to subsidiaries. This gives lenders oversight and control over how IP assets are distributed within the corporate group.

The presence of a robust J.Crew blocker in a loan agreement is designed to keep material assets within the collateral pool, and thereby protect lenders from borrowers' attempts to secure additional debt through unexpected transfers of IP. For this reason, J.Crew blocker provisions have been widely adopted in leveraged loan agreements.

Task The J.Crew blocker task requires determining whether a given provision in a loan agreement qualifies as a J.Crew blocker. To make this determination, the provision must be analyzed to assess whether it contains:

1. A prohibition on transferring IP assets to unrestricted subsidiaries

AND/OR

2. A requirement to obtain lender consent for IP transfers to any subsidiary.

If the provision includes one or both of these components, it can be classified as a J.Crew blocker. If not, the provision does not meet the criteria.

Construction process The dataset for this task was constructed by legal experts extracting real examples of provisions from public loan agreements. Each example was labeled as either meeting the criteria for a J.Crew blocker or not. The dataset contains 60 total examples, organized into two columns: "Text" (containing the clause in question) and "Label" (indicating whether the clause is a J.Crew Blocker provision). Each clause was analyzed and classified as a J.Crew Blocker provision ("Yes") or not ("No"). The construction process involved manually reviewing and annotating these samples, ensuring that each clause was accurately categorized. This process, carried out by legal experts, provides definitive answers to each sample, eliminating ambiguity.

Significance and value The ability to identify J.Crew blocker provisions is important for both lenders and borrowers in leveraged finance. For lenders, it helps ensure key protections are included in loan agreements. For borrowers, it provides insight into restrictions being placed on their activities. Given the widespread adoption of J.Crew blockers, this is a task that requires proficiency to actively participate in the leveraged loan market. The task serves as an important measure of an LLM's ability to understand and apply legal concepts, particularly those related to secured lending and intellectual property law. It also tests the LLM's capacity to analyze and interpret legal provisions. Given the increasing complexity and sophistication of financial transactions, the ability to accurately identify and understand such provisions is a valuable skill for any LLM. This task, therefore, provides a useful measure of progress for LLMs in their understanding and interpretation of complex legal clauses.

Clause	J.Crew Blocker Provision?
provided that (i) immediately before and after such designation, no Event of Default shall have occurred and be continuing, (ii) in the case of the designation of any Subsidiary as an Unrestricted Subsidiary, such designation shall constitute an Investment in such Unrestricted Subsidiary (calculated as an amount equal to the sum of (x) the fair market value of the Equity Interests of the designated Subsidiary and any of its Subsidiaries that are owned by Holdings or any Restricted Subsidiary, immediately prior to such designation (such fair market value to be calculated without regard to any Obligations of such designated Subsidiary or any of its Subsidiaries under the Guaranty Agreement) and (y) the aggregate principal amount of any Indebtedness owed by such Subsidiary and any of its Subsidiaries to Holdings or any of the Restricted Subsidiaries immediately prior to such designation, all calculated, except as set forth in the parenthetical to clause (x) above, on a consolidated basis in accordance with U.S. GAAP), and such Investment shall be permitted under Section 10.05, (iii) no Subsidiary may be designated as an Unrestricted Subsidiary if it or any of its Subsidiaries is a Restricted Subsidiary for the purpose of any Refinancing Notes Indenture, any Permitted Pari Passu Notes Document, any Permitted Pari Passu Loan Documents, any Permitted Junior Notes Document or other debt instrument, with a principal amount in excess of the Threshold Amount, (iv) following the designation of an Unrestricted Subsidiary as a Restricted Subsidiary, Holdings shall comply with the provisions of Section 9.12 with respect to such designated Restricted Subsidiary, (v) no Restricted Subsidiary may be a Subsidiary of an Unrestricted Subsidiary (and any Subsidiary of an Unrestricted Subsidiary that is acquired or formed after the date of designation shall automatically be designated as an Unrestricted Subsidiary) and (vi) in the case of the designation of any Subsidiary as an Unrestricted Subsidiary, each of (x) the Subsidiary to be so designated and (y) its Subsidiaries has not, at the time of designation, and does not thereafter, create, incur, issue, assume, guarantee or otherwise become directly or indirectly liable with respect to any Indebtedness pursuant to which the lender has recourse to any of the assets of Holdings or any Restricted Subsidiary (other than Equity Interests in an Unrestricted Subsidiary).	No
provided, that (i) immediately before and after such designation, no Event of Default exists (including after giving effect to the reclassification of Investments in, Indebtedness of and Liens on the assets of, the applicable Restricted Subsidiary or Unrestricted Subsidiary), (ii) as of the date of the designation thereof, no Unrestricted Subsidiary shall own any Capital Stock in any Restricted Subsidiary of the Borrower or hold any Indebtedness of or any Lien on any property of the Borrower or its Restricted Subsidiaries and (iii) no subsidiary may be designated as an Unrestricted Subsidiary if it owns intellectual property that is material to the business of the Borrower and its Restricted Subsidiaries, taken as a whole (such intellectual property, Material Intellectual Property), at the time of designation, other than in connection with transactions that have a bona fide business purpose so long as such transactions are not undertaken to facilitate a financing or a Restricted Payment or undertaken in connection with a liability management transaction. Notwithstanding anything contained in this Section 6.05 to the contrary, in no event shall (a) the Borrower or any Restricted Subsidiary be permitted to make or own any Investment in the Holdings direct or indirect equityholders constituting Material Intellectual Property (other than pursuant to a bona fide transition service or similar arrangement or in the same manner as other customers, suppliers or commercial partners of the relevant transferee generally) or (b) any Restricted Subsidiary transfer ownership of, or license on an exclusive basis, any Material Intellectual Property to any Unrestricted Subsidiary, other than in connection with transactions that have a bona fide business purpose and so long as such transactions are not undertaken to facilitate a financing or a Restricted Payment or undertaken in connection with a liability management transaction. Notwithstanding anything contained in this Section 6.06 to the contrary, in no event shall (a) the Borrower or any Restricted Subsidiary be permitted to make any Disposition of Material Intellectual Property to Holdings direct or indirect equityholders (other than pursuant to a bona fide transition service or similar arrangement or in the same manner as other customers, suppliers or commercial partners of the relevant transferee generally) or (b) any Restricted Subsidiary make any Disposition, constituting either a transfer of ownership or an exclusive license, of any Material Intellectual Property to any Unrestricted Subsidiary, other than in connection with transactions that have a bona fide business purpose and so long as such transactions are not undertaken to facilitate a financing or a Restricted Payment or undertaken in connection with a liability management transaction.	Yes

Table 21: Examples from `jcrew_blocker`.

F.4.3 Unfair Terms of Service

In LEGALBENCH, the Unfair Terms of Service task is denoted as `unfair_tos`.

Background An array of recent work has found that consumers rarely read terms of service agreements [97, 64]. As a result, consumers regularly sign agreements or contracts containing provisions that (1) they lack awareness of, and/or (2) would consider as "unfair" or "predatory." Reasons for this phenomenon include the sheer amount of time it would take to read every terms of service agreement, the obtuse language of these agreements, and the lack of actual recourse on an individual basis.

With reference to European consumer law, [82] identify eight categories of clauses in terms-of-service agreements which could be considered "potentially unfair":

- Arbirtration: clauses which mandated that all disputes between the parties would be resolved through arbitration.
- Unilateral change: clauses which allow the provider to modify the terms of service and/or the service itself.
- Content removal: clauses which give the provider a right to modify/delete a user's content
- Jurisdiction: clauses which specify a jurisdiction in which claims must be brought, regardless of where the user lives.
- Choice of law: clauses which specify the country's law which governs disputes arising under the contract, regardless of where the user lives.
- Limitation of liability: clauses which limit the liability of the service provider.
- Unilateral termination: clauses which empower the service provider to terminate/suspend the service at their discretion.
- Contract by using: clauses which stipulate that a consumer is bound by the terms of service simply by using the service.

A more detailed description of these categories can be found in [82].

Task The Unfair Terms of Service task requires an LLM to determine—given a clause from a terms of service agreeement—whether it belongs to one of the above eight categories, and if so, which one.

Construction process We use the version of data available in [21], which takes a subset from [82]. Unlike [21]—which frames the task as distinguishing "fair" from "unfair" clauses—we cast the task as 8-way multiclassification task across the original categories identified in [82].

Significance and value Unlike the CUAD and J.Crew Blocker task, the Unfair TOS task evaluates a LLM's ability to perform multiclass clause classification across a highly imbalanced dataset.

Clause	Clause type
you also acknowledge that a variety of evernote actions may impair or prevent you from accessing your content or using the service at certain times and/or in the same way , for limited periods or permanently , and agree that evernote has no responsibility or liability as a result of any such actions or results , including , without limitation , for the deletion of , or failure to make available to you , any content .	Arbitration
if you do not terminate your agreement before the date the revised terms become effective , your continued access to or use of the airbnb platform will constitute acceptance of the revised terms .	Choice of law
we may at any time and from time to time , in our sole discretion , change the fees and charges , or add new fees and charges , in relation to any of the products .	Content removal
you and academia.edu agree that any dispute , claim or controversy arising out of or relating to these terms or the breach , termination , enforcement , interpretation or validity thereof or the use of the site or services (collectively , “ disputes ”) will be settled by binding arbitration , except that each party retains the right : (i) to bring an individual action in small claims court and (ii) to seek injunctive or other equitable relief in a court of competent jurisdiction to prevent the actual or threatened infringement , misappropriation or violation of a party ’s copyrights , trademarks , trade secrets , patents or other intellectual property rights (the action described in the foregoing clause (ii) , an “ ip protection action ”) .	Contract by using
if we find that any shared content in your account violates our terms of service (including by violating another person ’s intellectual property or privacy rights) , we reserve the right to un-share or take down such content .	Jurisdiction
if you live in the european union : you agree that the laws of ireland , excluding conflict of laws rules , shall exclusively govern any dispute relating to this contract and/or the services .	Limitation of liability
oculus does not endorse or guarantee the opinions , views , advice or recommendations posted or sent by users .	Other
if you object to the changes , nintendo reserves the right to terminate this agreement or any portion of it upon reasonable notice and you will have to register again if you wish to continue using the nintendo account service under the new terms and conditions .	Unilateral change
unless you and we agree otherwise , in the event that the agreement to arbitrate above is found not to apply to you or to a particular claim or dispute , either as a result of your decision to opt out of the agreement to arbitrate or as a result of a decision by the arbitrator or a court order , you agree that any claim or dispute that has arisen or may arise between you and ebay must be resolved exclusively by a state or federal court located in salt lake county , utah .	Unilateral termination

Table 22: Examples from `unfair_tos`.

F.4.4 Contract QA

In LEGALBENCH, the Contract QA task is denoted as `contract_qa`.

Background Each of the above tasks evaluates the capacity for LLMs to learn to recognize a single type of clause, given a description of that clause and/or examples of it. The Contract QA task generalizes this across multiple clause types, evaluating an LLM's ability to recognize legal provisions that are *not* described in the prompt.

Task Each sample in the dataset consists of (1) a contract clause, and (2) a question asking if the clause is an example of a provision type (e.g., “Is this a severability clause?”). Across the dataset, the questions correspond to 22 different legal provisions. Questions and provisions are paired such that for each provision type, the LLM is presented with two clauses that are an example of the type, and two clauses which are not.

Class	**Number of samples**
Other	3454
Contract by using	15
Choice of law	38
Content removal	53
Unilateral change	70
Arbitration	98
Limitation of liability	28
Unilateral termination	32
Jurisdiction	25

Table 23: Test set class distribution.

Construction Process The data was manually extracted from a set of sample agreements contributed by a LegalTech vendor and from public sources. It represents a variety of contracts, such as:

- Vendor or Partner Data Protection Agreements (DPA)
- Master Services Agreements (MSA)
- Licensing Terms
- BIPA consents

Clause	Question	Answer
This Agreement shall be governed by and construed in accordance with the laws of the State of New York, without giving effect to any choice of law or conflict of law provisions.	Does the clause discuss BIPA consent?	No
If a dispute arises between the parties under this Agreement that cannot be resolved through good faith negotiations within a reasonable period of time, such dispute shall be escalated to an executive officer of each party for resolution. If such executive officers are unable to resolve such dispute within a reasonable period of time after escalation, either party may pursue any available legal remedies.	Does the clause discuss how disputes may be escalated?	Yes

Table 24: Examples for `contract_qa`.

F.5 Consumer Contracts QA

In LEGALBENCH, the Consumer Contracts QA task is denoted as `consumer_contracts_qa`.

Background Consumer contracts govern many economic and social activities, ranging from retail purchases and online search to social media and entertainment. These contracts can affect consumers' access to services, control terms of payment, and determine the remedies available when consumers' rights are violated. Despite the importance of these legal agreements, consumers typically lack the time, expertise, and incentive to properly examine how consumer contracts impact their rights and interests. This issue is known as the "no-reading" problem [8, 7]. LLMs may offer a solution. By reading consumer contracts and explaining their legal ramifications, LLMs could enable consumers to better understand and exercise their legal rights in many everyday contexts.

Task The Consumer Contracts QA task, first introduced in [73], aims to examine the degree to which an LLM can understand certain consumer contracts. Specifically, the task is comprised of 200 yes/no legal questions relating to the terms of service of popular websites. Examples of questions are provided in the table below.

Contract: Content Removal and Disabling or Terminating Your Account We can remove any content or information you share on the Service if we believe that it violates these Terms of Use, our policies (including our Instagram Community Guidelines), or we are permitted or required to do so by law. We can refuse to provide or stop providing all or part of the Service to you (including terminating or disabling your your access to the Facebook Products and Facebook Company Products) immediately to protect our community or services, or if you create risk or legal exposure for us, violate these Terms of Use or our policies (including our Instagram Community Guidelines), if you repeatedly infringe other people's intellectual property rights, or where we are permitted or required to do so by law. We can also terminate or change the Service, remove or block content or information shared on our Service, or stop providing all or part of the Service if we determine that doing so is reasonably necessary to avoid or mitigate adverse legal or regulatory impacts on us. If you believe your account has been terminated in error, or you want to disable or permanently delete your account, consult our Help Center.When you request to delete content or your account, the deletion process will automatically begin no more than 30 days after your request. It may take up to 90 days to delete content after the deletion process begins. While the deletion process for such content is being undertaken, the content is no longer visible to other users, but remains subject to these Terms of Use and our Data Policy. After the content is deleted, it may take us up to another 90 days to remove it from backups and disaster recovery systems. Content will not be deleted within 90 days of the account deletion or content deletion process beginning in the following situations: where your content has been used by others in accordance with this license and they have not deleted it (in which case this license will continue to apply until that content is deleted); or where deletion within 90 days is not possible due to technical limitations of our systems, in which case, we will complete the deletion as soon as technically feasible; or where deletion would restrict our ability to: investigate or identify illegal activity or violations of our terms and policies (for example, to identify or investigate misuse of our products or systems); protect the safety and security of our products, systems, and users; comply with a legal obligation, such as the preservation of evidence; or comply with a request of a judicial or administrative authority, law enforcement or a government agency; in which case, the content will be retained for no longer than is necessary for the purposes for which it has been retained (the exact duration will vary on a case-by-case basis). If you delete or we disable your account, these Terms shall terminate as an agreement between you and us, but this section and the section below called "Our Agreement and What Happens if We Disagree" will still apply even after your account is terminated, disabled, or deleted. **Question**: According to the terms, 30 days after Ive asked to delete content, can other users see that content? **Answer**: No
Contract: 16. Termination You may terminate these Terms at any time and for any reason by deleting your Account and discontinuing use of all Services. If you stop using the Services without deactivating your Account, your Account may be deactivated due to prolonged inactivity. We may suspend or terminate your Account, moderator status, or ability to access or use the Services at any time for any or no reason, including for violating these Terms or our Content Policy. The following sections will survive any termination of these Terms or of your Account: 4 (Your Content), 6 (Things You Cannot Do), 10 (Indemnity), 11 (Disclaimers), 12 (Limitation of Liability), 13 (Governing Law and Venue), 16 (Termination), and 17 (Miscellaneous). 17. Miscellaneous These Terms constitute the entire agreement between you and us regarding your access to and use of the Services. Our failure to exercise or enforce any right or provision of these Terms will not operate as a waiver of such right or provision. If any provision of these Terms is, for any reason, held to be illegal, invalid, or unenforceable, the rest of the Terms will remain in effect. You may not assign or transfer any of your rights or obligations under these Terms without our consent. We may freely assign any of our rights and obligations under these Terms. **Question**: Will certain terms remain in force notwithstanding a users termination of the service? **Answer**: Yes

Table 25: Task examples.

In addition to the original 200 questions, the task includes an alternatively worded version of all 200 questions. While each question's content is substantially the same across both versions of the question, the alternatively worded questions are, by design, less readable, that is, more difficult for a human to read. Comparing performance across the original questions and the alternatively worded questions can help assess an LLM's brittleness in performing the task at hand. An example is provided in the table below:

Original wording	**Alternative wording**
Am I allowed to be paid for writing a Wikipedia article, assuming I disclose who's paying me?	Are Wikipedia contributors permitted to receive payment in respect of their contributions, provided they disclose the identity of the person or institution providing such payment?

Table 26: Example of reworded question.

Construction process The task was introduced in [73]. To construct the dataset, an attorney drafted 200 yes/no questions relating to the terms of service of the 20 most-visited U.S. websites (10 questions per document), as well as an alternatively worded version of all 200 questions. The questions relate to a wide range of legal issues arising in the terms of service, including eligibility to access services, payment for services, limitations of liability, intellectual property rights, and dispute resolution procedures. Answers to all questions can be obtained from the applicable terms of service.

Significance and value Given the ubiquity of consumer contracts, LLMs capable of reading these documents and communicating their contents to consumers might offer significant benefits. These benefits, however, are contingent on a model's accuracy and reliability. LLMs that misinterpret the provisions of consumer contracts may hinder consumers' ability to understand and exercise their contractual rights. The Consumer Contracts QA task is a preliminary attempt at evaluating the ability of LLMs to read certain consumer contracts.

F.6 Contract NLI Tasks

In LEGALBENCH, the Contract NLI tasks are denoted as `contract_nli_*`.

Task The Contract NLI tasks require a LLM—given an excerpt of a contract and an assertion about the legal effect of that excerpt—to determine whether the assertion is supported or unsupported by the excerpt.

Construction process These tasks are constructed by transforming data released by [74]. The original dataset consists of 607 contracts and 17 assertions (e.g., "Receiving Party shall not disclose the fact that Agreement was agreed or negotiated "). Each contract is labeled for each assertion as supporting, negating, or not mentioning the assertion. Please refer to the original paper for details on annotation.

We restructure this dataset for a short-context LLM setting. Specifically, we treat each assertion as a separate task, where the objective is to determine whether a contract excerpt is supportive (or not) of the assertion. For each instance where a contract is supportive of an assertion, [74] has annotated the excerpt of the contract that is supportive. When creating a task, we use the supportive excerpts for the assertion from the test set as positive instances. To generate negative instances, we combine excerpts where the assertion is contradicted with a random sample of excerpts associated with other assertions. We treat both groups of excerpts as instances which are "unsupportive" of the assertion. We transform the assertion into a Yes/No question, where the LLM is asked to determine if a clause satisfies the assertion.

Table 27 lists each task, the assertion associated with the task, and an example of an excerpt which supports the assertion.

Table 27: ContractNLI Tasks

Task
Task name: contract_nli_return_of_confidential_information **Question**: Identify if the clause provides that the Receiving Party shall destroy or return some Confidential Information upon the termination of Agreement. **Example**: Upon receipt by the Recipient of a written demand from the Disclosers: 8.1.1 the Recipient must return or procure the return to the Disclosers or, as the Disclosers may require, destroy or procure the destruction of any and all materials containing the Confidential Information together with all copies; 8.1.2 if the Disclosers requires, the Recipient must provide the Disclosers with a certificate or such other evidence as the Disclosers may reasonably require duly signed or executed by an officer of the Recipient confirming that the Recipient has complied with all of its obligations under this Agreement including about return, destruction and deletion of Confidential Information and media; 8.1.3 the Recipient must delete or procure the deletion of all electronic copies of Confidential Information; and 8.1.4 the Recipient must make, and procure that the Authorised Persons shall make, no further Use of the Confidential Information.
Task name: contract_nli_no_licensing **Question**: Identify if the clause provides that the Agreement shall not grant Receiving Party any right to Confidential Information. **Example**: No license to the receiving party under any trade secrets or patents or otherwise with respect to any of the Proprietary Information is granted or implied by conveying proprietary Information or other information to such party, and none of the information transmitted or exchanged shall constitute any representation, warranty, assurance, guaranty or inducement with respect to the infringement of patents or other rights of others.
Task name: contract_nli_confidentiality_of_agreement **Question**: Identify if the clause provides that the Receiving Party shall not disclose the fact that Agreement was agreed or negotiated. **Example**: In addition, except as permitted herein, Recipient shall not disclose the fact that the parties are exchanging Confidential Information and having discussions. In connection therewith, it is agreed that no public release or disclosure of any contemplated transaction shall be made except by a mutually agreed disclosure except that each party may make such disclosure if advised by its outside securities counsel in writing that such disclosure is required; PROVIDED, HOWEVER, that in such event such party will notify the other party that it intends, as a preliminary matter, to take such action and the outside securities counsel of such party shall first discuss the mater with the outside securities counsel of the other party before any definitive decision is made on the disclosure.
Task name: contract_nli_explicit_identification **Question**: Identify if the clause provides that all Confidential Information shall be expressly identified by the Disclosing Party. **Example**: 1. As used herein, the term "Proprietary Information" refers to any and all Information of a confidential, proprietary, or secret nature which is applicable to or related In any way to (i) the business, present or future, of the Disclosing Party, (ii) the research and development or investigations of the Disclosing Party or (iii) the business of any customer of the Disclosing Party; provided, in each case, that such information is delivered to the Receiving Party by the Disclosing Party and (a) is marked or identified in writing as "Confidential", (b) if verbal or visual disclosure, is identified as "Confidential" in a writing within ten (10) business days of such disclosure, or
Task name: contract_nli_survival_of_obligations **Question**: Identify if the clause provides that ome obligations of Agreement may survive termination of Agreement.

Table 27 – continued from previous page

Task
Example: b. This Agreement shall be valid when signed by duly authorised representatives of the Parties and shall be binding on each Party for 10 (ten) years as from the date of signature of the last signatory, even if at the end of the negotiations a data sharing agreement is not signed between the Parties, or until such time as the Information enters into the public domain.
Task name: contract_nli_permissible_development_of_similar_information **Question**: Identify if the clause provides that the Receiving Party may independently develop information similar to Confidential Information. **Example**: "Confidential Information" of a disclosing party ("Discloser") means the following, regardless of its form and including copies made by the receiving party ("Recipient"), whether the Recipient becomes aware of it before or after the date of this Agreement: except where that information is: Independently developed by the Recipient without use, directly or indirectly of Confidential Information received from the Discloser.
Task name: contract_nli_permissible_post-agreement_possession **Question**: Identify if the clause provides that the Receiving Party may retain some Confidential Information even after the return or destruction of Confidential Information. **Example**: 9. Upon the Disclosing Party's written request, the Receiving Party shall (at the Receiving Party's election) promptly return or destroy (provided that any such destruction shall be certified by a duly authorized Representative of the Receiving Party) all Confidential Information of the Disclosing Party and all copies, reproductions, summaries, analyses or extracts thereof or based thereon (whether in hard-copy form or an intangible media, such as electronic mail or computer files) in the Receiving Party's possession or in the possession of any Representative of the Receiving Party; provided, however: (i) that if a legal proceeding has been instituted to seek disclosure of the Confidential Information, such material shall not be destroyed until the proceeding is settled or a final judgment with respect thereto has been rendered; (ii) that the Receiving Party shall not, in connection with the foregoing obligations, be required to identify or delete Confidential Information held electronically in archive or back-up systems in accordance with general systems archiving or backup policies; and (iii) that the Receiving Party shall not be obligated to return or destroy Confidential Information of the Disclosing Party to the extent the Receiving Party is required to retain a copy pursuant to applicable law, and further provided that the Receiving Party will not, and the Receiving Party will use reasonable measures to cause its employees not to, access such Confidential Information so archived or backed-up.
Task name: contract_nli_inclusion_of_verbally_conveyed_information **Question**: Identify if the clause provides that Confidential Information may include verbally conveyed information. **Example**: I acknowledge that The Business Partnership has provided, and/or has agreed to provide in the future, to me information of a confidential or proprietary nature (the Confidential Information) Confidential Information shall mean any information or data relating to any clients of The Business Partnership business or affairs disclosed whether in writing, orally or by any other means.
Task name: contract_nli_sharing_with_third-parties **Question**: Identify if the clause provides that the Receiving Party may share some Confidential Information with some third-parties (including consultants, agents and professional advisors). **Example**: Receiving Party shall carefully restrict access to Sensitive Information to employees, contractors and third parties as is reasonably required and shall require those persons to sign nondisclosure restrictions at least as protective as those in this Agreement.
Task name: contract_nli_permissible_copy **Question**: Identify if the clause provides that the Receiving Party may create a copy of some Confidential Information in some circumstances. **Example**: If any party makes copies of the Confidential Information of the other party, such copies shall also constitute Confidential Information and any and all confidential markings on such documents shall be maintained.
Task name: contract_nli_notice_on_compelled_disclosure **Question**: Identify if the clause provides that the Receiving Party shall notify Disclosing Party in case Receiving Party is required by law, regulation or judicial process to disclose any Confidential Information.

Table 27 – continued from previous page

Task
Example: If the Receiving Party or its Representatives are requested or required in any judicial, arbitral or administrative proceeding or by any governmental or regulatory authority to disclose any Evaluation Material (whether by deposition, interrogatory, request for documents, subpoena, civil investigative demand, or otherwise), or the Receiving Party is so requested or required to disclose any of the facts disclosure of which is prohibited under paragraph (3)(e) of this Agreement, the Receiving Party shall give the Furnishing Party prompt notice of such request so that the Furnishing Party may seek an appropriate protective order or other appropriate remedy and/or waive compliance with the provisions of this Agreement, and, upon the Furnishing Party's request and at the Furnishing Party's expense, shall reasonably cooperate with the Furnishing Party in seeking such an order. (d) Notice If either Party proposes to make any disclosure in reliance on clause (i) above, the disclosing Party shall, to the extent practicable, provide the other Party with the text of the proposed disclosure as far in advance of its disclosure as is practicable and shall in good faith consult with and consider the suggestions of the other Party concerning the nature and scope of the information it proposes to disclose. Notwithstanding the foregoing, a Party may make such public announcement or public statement if in the opinion of such Party's outside counsel or General Counsel, such public announcement or public statement is necessary to avoid committing a violation of law or of any rule or regulation of any securities association, stock exchange or national securities quotation system on which such Party's securities are listed or trade. In such event, the disclosing Party shall use its reasonable best efforts to give advance notice to the other Party and to consult with the other Party on the timing and content of any such public announcement or public statement.
Task name: contract_nli_permissible_acquirement_of_similar_information **Question**: Identify if the clause provides that the Receiving Party may acquire information similar to Confidential Information from a third party. **Example**: For the purposes of this Agreement, the term "Confidential Information" shall mean all trade secrets and confidential or proprietary information (and any tangible representation thereof) owned, possessed or used in connection with The Company Business or by the Buyer Parties and its Affiliates; provided, however, that "Confidential Information" does not include information which is or becomes generally available to the public other than as a result of a disclosure by a Seller Party..
Task name: contract_nli_sharing_with_employees **Question**: Identify if the clause provides that the Receiving Party may share some Confidential Information with some of Receiving Party's employees. **Example**: We and our representatives will keep the Evaluation Materials completely confidential; provided, however, that (i) any of such information may be disclosed to those of our directors, officers, employees, agents, representatives (including attorneys, accountants and financial advisors), lenders and other sources of financing (collectively, "our representatives") who we reasonably determine need to know such information for the purpose of evaluating a Possible Transaction between us and the Company (it being understood that our representatives shall be informed by us of the confidential nature of such information and shall be directed by us, and shall each agree to treat such information confidentially) and
Task name: contract_nli_limited_use **Question**: Identify if the clause provides that the Receiving Party shall not use any Confidential Information for any purpose other than the purposes stated in Agreement. **Example**: 2.1. A Receiving Party agrees: 2.1.2. to use the Confidential Information of the other solely in, and to the extent necessary for the Purpose and not to copy or use any Confidential Information of the other save to the extent necessary for the Purpose;

Significance and value The Contract NLI tasks evaluate an LLM's capacity to reason over the rights and obligations created by a contract. The ability to perform this skill is essential to many types of legal work.

F.7 Corporate Lobbying

In LEGALBENCH, the Corporate Lobbying task is denoted as `corporate_lobbying`.

Background A significant amount of effort is devoted to identifying developing sources of law which implicate client or issue interests. Examples of such sources include: legislative bills, proposed regulations, or in-progress litigation. Identifying these sources serves multiple purposes. From a scholarly standpoint, researchers often aggregate sources into issue-focused databases, enabling them to identify emerging trends or patterns across different sources [37]. From an advocacy standpoint, identifying sources allows affected groups to better understand how their rights or obligations may be affected, and how to focus efforts on interacting with courts, legislatures, and other governmental bodies [15, 83, 44].

Task The Corporate Lobbying task requires an LLM to determine whether a proposed Congressional bill may be relevant to a company based on a company's self-description in its SEC 10K filing. The following information about a bill and a company are available:

- The title of the bill.
- A summary of the bill.
- The name of the company.
- A description of the company.

We expect higher accuracy of LLM predictions if we were to provide the model with more data about a bill, and especially if we provide it with more data about a company. Proprietary applications of this approach could leverage significant internal company data. More expensive deployments could leverage the full text of the bill

Construction process This data was manually labeled. This work was an extension of the research described in [92].

Significance and value Determining whether a particular bill is relevant for a company requires (1) identifying the legal consequences of the bill, and (2) whether those consequences are relevant to a company's business model, structure, or activities. As discuss above, this type of prognostication is a common legal practice. For instance, law firms regularly publish "client alerts" which seek to keep clients updated on new legal developments [122].

Class	Number of samples
No	345
Yes	145

Table 28: Test set class distribution

Field	Text
Bill Title	A bill to provide standards relating to airline travel by Federal employees for official business.
Bill Summary	Fly Smart Act This bill establishes standards for airline travel by federal employees for official business, including a general requirement to use coach-class accommodations and a ban on military aircraft for domestic official travel. It allows use of first-class and business class for federal employees under certain circumstances, such as to accommodate a disability or special need or because of exceptional security circumstances
Company Name	Alaska Air Group, Inc.
Company Description	Virgin America has been a member of Air Group since it was acquired in 2016. In 2018, Virgin America and Alaska combined operating certificates to become a single airline, and legally merged into a single entity. The Company also includes McGee Air Services, an aviation services provider that was established as a wholly-owned subsidiary of Alaska in 2016. Together with our regional partner airlines, we fly to 115 destinations with over 1,200 daily departures through an expansive network across the United States, Mexico, Canada, and Costa Rica. With global airline partners, we provide our guests with a network of more than 900 destinations worldwide. Our adjusted net income was $554 million, which excludes merger-related costs, special items and mark-to-market fuel hedge adjustments. Refer to "Results of Operations" in Management's Discussion and Analysis for our reconciliation of Non-GAAP measures to the most directly comparable GAAP measure. Mainline - includes scheduled air transportation on Alaska's Boeing or Airbus jet aircraft for passengers and cargo throughout the U.S., and in parts of Canada, Mexico, and Costa Rica. other third-party carriers' scheduled air transportation for passengers across a shorter distance network within the U.S. under capacity purchase agreements (CPA). Horizon - includes the capacity sold to Alaska under CPA. Expenses include those typically borne by regional airlines such as crew costs, ownership costs and maintenance costs. We believe our success depends on our ability to provide safe air transportation, develop relationships with guests by providing exceptional customer service and low fares, and maintain a low cost structure to compete effectively. In 2018 , we focused much of our energy on the integration of Virgin America, completing over 95% of our integration milestones. In January 2018, Alaska and Virgin America received a Single Operating Certificate (SOC) from the Federal Aviation Administration (FAA), which recognizes Alaska and Virgin America as one airline. In April 2018, we transitioned to a single Passenger Service System (PSS), which allows us to provide one reservation system, one website and one inventory of flights to our guests. This transition to a single PSS enables us to unlock many of the revenue synergies expected from the acquisition, and to provide consistent branding to our guests at all airport gates, ticketing, and check-in areas. The two most important milestones we have yet to complete include combining the maintenance operations of Boeing and Airbus, and reconfiguring our Airbus fleet. In 2018 , we painted 33 Airbus aircraft with the Alaska livery and we are in process of reconfiguring all Airbus aircraft to achieve a cabin experience for our guests that is consistent with our Boeing fleet. In early 2019, we will also complete the integration of our crew management systems and aim to reach a collective bargaining agreement with our aircraft technicians, the last remaining labor group that has not yet reached a joint collective bargaining agreement. With the integration largely behind us, we remain committed to our vision to become the favorite airline for people on the West Coast. The acquisition of Virgin America positioned us as the fifth largest airline in the U.S., with an unparalleled ability to serve West Coast travelers. ' evolving needs by offering a relevant network and schedule, upgrading our onboard offerings, and retaining our unique West Coast vibe. Some of the more notable product enhancements underway include adding high-speed satellite connectivity to our entire Boeing and Airbus fleets, updating and expanding our airport lounges, and working with the Port of Seattle to open a state-of-the-art 20-gate North Satellite Concourse 4 at Sea-Tac Airport, including a 15,000 square-foot flagship lounge. We have also introduced new food and beverage menus, which include more fresh, local, and healthy offerings including salads, protein plates, and fresh snacks, as well as new beverage offerings, including craft beers, juices and an updated wine selection. We are also active in the communities we serve and strive to be an industry leader in environmental and community stewardship.

Table 29: An example of a relevant bill for `corporate_lobbying`.

F.8 Definition Tasks

In LEGALBENCH, the Definition Tasks are denoted as `definition_classification` and `definition_extraction`.

Background Judicial opinions regularly involve *definition*, assigning a particular meaning to words or phrases (Let us define words and phrases as "terms"). Definition of terms can occur when judges introduce or discuss legal concepts (e.g. *parol evidence*), and it frequently occurs when judges interpret terms in legal texts. This can include language from past judicial opinions and language appearing in legal texts like contracts, statutes, and the Constitution. Historically, interpreters have often evaluated the definition(s) of individual words. For example, in interpreting the meaning of "keep and bear arms" in the Second Amendment, courts consider the definition(s) of individual words (like "bear"). This approach—focusing on terms' definitions—has only increased in recent decades with the rise of textualist approaches to constitutional and statutory interpretation.

Judicial opinions define a wide range of terms, including ordinary terms, legal terms, and scientific terms. They also appeal to a wide range of defining sources, including ordinary dictionaries, legal dictionaries, and legal texts. For an example of the last, consider statutory definitions: 1 U.S.C. 1 offers generally applicable definitions of many frequent statutory terms.

It is useful for lawyers to identify *when definition occurs* (definition classification), as well as *which terms* have been defined (definition extraction). These tasks might seem simple at first. There are some intuitively plausible indicators of definition classification and extraction. For example, defined terms often (but not always) appear in quotation marks or near a citation to a dictionary.

However, these tasks are not entirely straightforward. Indicators like quotation will not lead to perfect definition classification and extraction. Consider for example, this sentence from the dataset related to the definition of "confidential": *The term "confidential" meant then, as it does now, "private" or "secret." Webster's Seventh New Collegiate Dictionary 174 (1963).*[18] As another example from the dataset, consider this definition of "brought": *But a natural reading of § 27's text does not extend so far. "Brought" in this context means "commenced," Black's Law Dictionary 254 (3d ed. 1933).*[19] Other examples exclusively quote the definition, rather than defined terms: *Stare decisis ("to stand by things decided") is the legal term for fidelity to precedent. Black's Law Dictionary 1696 (11th ed. 2019).*[20] In all of these examples, the presence of a dictionary would not indicate which term is extracted. In other examples, there is no dictionary cited; there is not a perfect correlation between dictionary citation and classification of a sentence as a defining one.[21]

Tasks The Definition Classification task requires an LLM to determine–given an excerpt from a Supreme Court opinion–whether the excerpt is defining any term (Yes/No). The Definition Extraction task requires an LLM to determine–given an excerpt from a Supreme Court opinion–which term the excerpt is defining (Open-ended response).

Sentence	Definition sentence?
The risk of that consequence ought to tell us that something is very wrong with the Court's analysis.	No
This term has long referred to a class of expenses commonly recovered in litigation to which attorney's fees did not traditionally belong. See Black's Law Dictionary 461 (1891) (defining "expensae litis" to mean "generally allowed" costs); 1 J. Bouvier, Law Dictionary 392 (1839) (defining the term to mean the "costs which are generally allowed to the successful party"); id., at 244 (excluding from the definition of "costs" the "extraordinary fees [a party] may have paid counsel").	Yes

Table 30: Examples for `definition_classification`.

Construction process An original hand-coded dataset was constructed to study how the Supreme Court relies on dictionaries over time. Any case citing a dictionary was included in the dataset, and human coders identified relevant excerpts that defined terms and *which* terms were defined.

That dataset has been repurposed for the task here. For the definition extraction task, the original dataset includes the relevant information (excerpts, with the defined term coded separately).

[18] Food Mktg. Inst. v. Argus Leader Media, 139 S. Ct. 2356, 2363 (2019).

[19] Merrill Lynch, Pierce, Fenner & Smith Inc. v. Manning, 136 S. Ct. 1562, 1568 (2016).

[20] June Medical Services L.L.C. v. Russo, 140 S. Ct. 2103, 2134 (2020).

[21] E.g. "And "remuneration" means "a quid pro quo," "recompense" or "reward" for such services. Id., at 1528." BNSF Ry. Co. v. Loos, 139 S. Ct. 893, 905 (2019).

Sentence	Defined term
The term "plaintiff" is among the most commonly understood of legal terms of art: It means a "party who brings a civil suit in a court of law." Black's Law Dictionary 1267 (9th ed. 2009) see also Webster's Third New International Dictionary 1729 (1961)"	plaintiff
The ordinary understanding of law enforcement includes not just the investigation and prosecution of offenses that have already been committed, but also proactive steps designed to prevent criminal activity and to maintain security.	law enforcement

Table 31: Examples for `definition_extraction`.

For the definition classification task, the original dataset includes examples of language defining terms. To create a set of non-defining language, Neel Guha randomly selected similarly long excerpts of text from the same Supreme Court opinions. Kevin Tobia analyzed those randomly selected excerpts, identifying any that include definitions (for removal). The resulting dataset has 691 sentences which define sentences, and 646 sentences which do not.

Significance and value This is not a particularly difficult task for human lawyers, and it is unlikely that LLMs would replace lawyers as experts in this process. However, it is possible that LLMs successful in these tasks could provide beneficial legal research roles (e.g. quickly identifying all prior definitions of a specific term in a particular jurisdiction).

Moreover, the definition extraction task serves as a useful test of LLMs abilities, given the task's open-ended nature. The task is not limited to a small set of possible answers (e.g. Yes, No). Rather, it requires identifying which term of all terms in an excerpt is defined. Most of these choices will admit of over ten possible answers (i.e. excerpts of over ten words). Moreover, there is great variety in the language used across the examples. There are hundreds of possible answers, across all items.

F.9 Diversity Jurisdiction

In LEGALBENCH, the Diversity Jurisdiction tasks are denoted as `diversity_*`.

Background Diversity jurisdiction is one of two ways in which a federal court may have jurisdiction over a lawsuit pertaining to state law. Diversity jurisdiction exists when there is (1) complete diversity between plaintiffs and defendants, and (2) the amount-in-controversy (AiC) is greater than $75,000.

"Complete diversity" requires that there is no pair of plaintiff and defendant that are citizens of the same state. However, it is acceptable for multiple plaintiffs to be from the same state, or for multiple defendants to be from the same state.

The AiC requirement allows for certain forms of aggregation. Specifically, if plaintiff **A** asserts two independent claims against defendant **B**, the value of the claims may be added together when considering if the AiC requirement is met. However, a plaintiff may not aggregate the value of claims against two separate defendants, and two plaintiffs may not aggregate claims against the same defendant.

Tasks We define six different tasks, each of which tests the diversity jurisdiction rule under a different pattern of facts. The diversity jurisdiction tasks are:

- `diversity_1`: The fact patterns consists of one plaintiff, one defendant, and one claim per plaintiff-defendant pair.
- `diversity_2`: The fact patterns consists of one plaintiff, two defendants, and one claim per plaintiff-defendant pair.
- `diversity_3`: The fact patterns consists of one plaintiff, one defendant, and two claims per plaintiff-defendant pair.
- `diversity_4`: The fact patterns consists of two plaintiffs, one defendant, and one claim per plaintiff-defendant pair.
- `diversity_5`: The fact patterns consists of two plaintiffs, one defendant, and two claims per plaintiff-defendant pair.
- `diversity_6`: The fact patterns consists of two plaintiffs, two defendants, and two claims per plaintiff-defendant pair.

Construction process We programmatically construct a dataset to test the diversity jurisdiction. We generate randomness over the names of the parties, the claims, and the amounts.

Significance and value It is extremely unlikely LLMs would ever be used to evaluate diversity jurisdiction in practical settings. However, because the task is considered extremely simplistic—and one that first year law students are expected to perform perfectly—it offers a useful evaluation benchmark for LLMs. The structure of the task is potentially non-trivial for LLMs, as it requires identifying the relationships between parties (i.e., who are plaintiffs and defendants), understanding which claims may be aggregated, and computing whether the aggregated amounts meet the AiC requirement.

Task	Facts	Diversity Jurisdiction?
diversity_1	Oliver is from Oregon. William is from Oregon. Oliver sues William for defamation for $3,000.	No
diversity_1	James is from South Dakota. Sophia is from Virginia. James sues Sophia for negligence for $9,010,000.	Yes
diversity_2	Benjamin is from South Carolina. Amelia is from Indiana. Mia is from South Carolina. Benjamin sues Amelia and Mia each for wrongful eviction for $22,000.	No
diversity_2	James is from Colorado. Elijah is from West Virginia. Theodore is from Washington. James sues Elijah and Theodore each for negligence for $2,864,000.	Yes
diversity_3	Ava is from Rhode Island. Theodore is from Rhode Island. Ava sues Theodore for securities fraud for $70,000 and trespass for $6,000.	No
diversity_3	Charlotte is from Colorado. Harper is from Oklahoma. Charlotte sues Harper for breach of contract for $74,000 and securities fraud for $88,000.	Yes
diversity_4	Harper is from New Jersey. Benjamin is from Colorado. Isabella is from Colorado. Harper and Benjamin both sue Isabella for breach of contract for $6,165,000.	No
diversity_4	Noah is from Indiana. Sophia is from West Virginia. Benjamin is from Montana. Noah and Sophia both sue Benjamin for defamation for $3,996,000.	Yes
diversity_5	Noah is from Idaho. Elijah is from Connecticut. Theodore is from Wyoming. Noah and Elijah both sue Theodore for medical malpractice for $57,000 and legal malpractice for $16,000.	No
diversity_5	Charlotte is from Oregon. Mia is from Virginia. Elijah is from Tennessee. Charlotte and Mia both sue Elijah for trademark infringement for $57,000 and medical malpractice for $20,000.	Yes
diversity_6	Lucas is from South Dakota. Amelia is from New Hampshire. Benjamin is from South Dakota. Benjamin is from South Dakota. Lucas and Amelia both sue Benjamin for negligence for $16,000 and wrongful eviction for $76,000. Lucas and Amelia both sue Olivia for medical malpractice for $3,000 and breach of contract for $76,000.	No
diversity_6	Emma is from Kansas. Noah is from Delaware. Elijah is from South Dakota. Elijah is from New Jersey. Emma and Noah both sue Elijah for trademark infringement for $4,000 and trespass for $85,000. Emma and Noah both sue Liam for negligence for $10,000 and defamation for $67,000.	Yes

Table 32: Examples for the Diversity Tasks.

F.10 Function of Decision Section

In LEGALBENCH, the Function of Decision Section task is denoted as `function_of_decision_section`.

Background In common-law legal systems, written judicial decisions serve two functions. First, they resolve the dispute that litigants brought before the court and explain the reason for the court's decision. Second, they become new law, binding on future parties and future courts should another case arise that presents sufficiently similar facts.

Because judicial decisions not only describe the law, but are themselves the law, lawyers in common-law legal systems must be able to read and digest case law to extract key legal principles and apply those principles to their own cases. This skill takes time and practice to develop.

Importantly, not every word in a judicial decision is binding, only the facts and reasoning that were required for the court to reach its decision. Thus, lawyers must distinguish important from trivial facts across numerous past decisions before they can conclude what the law on a particular issue is. One of the most foundational case-reading skills is the ability to review a legal decision and identify the function that each section of the decision serves. In the American legal education system, this skill is taught beginning in the first year of law school, often by encouraging students to identify the function of each section of a decision. A typical classification scheme is as follows:

- Facts: A section of the decision that recounts the historical events and interactions between the parties that gave rise to the dispute.
- Procedural History: A section of the decision that describes the parties' prior legal filings and prior court decisions that led up to the issue to be resolved by the decision.
- Issue: A section of the decision that describes a legal or factual issue to be considered by the court.
- Rule: A section of the decision that states a legal rule relevant to resolution of the case.
- Analysis: A section of the decision that evaluates an issue before the court by applying governing legal principles to the facts of the case
- Conclusion: A section of the decision that articulates the court's conclusion regarding a question presented to it.
- Decree: A section of the decision that announces and effectuates the court's resolution of the parties' dispute, for example, granting or denying a party's motion or affirming, vacating, reversing, or remanding a lower court's decision.

Identifying the function of sections within judicial decisions is a fundamental skill for lawyers in common-law legal systems. Without it, precedent-based legal reasoning would be impossible.

Task The Function of Decision Sections task requires an LLM to determine–given a one-paragraph excerpt of a legal decision–which of the seven functions above that paragraph serves in the context of the entire decision.

Construction process We created a dataset of paragraphs from legal decisions, classified into one of the seven functions above. Paragraphs were taken from decisions in West Publishing's fourth Federal Reporter series, which publishes the decisions of the United States Courts of Appeals. To avoid selection bias and achieve a degree of randomness, paragraphs were selected from sequential decisions, in the order they appeared, spanning all areas of civil and criminal law that fall within the jurisdiction of the federal courts.

Significance and value Beginning law students may initially have trouble identifying the function of a particular section within a judicial opinion, but it quickly becomes a simple task. LLMs would not be called on to perform this task in the actual practice of law, but because it is a foundational legal reasoning skill, it provides a useful measure of reasoning progress for LLMs.

Class	Number of samples
Facts	49
Procedural History	58
Issue	51
Rule	56
Analysis	56
Conclusion	50
Decree	47

Table 33: Test set class distribution.

Excerpt	Function
The Commission's notice and orders, however, are to the contrary. From the very outset, the Commission has made clear that the Governance Order was no more than a call for a proposal that would then be subject to further notice, comment, and revision.	Analysis
Donna and Hurley contend that the Supreme Court's decision in Honeycutt v. United States, —— U.S. ——, 137 S. Ct. 1626, 198 L.Ed.2d 73 (2017), should be applied retroactively to invalidate the forfeiture judgments against them.	Conclusion
For the reasons stated, we affirm the district court's judgment.	Decree
"The Game of Life" is a classic family board game, introduced in 1960 by the Milton Bradley Company to great success. This case involves a long-running dispute between Rueben Klamer, a toy developer who came up with the initial concept of the game, and Bill Markham, a game designer whom Klamer approached to design and create the actual game prototype. Eventually, their dispute (which now involves various assignees, heirs, and successors-in-interest) reduced to one primary issue: whether the game qualified as a "work for hire" under the Copyright Act of 1909. If it did, Markham's successors-in-interest would not possess the termination rights that would allow them to reassert control over the copyright in the game. After considering the evidence produced at a bench trial, the district court concluded that the game was, indeed, such a work. Plaintiff-appellants, who all trace their interest in the game to Markham, challenge that determination. We affirm.	Facts
Officers of the Puerto Rico Police Department watched Julio Casiano-Santana ("Casiano") engage in a drug deal. They arrested him, recovering a loaded pistol and three bags of crack cocaine from the scene. Casiano was charged with possession of a firearm in furtherance of a drug trafficking crime, 18 U.S.C. § 924(c)(1)(A)(i), two counts of possession with intent to distribute controlled substances, 21 U.S.C. § 841(a)(1) and (b)(1)(C), and possession of a firearm by a convicted felon, 18 U.S.C. § 922(g)(1).	Issue
On remand, the district court held a new sentencing hearing, in which Lawrence allocuted. Resentencing Transcript at 11–12, United States v. Lawrence, No. 03-cr-00092-CKK (D.D.C. Oct. 5, 2009), ECF No. 103. Lawrence told the court that, while incarcerated, he had "been trying to do the right things as far as * * * becoming a man so I can provide for my son, he's 11 and very big." Id. Lawrence's mother was "getting old" and does "the best that she can[,]" but his son had "health issues as far as * * * weight gain and a lot of other things." Id. at 12. Lawrence explained that he "just want[ed] a chance to be a father" to his son, and that he "was just hoping that it's possible that * * * I can get out in his life before * * * the streets * * * or anything that maybe I have done affect him[.]" Id. He said he wanted to "be a productive citizen[,]" and noted that he "read the Bible" and "attended church, school, [and] college." Id. He admitted that he had "gotten into some altercations," but "not because I wanted to, but it's prison, and you know, there's all types of people in prison." Id. While "making no excuses" for his actions, he said he "was just hoping the Court would have leniency" in his "particular case." Id.	Procedural History
The border between interpretation and bare consultation can be hazy and, therefore, "difficult to plot." Lawless, 894 F.3d at 18 (citing Livadas, 512 U.S. at 124 n.18, 114 S.Ct. 2068). This case, however, does not closely approach the border: on their face, Rose's state-law claims require more than bare consultation of the CBA. They substantially depend on construing the terms of the agreement (the CBA) that RTN and the Union negotiated. We explain briefly.	Rule

Table 34: Examples for function_of_decision_section

F.11 Hearsay

In LEGALBENCH, the hearsay task is denoted as `hearsay`.

Background The Federal Rules of Evidence dictate that "hearsay" evidence is inadmissible at trial. Hearsay is defined as an "out-of-court statement introduced to prove the truth of the matter asserted." In determining whether a piece of evidence meets the definition of hearsay, lawyers ask three questions:

1. Was there a statement? The definition of statement is broad, and includes oral assertions, written assertions, and non-verbal conduct intended to communicate (i.e. *assert*) a message. Thus, for the purposes of the hearsay rule, letters, verbal statements, and pointing all count as statements.
2. Was it made outside of court? Statements not made during the trial or hearing in question count as being out-of-court.
3. Is it being introduced to prove the truth of the matter asserted? A statement is introduced to prove the truth of the matter asserted if its truthfulness is essential to the purpose of its introduction. Suppose that at trial, the parties were litigating whether Alex was a soccer fan. Evidence that Alex told his brother "I like soccer," would be objectionable on hearsay grounds, as (1) the statement itself asserts that Alex likes soccer, and (2) the purpose of introducing this statement is to prove/disprove that Alex likes soccer. In short, the truthfulness of the statement's assertion is central to the issue being litigated. However, consider if one of the parties wished to introduce evidence that Alex told his brother, "Real Madrid is the greatest soccer team in the world." This statement would **not** be hearsay. It's assertion—that Real Madrid is the greatest soccer team in the world—is unrelated to the issue being litigated. Here, one party is introducing the statement not to prove what the statement says, but to instead show that a particular party (i.e. Alex) was the speaker of the statement.

Task Given a legal issue and a piece of prospective evidence, the LLM must determine whether the evidence constitutes hearsay under the above test.

We note that in practice, many pieces of evidence which are hearsay are nonetheless still admissible under one of the many hearsay exception rules. We ignore these exceptions for our purposes, and leave the construction of benchmarks corresponding to these exceptions for future work.

Construction process We create the hearsay dataset by hand, drawing inspiration from similar exercises available in legal casebooks and online resources. The dataset consists of 5 slices, where each slice tests a different aspect of the hearsay rule. We randomly select 1 sample from each slice to be in the train set. The remainder of the slice constitutes the test set (for a total of 95 samples). The slices (with test set counts) are:

- Statement made in court ($n = 14$): Fact patterns where the statement in question is made during the course of testimony at trial. Thus, the statement is not hearsay.
- Non-assertive conduct ($n = 19$): Fact patterns where the evidence does not correspond to a statement. Hence, the hearsay rule is inapplicable.
- Standard hearsay ($n = 29$): Fact patterns where there is an oral statement, it is said out of court, and it is introduced to prove the truth of the matter asserted. Thus, these fact patterns correspond to hearsay.
- Non-verbal hearsay ($n = 12$): Fact patterns where the statement is hearsay, but made in writing or through assertive conduct (e.g. pointing).
- Not introduced to prove truth ($n = 20$): Fact patterns where an out-of-court statement is introduced to prove something other than what it asserts.

Significance and value The hearsay rule is commonly taught in law school as part of Evidence. Law students are expected to understand the rule, and how to apply it. The hearsay task is interesting for LLM evaluation because it emphasizes multi-step reasoning—the test for hearsay encompasses several different steps, where each step differs in difficulty. These include:

- Event detection: The LLM must determine whether the fact pattern mentions a statement being made.
- Spatial reasoning: The LLM must determine whether the statement was made inside a court room.
- Argument extraction: The LLM must determine what the statement is asserting.
- Argument relevance: The LLM must finely determine whether the assertion is relevant to the issue being litigated.

Facts	**Hearsay?**
On the issue of whether Will knew that the company intended to announce its drug trials had been cancelled, the fact that he told the jury that "he didn't know the first thing about how medicines worked."	No
On the issue of whether Gerald was alive immediately after being attacked by Kathryn, Gerald's statement, "I was attacked by Kathryn."	No
On the issue of whether Susan was familiar with Shakespeare, the fact that she had once played the role of Macbeth and recieved a standing ovation after her monologue.	No
To prove that the insured under a life policy is dead, his wife offers a death certificate.	Yes
On the issue of whether Albert bought a knife, Angela testified that he shook his head when she asked him.	Yes
On the issue of whether the brakes were faulty, Amy testifies that she heard Arthur claim that he thought something was wrong with the car.	Yes

Table 35: Examples for hearsay

F.12 Insurance Policy Interpretation

In LEGALBENCH, the Insurance Policy Interpretation task is denoted as `insurance_policy_interpretation`.

Background Insurance disputes often arise when parties disagree on whether a claim is covered under a certain insurance policy. To study such disagreements in interpretation, researchers at Stanford recruited crowdsource workers to review a pair of an insurance policy and a claim and respond whether they believe the claim is covered. A policy-claim pair whose applicability the workers disagree with each other on suggests ambiguity in the policy text.

Task The Insurance Interpretation task requires an LLM to review a pair of an insurance policy and a claim and determine whether the policy clearly covers the claim, clearly does not cover it, or if it is unclear whether it covers it or not.

Construction process The clause-claim pairs are manually constructed before being reviewed by crowdsource workers [139]. To convert the numbers of Covered/Not_Covered/Can't_Decide responses to discrete labels, we first calculate the 95% multinomial confidence interval of the proportion of each response. We then choose the label for which the confidence interval lower bound is greater than or equal to .5. If no label has a lower bound $\geq .5$, we classify the policy-claim pair as "It's ambiguous." This conversion process ensures that individual crowdsource workers do not arbitrarily sway the labels. Examples for each label can be found in Table 36.

Policy: Harper's insurance covers damage from "House Removal," which includes "damage to belongings that occurs while being stored by professional removal contractors." **Claim**: Harper is moving to a new home on the other side of town. Because her old home has already sold and her new home is not yet ready for her to move in, she checks into a hotel and asks a professional moving company to store some of her belongings at the company warehouse. A couple days before she is set to move in, the warehouse floods, which ruins the items that the movers were storing for Harper. Harper files a claim with her insurance company for the damage to her belongings. **Label**: Covered
Policy: Denise's insurance covers damage from "House Removal," defined as "damage to belongings caused while being removed by professional removal contractors from the home." **Claim**: Denise is moving to a new home on the other side of town. She asks her uncle, a retired professional mover, to help move her belongings out of her current home. During the move, her uncle's truck is involved in a minor accident that damages several pieces of her furniture and other belongings. Denise files a claim with her insurance company for the damage to her belongings. **Label**: Not Covered
Policy: Jason has insurance coverage against loss and damage from "Identity Theft," which excludes "identity theft connected with the policyholder's business." **Claim**: Jason is a successful car salesman. One day, while Jason is at home, hackers manage to infiltrate Jason's home WiFi network. The hackers steal Jason's social security number and open a number of fraudulent lines of credit in his name. To resolve the fraud, Jason must spend thousands of dollars in legal fees. Jason files a claim with his insurance company for his losses. **Label**: It's ambiguous.

Table 36: Examples for `insurance_interpretation`.

Significance and value The ability to determine whether an insurance claim is covered under a given policy can significantly reduce claim processing time. It can also shine light on potential ambiguity in existing policies. Additionally, this task represents one of the rare benchmarks where an LLM is required to predict laypeople's legal interpretations, as we retrieve the ground truth labels based on crowdsourced responses.

F.13 International Citizenship Questions

In LEGALBENCH, the International Citizenship Questions task is denoted as `international_citizenship_questions`.

Background The GLOBALCIT Citizenship Law Dataset is a valuable resource that comprehensively categorizes citizenship acquisition and loss methods in 190 countries. It enables cross-country comparisons and offers insights into global trends in citizenship laws and examines 28 different ways in which citizenship can be acquired, as well as 15 ways laws allow citizenship to be lost. The original dataset is formulated as a tabular survey dataset. We change this survey format into Yes/No questions about specific countries and their laws as of 2020 resulting in 9300 question- answer pairs.

Task The model must answer yes/no questions about global citizenship law.

Question	Answer
Consider the country of Central African Republic. Does the country provide for acquisition of citizenship by a person who is in the public service (including military service) and, if so, under which conditions?	No
Consider the country of Bolivia. Does the country provide for involuntary loss of citizenship by a person who is adopted by or in legal guardianship of a citizen of another country and, if so, under which conditions?	No
Consider the country of Denmark. Which residence conditions does the country provide for residence-based acquisition?	Yes
Consider the country of Germany. Does the country require the demonstration of civic knowledge or cultural integration for residence-based acquisition?	Yes

Table 37: Examples for international_citizenship_questions

Contruction process We download the GLOBALCIT Citizenship Law Dataset [137] and craft a script that converts the tabular survey data into yes/no questions, appending information about the country and the time at which the survey was created.

Significance and value Understanding knowledge about the law globally is important to evaluate. To successfully answer legal reasoning questions globally language models must be able to retrieve a rule and then reason about it.

F.14 Learned Hand Tasks

In LEGALBENCH, the Learned Hand tasks are denoted as `learned_hand_*`.

Background A person may experience problems in many areas of their lives – their family, work, finances, housing, education, driving, and more – which legal professionals would recognize as being 'legal issues'. The person may not know that a problem with a credit card company, a landlord, a spouse, or an employer is a 'legal issue', or what terminology or categorization a lawyer would use to make sense of it.

The designation of a legal issue means that a person may benefit from getting specialized guidance from a legal professional to resolve this problem because they can guide them on their rights, liabilities, possible options, procedures, and specialized legal tasks. Not all people may want to pursue legal services to resolve a legal issue. But legal issue-spotting can help them both make sense of the problem they are experiencing, and what services and laws might be available if they wish to make use of them.

Legal professionals typically carry out issue-spotting during an intake process. They receive a person's verbal or written description of the situation they are in. Then the professional identifies the main legal issues that are apparent in this situation, starting at the main top-level categories and then sometimes proceeding to identify more specific sub-categories of legal issues. For example, the professional may identify that a person's situation involves a legal issue with the top-level category of 'housing' and specific sub-categories of 'possible eviction for non-payment of rent' and 'poor living conditions of their rental'.

The professional may identify multiple overlapping legal issue categories in one situation. For example, the professional may identify that a person has a housing law issue and a family law issue if their landlord is threatening to evict them because of the police being called to the rental home because of a domestic violence incident.

The main categories of legal issues that professionals would identify in people's situations are:

- **Benefits**: A situation would have a benefits issue if it involves the person attempting to resolve a problem with applying for, receiving, or discontinuing public benefits and social services from the government. This could include benefits that support them regarding food, disability, old age, housing, health, unemployment, child care, or other social needs.
- **Business**: A situation would have a business issue if the person is running a small business or nonprofit, and encounters a problem around incorporation, licenses, taxes, regulations, bankruptcies, or natural disasters. This category is not meant to apply to larger corporate legal issues, but rather the kinds of business problems that an individual might bring to a legal professional for help.
- **Consumer**: A situation would have a consumer legal issue if the person was dealing with problems around debt and money, insurance, consumer goods and contracts, taxes, or small claims about the quality of service.
- **Courts**: A situation would be categorized as a courts issue if the person is dealing with a problem around how to interact with the court system or with lawyers more broadly. This may involve the person attempting to follow legal procedures, court rules, or filing requirements, or it may involve them attempting to hire, manage, or address lawyers.
- **Crime**: A situation would have a crimes issue if the person is dealing with the criminal justice system as a defendant, victim, or family member. They may be experiencing problems around being investigated, searched, or charged with a crime, or going to a criminal trial and prison, or being a victim of a crime.
- **Divorce**: A situation would be categorized as a divorce issue if a person is dealing with a separation, divorce, or annulment while splitting with a spouse or partner. The problem may involve separation, spousal support, splitting money and property, child support, visitation, or following the related court process.
- **Domestic Violence**: A situation would have a domestic violence issue if the person is dealing with abuse with a partner, family member, or other intimate acquaintance. The situation may involve understanding rights and laws related to domestic violence, getting protective orders, enforcing them, reporting abuse, and dealing with collateral consequences to housing, finances, employment, immigration, and education.
- **Education**: A situation has an education issue if the person is dealing with a problem around school for themselves or a family member. The situation may involve accommodations for special needs, discrimination, student debt, discipline, or other issues in education.
- **Employment**: A situation would be identified as having an employment issue if the person has a problem with a job, including during the application process, during the job, or after ending employment. Problems may include discrimination, harassment, payment, unionizing, pensions, termination, drug testing, background checks, worker's compensation, classification as a contractor, or more.
- **Estates**: A situation would have an estates issue if a person is dealing with an estate, wills, or guardianship. This may include issues around end-of-life planning, health and financial directives, trusts, guardianships, conservatorships, and other estate issues that people and family deal with.
- **Family**: A situation would have a family law issue if a person is dealing with an issue involving a family member. This may include issues around divorce, child custody, domestic violence, adoption, paternity, name change, and other family issues.

- **Health**: A situation would be categorized as a health law issue if the person is dealing with problems around accessing health services or protecting their rights in medical settings. This may involve problems with accessing health care, paying for care, getting public benefits for care, privacy of medical records, problems with quality of care, or other issues.
- **Housing**: A situation would have a housing law issue if a person is dealing with problems around the housing where they live, or that they own. These include problems with renting a home, eviction, living conditions, discrimination, foreclosure, post-disaster housing, housing assistance, and more.
- **Immigration**: A situation would have an immigration issue if a person is not a full citizen in the US and is dealing with problems related to their status. This may include understanding visa options, working as an immigrant, political asylum, border searches, deportation, human trafficking, refugees, immigration court hearings, and more.
- **Torts**: A situation would be categorized as having a torts issue if the person is dealing with an accident or conflict with another person that involves some perceived harm. These problems may include a car accident, conflicts with neighbors, dog bites, bullying, harassment, data privacy breaches, being sued, or suing someone else.
- **Traffic**: A situation would have a traffic law issue if the person is experiencing a problem with traffic, parking, or car ownership. This might include problems with getting ticketed, getting or reinstating a driver's license, car accidents, purchasing a car, repossession, and more.

This set of categories is commonly used by legal professionals as they triage potential clients. The LIST Taxonomy from Stanford Legal Design Lab has formalized these categories into a machine-readable taxonomy, available at `https://taxonomy.legal/`. The LIST taxonomy builds on the taxonomies built by legal aid groups, like the Legal Services Corporation categories list that most legal aid groups use to encode their matters, signifying what issues they helped clients with.[22]. LIST also builds off of the legal aid community's National Subject Matter Index, which was a more extensive list of categories to further assist legal aid groups in tracking the issues they helped people with.[23].

Performing the Legal Issue-Spotting task requires parsing through informal wording and structures that a person may use to convey the situation they are struggling with. Typically the person is writing this narrative down in an informal, quick manner (like in an online intake form on a legal services website) or speaking it aloud (like on an intake hotline, or during an in-person interview). The narratives are not typically structured into a concise order. They use informal terminology rather than legal terms.

Task The Legal Issue-spotting task requires an LLM to consider a person's narrative about their situation. The LLM must use this narrative to determine which legal issue category (or categories) apply to the person's situation.

Construction process There is a crowdsourced dataset, created via the online labeling game Learned Hands, that has established when and how these legal issue categories apply to people's narratives. The narratives are drawn from the subreddit r/legaladvice, in which people share several lines or paragraphs about the situation they are dealing with, which they think might involve legal issues. The moderators of the subreddit are active in managing activity, so the posts do not contain personal identifying information or off-topic postings.

The Stanford Legal Design Lab and Suffolk LIT Lab built the Learned Hands game so that law students and lawyers could read narratives one by one, and then answer a series of yes-no-skip questions about what legal issue seems to be present. Once there are sufficient consistent votes for a certain label to apply, or to not apply, to a narrative, then the label is finalized. A narrative may have more than one label, as mentioned above.

Significance and value The legal issue categorization helps the professional triage the person to the right services, resources, and procedures that can assist them in resolving the legal issue. If an LLM is able to identify legal issues in people's informal narratives, this demonstrates an ability to perform a key task in people's justice journeys. Issue-spotting by an LLM may be able to help a person who is just starting to explore whether or how they should engage with legal services, courts, or exercising their rights.

The issue-spotting task may be provided online, when people are visiting legal help websites and trying to find what guide, form, or service would best help them with their problem. Or it may be integrated into the intake process that paralegals or justice workers carry out over hotlines or in-person, to speed up the often lengthy intake process.

[22] See the LSC's list at `https://www.lsc.gov/i-am-grantee/grantee-guidance/lsc-reporting-requirements/case-service-reporting/csr-handbook-2017`

[23] See the NSMI database at `https://nsmi.lsntap.org/`

F.15 Legal Reasoning Causality

In LEGALBENCH, the Legal Reasoning Causality task is denoted as `legal_reasoning_causality`.

Background In many legal domains, systematic evidential barriers hinder the substantiation of causal claims through direct evidence. To address these shortcomings, courts have recognized the power of statistical evidence in establishing causation in various contexts, such as product liability,[24] medical malpractice,[25] discrimination,[26] and more. For instance, when pursuing a labor discrimination claim, the plaintiff must establish that her protected trait was the underlying reason for the alleged discriminatory decision (e.g., firing or not hiring). However, direct evidence of discriminatory intent rarely exists, so it is often nearly impossible to refute the possibility that other (legitimate) differences between two employees or candidates were the cause for favoring one over the other. In such cases, litigants can and often do try to substantiate a causal link between the plaintiff's group affiliation and the defendant's behavior through statistical analysis. For instance, plaintiffs might send fictitious resumes that differ only by the suspected demographic characteristic,[27] akin to a field experiment. Likewise, statistical analysis of observational data that controls for the major factors affecting the employment practice can be used to demonstrate whether a specific social group suffers from inferior outcomes (relative to some control group) vis-à-vis a particular employer, landlord, or lender that engages with a sufficiently large number of employees or customers.[28]

Task The "causal reasoning" task requires an LLM to determine whether the court's reasoning regarding the finding of whether a causal link exists between the plaintiff's protected trait and the allegedly discriminatory decision relied on either statistical or direct-probative evidence. It requires understanding the types of words that are used to describe statistical evidence in any given context (regression, correlation, variables, control, and more), and the extent to which those words relate to substantiating a finding of causality (as opposed to other legal components).

Construction process We manually created a dataset of fifty-nine excerpts from court decisions in lawsuits alleging labor market discrimination filed in US Federal District Courts. First, fifty-nine court decisions involving claims of labor discrimination were identified using the LexisNexis database. Second, the passages in which the finding of causality appeared were identified and extracted. Third, we coded the passages as either relying on statistical evidence (e.g., regression analysis, findings of correlation, etc.) or on direct evidence (e.g., witnesses, documents, etc.).

We selected two random samples from each class to use as part of the train split.

Significance and value The potential of LLMs to identify different types of legal reasoning in general, and the finding of causality in particular, has implications both for the legal profession and for the academic study of law and judicial decision-making. First, algorithmic tools are gradually being utilized by lawyers to assist them in preparing for litigation. Specifically, given the heterogeneity among judges, a key element of a successful litigation strategy is a lawyer's ability to construct their arguments based on the specific inclinations of the judge assigned to the case. Gaining an accurate understanding of judges' unique mode of reasoning (including, e.g., the types of evidence they tend to rely on), based on their prior decisions, is crucial for winning any lawsuit. Second, databases consisting of court decisions are the most common source for studying the law and judicial decision-making in legal academia. However, these databases are typically limited to rather technical information, such as the names of the parties and the judge(s), the legal area of the case, and the like. The essential part of any judicial opinion – the legal reasoning – is typically treated as a black box. An LLM that could classify the various types of legal reasoning – e.g., what evidence is used to establish causation – can facilitate studying judicial decision- making in ways currently not feasible at large scales.

[24] See, e.g, Neurontin v. Pfizer, 712 1st Cir. 52 (2013).

[25] O'Neal v. St. John Hosp. & Med. Ctr, 487 Mich SC, 485 (2010).

[26] See, e.g., International Broth. of Teamsters v. U.S., 431 U.S. 324 ("[i]n many cases the only available avenue of proof is the use of racial statistics to uncover clandestine and covert discrimination by the employer or union involved"); Bazemore v. Friday, 478 U.S 385 (1986); Marcus Jones v. Lee Way Motor Freigh, 431 10 th cir 245 (1970) ("In racial discrimination cases, statistics often demonstrate more than the testimony of many witnesses").

[27] Havens Realty Corp. v. Coleman, 455 U.S. 363, 374-75, 71 L. Ed. 2d 214, 102 S. Ct. 1114 (1982).

[28] See Bazemore v. Friday, 478 U.S 385 (1986) ("although it need not include every conceivable factor. Given the frequency of employment discrimination litigation in the contemporary United").

Excerpt	Relies on statistical evidence?
However, a review of the "over base level" numbers of the four comparators and Escalera in core endourology reflects significant differences in the severity of losses between the comparators and Escalera during the January through June 2015 time period. Escalera's utilization of different time periods for each comparator within 2015 is not appropriate when examining the team managers' performances given Bard Medical's Solo/Skylite production products. Using the same time frame for each comparator, the record reflects that between January through June of 2015, Kunzinger was $55,626.89 below base, Santoro was $160,651.77 above base, Peters was $20,070.56 above base, and Martin was $79,932.38 above base. (Ottley Dep. Exs. 3, 12, 14, 16.) These numbers demonstrate that the "losses" experienced by the comparators during the same time period as Escalera are not substantially identical. Escalera's loss of base was $68,799.06 more than [**25] the closest comparator he identified. Additionally, comparing the "over base level" numbers of the comparators and Escalera between January through October 2015 reflects that at the time Escalera was terminated he had suffered significantly more loss over base than his identified comparators: Escalera was [*805] $174,792.44 below base, Kunzinger was $101,132.60 below base,3 Santoro was $110,078.73 above base, Peters was $31,876.80 below base, and Martin was $1,611.79 below base. Because of these significant differences in losses, no reasonable jury could find that these four comparators and Escalera are similarly situated in all relevant respects.	No
Equally without evidentiary significance is the statistical analysis of the list of 17; indeed, the analysis was not even admissible under HN4 the standard of Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 125 L. Ed. 2d 469, 113 S. Ct. 2786 (1993), governing the admissibility of expert testimony, which requires the district judge to satisfy himself that the expert is being as careful as he would be in his regular professional work outside his paid litigation consulting. E.g., Braun v. Lorillard Inc., 84 F.3d 230, 234-35 (7th Cir. 1996); Rosen v. Ciba-Geigy Corp., 78 F.3d 316, 318 (7th Cir. 1996); Daubert v. Merrell Dow Pharmaceuticals, Inc., 43 F.3d 1311, 1316-19 (9th Cir. 1995); cf. Mid-State Fertilizer Co. v. [**7] Exchange National Bank, 877 F.2d 1333, 1339 (7th Cir. 1989). Although the expert used standard statistical methods for determining whether there was a significant correlation between age and retention for the 17 persons on the list, see Michael O. Finkelstein & Bruce Levin, Statistics for Lawyers 157 (1990) (Fisher's exact test), the omission of Sebring and Shulman from the sample tested was arbitrary. The expert should at least have indicated the sensitivity of his analysis to these omissions. More important is the expert's failure to correct for any potential explanatory variables other than age. Completely ignored was the more than remote possibility that age was correlated with a legitimate job-related qualification, such as familiarity with computers. Everyone knows that younger people are on average more comfortable with computers than older people are, just as older people are on average more comfortable with manual-shift cars than younger people are. Three weeks of training might go some distance toward closing the computer-literacy gap, yet it would be more surprising than otherwise if so short a period of training could close the gap completely. The expert could easily [**8] have inquired about the feasibility of ascertaining through discovery the history of the use of computers by each of the employees on the list of 17.	Yes

Table 38: Examples for `legal_reasoning_causality`.

Class	Number of samples
Yes	31
No	24

Table 39: Test set class distribution.

F.16 MAUD Tasks

In LEGALBENCH, the MAUD tasks are denoted as `maud_*`.

Background We adapt the Merger Agreement Understanding Dataset (MAUD) for LEGALBENCH. MAUD consists of over 37,000 expert-annotated examples for a set of legal reading comprehension tasks based on the American Bar Association's 2021 Public Target Deal Point Study. In the Study, lawyers review merger agreements and identify key legal clauses ("deal points") within those contracts. The lawyers then specify the nature of the clauses by answering a predetermined set of questions that cover a wide range of topics, including conditions to closing, the definition of material adverse effect, and remedies to breach of contract. MAUD's multiple-choice format, according to [142], assesses an LLM's ability to interpret the meaning of specialized legal language.

Task The tasks take advantage of MAUD's reading comprehension component. They require an LLM—given a key legal clause and a set of descriptions for the clause—to choose the option that best describes the clause.

Construction process These tasks are constructed by transforming the abridged dataset released by [142]. The abridged dataset contains 14,928 examples with deal points extracted from 94 merger agreements covering 92 multiple-choice questions. We narrow down to 57 questions by filtering out the ones with fewer than 50 examples. Each example consists of the text of a deal point, the question, options, and the answer key.

We create translations that map the questions into human-readable multiple-choice prompts. For instance, the prompt for the question "Accuracy of Target 'General' R&W: Bringdown Timing" is "When are representations and warranties required to be made according to the bring down provision?" It is then followed by an enumeration of the options for the LLM to choose among.

We focus on MAUD's abridged examples because we are interested in assessing an LLM's legal reading comprehension capability rather than its ability to extract relevant text segment given a complete deal point. Additionally, inputs of examples from the main dataset, which contains complete deal point texts, are oftentimes far longer than what an average open-source LLM could ingest at once, rendering them unsuitable for benchmarking purposes.

The table below lists the question and options for each MAUD-based LEGALBENCH task along with an example input-answer pair.

Table 40: MAUD Tasks

Task
Task name: `maud_type_of_consideration` **Question**: What type of consideration is specified in this agreement? **Options**: A: All Cash; B: All Stock; C: Mixed Cash/Stock; D: Mixed Cash/Stock: Election **Example**: each Share <omitted> shall be converted into the right to receive the Offer Price in cash, without interest (the "Merger Consideration"), minus any withholding of Taxes required by applicable Laws in accordance with Section 3.6(d) (Page 20) **Answer**: A
Task name: `maud_accuracy_of_target_general_rw_bringdown_timing_answer` **Question**: When are representations and warranties required to be made according to the bring down provision? **Options**: A: At Closing Only; B: At Signing & At Closing **Example**: Section 7.2 Conditions to Obligations of Parent and Acquisition Sub to Effect the Merger. The obligations of Parent and Acquisition Sub to effect the Merger are, in addition to the conditions set forth in Section 7.1, further subject to the satisfaction or (to the extent not prohibited by Law) waiver by Parent at or prior to the Effective Time of the following conditions: (a) each of the representations and warranties of the Company contained in this Agreement, without giving effect to any materiality or "Company Material Adverse Effect" or similar qualifications therein, shall be true and correct as of the Closing Date, except for such failures to be true and correct as would not, individually or in the aggregate, have a Company Material Adverse Effect (except to the extent such representations and warranties are expressly made as of a specific date, in which case such representations and warranties shall be so true and correct as of such specific date only); (Page 67) **Answer**: A
Task name: `maud_accuracy_of_target_capitalization_rw_(outstanding_shares)_bringdown_standard_answer` **Question**: How accurate must the capitalization representations and warranties be according to the bring down provision? **Options**: A: Accurate in all material respects; B: Accurate in all respects; C: Accurate in all respects with below-threshold carveout; D: Accurate in all respects with de minimis exception

Table 40 – continued from previous page

Task
Example: Conditions to the Offer Notwithstanding any other term of the Offer or this Agreement to the contrary, Merger Sub will not be required to accept for payment or, subject to any applicable rules and regulations of the SEC, including Rule 14e-l(c) under the Exchange Act (relating to Merger Sub's obligation to pay for or return tendered Shares promptly after the termination or withdrawal of the Offer), to pay for any Shares tendered pursuant to the Offer, and may delay the acceptance for payment of or, subject to any applicable rules and regulations of the SEC, the payment for, any tendered Shares, and (subject to the provisions of this Agreement) may terminate the Offer and not accept for payment any tendered Shares, at any scheduled Expiration Date (as it may have been extended pursuant to Section 2.1 of this Agreement) if <omitted> (ii) any of the additional conditions set forth below are not satisfied or waived in writing by Parent at the Expiration Time: <omitted> (d) Representations and Warranties. Each of the representations and warranties set forth in: <omitted> (iv) this Agreement (other than those set forth in the foregoing clauses (i), (ii) and (iii) of this clause (d) of Annex I), without giving effect to any "materiality" or "Material Adverse Effect" qualifiers or qualifiers of similar import set forth therein, shall be true and correct as of the consummation of the Offer as though made as of the consummation of the Offer (Page 107) **Answer**: D
Task name: `maud_accuracy_of_fundamental_target_rws_bringdown_standard` **Question**: How accurate must the fundamental representations and warranties be according to the bring down provision? **Options**: A: Accurate at another materiality standard (e.g., hybrid standard); B: Accurate in all material respects; C: Accurate in all respects **Example**: (b) Additional Conditions to Obligation of Parent and Merger Sub. <omitted> the representations and warranties of the Company set forth in Article 3 shall be true and correct <omitted> at and as of the Closing as if made at and as of such time (Page 11) **Answer**: A
Task name: `maud_ability_to_consummate_concept_is_subject_to_mae_carveouts` **Question**: Is the "ability to consummate" concept subject to Material Adverse Effect (MAE) carveouts? **Options**: A: No; B: Yes **Example**: "Material Adverse Effect" means, with respect to Huntington, TCF or the Surviving Corporation, as the case may be, any effect, change, event, circumstance, condition, occurrence or development that, either individually or in the aggregate, has had or would reasonably be likely to have a material adverse effect on (i) the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole (provided, however, that, with respect to this clause (i), Material Adverse Effect shall not be deemed to include the impact of (A) changes, after the date hereof, in U.S. generally accepted accounting principles ("GAAP") or applicable regulatory accounting requirements, (B) changes, after the date hereof, in laws, rules or regulations (including the Pandemic Measures) of general applicability to companies in the industries in which such party and its Subsidiaries operate, or interpretations thereof by courts or Governmental Entities, (C) changes, after the date hereof, in global, national or regional political conditions (including the outbreak of war or acts of terrorism) or in economic or market (including equity, credit and debt markets, as well as changes in interest rates) conditions affecting the financial services industry generally and not specifically relating to such party or its Subsidiaries (including any such changes arising out of the Pandemic or any Pandemic Measures), (D) changes, after the date hereof, resulting from hurricanes, earthquakes, tornados, floods or other natural disasters or from any outbreak of any disease or other public health event (including the Pandemic), (E) public disclosure of the execution of this Agreement, public disclosure or consummation of the transactions contemplated hereby (including any effect on a party's relationships with its customers or employees) (it being understood that the foregoing shall not apply for purposes of the representations and warranties in Sections 3.3(b), 3.4, 4.3(b) or 4.4) or actions expressly required by this Agreement or that are taken with the prior written consent of the other party in contemplation of the transactions contemplated hereby, or (F) a decline in the trading price of a party's common stock or the failure, in and of itself, to meet earnings projections or internal financial forecasts, but not, in either case, including any underlying causes thereof; except, with respect to subclauses (A), (B), (C) or (D), to the extent that the effects of such change are materially disproportionately adverse to the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole, as compared to other companies in the industry in which such party and its Subsidiaries operate) or (ii) the ability of such party to timely consummate the transactions contemplated hereby. (Page 18) **Answer**: A
Task name: `maud_fls_(mae)_standard` **Question**: What is the Forward Looking Standard (FLS) with respect to Material Adverse Effect (MAE)? **Options**: A: "Could" (reasonably) be expected to; B: "Would"; C: "Would" (reasonably) be expected to; D: No; E: Other forward-looking standard

Table 40 – continued from previous page

Task
Example: "Material Adverse Effect" means, with respect to BancorpSouth, Cadence or the Surviving Entity, as the case may be, any effect, change, event, circumstance, condition, occurrence or development that, either individually or in the aggregate, has had or would reasonably be expected to have a material adverse effect on (i) the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries taken as a whole (provided, however, that, with respect to this clause (i), Material Adverse Effect shall not be deemed to include the impact of (A) changes, after the date hereof, in U.S. generally accepted accounting principles ("GAAP") or applicable regulatory accounting requirements, (B) changes, after the date hereof, in laws, rules or regulations (including the Pandemic Measures) of general applicability to companies in the industries in which such party and its Subsidiaries operate, or interpretations thereof by courts or Governmental Entities (as defined below), (C) changes, after the date hereof, in global, national or regional political conditions (including the outbreak of war or acts of terrorism) or in economic or market (including equity, credit and debt markets, as well as changes in interest rates) conditions affecting the financial services industry generally and not specifically relating to such party or its Subsidiaries (including any such changes arising out of a Pandemic or any Pandemic Measures), (D) changes, after the date hereof, resulting from hurricanes, earthquakes, tornados, floods or other natural disasters or from any outbreak of any disease or other public health event (including a Pandemic), (E) public disclosure of the execution of this Agreement, public disclosure or consummation of the transactions contemplated hereby (including any effect on a party's relationships with its customers or employees) or actions expressly required by this Agreement or that are taken with the prior written consent of the other party in contemplation of the transactions contemplated hereby, or (F) a decline in the trading price of a party's common stock or the failure, in and of itself, to meet earnings projections or internal financial forecasts (it being understood that the underlying causes of such decline or failure may be taken into account in determining whether a Material Adverse Effect has occurred), except to the extent otherwise excepted by this proviso); except, with respect to subclauses (A), (B), (C), or (D) to the extent that the effects of such change are materially disproportionately adverse to the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole, as compared to other companies in the industry in which such party and its Subsidiaries operate), or (ii) the ability of such party to timely consummate the transactions contemplated hereby. (Page 19) **Answer**: C
Task name: `maud_general_economic_and_financial_conditions_subject_to_disproportionate_impact_modifier` **Question**: Do changes caused by general economic and financial conditions that have disproportionate impact qualify for Material Adverse Effect (MAE)? **Options**: A: No; B: Yes **Example**: "Material Adverse Effect" means, with respect to Huntington, TCF or the Surviving Corporation, as the case may be, any effect, change, event, circumstance, condition, occurrence or development that, either individually or in the aggregate, has had or would reasonably be likely to have a material adverse effect on (i) the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole (provided, however, that, with respect to this clause (i), Material Adverse Effect shall not be deemed to include the impact of (A) changes, after the date hereof, in U.S. generally accepted accounting principles ("GAAP") or applicable regulatory accounting requirements, (B) changes, after the date hereof, in laws, rules or regulations (including the Pandemic Measures) of general applicability to companies in the industries in which such party and its Subsidiaries operate, or interpretations thereof by courts or Governmental Entities, (C) changes, after the date hereof, in global, national or regional political conditions (including the outbreak of war or acts of terrorism) or in economic or market (including equity, credit and debt markets, as well as changes in interest rates) conditions affecting the financial services industry generally and not specifically relating to such party or its Subsidiaries (including any such changes arising out of the Pandemic or any Pandemic Measures), (D) changes, after the date hereof, resulting from hurricanes, earthquakes, tornados, floods or other natural disasters or from any outbreak of any disease or other public health event (including the Pandemic), (E) public disclosure of the execution of this Agreement, public disclosure or consummation of the transactions contemplated hereby (including any effect on a party's relationships with its customers or employees) (it being understood that the foregoing shall not apply for purposes of the representations and warranties in Sections 3.3(b), 3.4, 4.3(b) or 4.4) or actions expressly required by this Agreement or that are taken with the prior written consent of the other party in contemplation of the transactions contemplated hereby, or (F) a decline in the trading price of a party's common stock or the failure, in and of itself, to meet earnings projections or internal financial forecasts, but not, in either case, including any underlying causes thereof; except, with respect to subclauses (A), (B), (C) or (D), to the extent that the effects of such change are materially disproportionately adverse to the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole, as compared to other companies in the industry in which such party and its Subsidiaries operate) or (ii) the ability of such party to timely consummate the transactions contemplated hereby. (Page 18) **Answer**: A
Task name: `maud_change_in_law__subject_to_disproportionate_impact_modifier` **Question**: Do changes in law that have disproportionate impact qualify for Material Adverse Effect (MAE)? **Options**: A: No; B: Yes

Table 40 – continued from previous page

Task
Example: "Material Adverse Effect" means, with respect to SVB Financial, Boston Private or the Surviving Corporation, as the case may be, any effect, change, event, circumstance, condition, occurrence or development that, either individually or in the aggregate, has had or would reasonably be expected to have a material adverse effect on (i) the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries taken as a whole (provided, however, that, with respect to this clause (i), Material Adverse Effect shall not be deemed to include the impact of (A) changes, after the date hereof, in U.S. generally accepted accounting principles ("GAAP") or applicable regulatory accounting requirements, (B) changes, after the date hereof, in laws, rules or regulations of general applicability to companies in the industries in which such party and its Subsidiaries operate, or interpretations thereof by courts or Governmental Entities, (C) changes, after the date hereof, in global, national or regional political conditions (including the outbreak of war or acts of terrorism) or in economic or market (including equity, credit and debt markets, as well as changes in interest rates) conditions affecting the financial services industry generally and not specifically relating to such party or its Subsidiaries, (D) changes, after the date hereof, resulting from hurricanes, earthquakes, tornados, floods or other natural disasters or from any outbreak of any disease or other public health event (including the COVID-19 pandemic and the implementation of the Pandemic Measures), (E) public disclosure or consummation of the transactions contemplated hereby or actions expressly required by this Agreement or that are taken with the prior written consent of the other party in contemplation of the transactions contemplated hereby (it being understood and agreed that this clause (E) shall not apply with respect to any representation or warranty that is intended to address the consequences of the execution, announcement or performance of this Agreement or consummation of the Merger) or (F) the failure, in and of itself, to meet earnings projections or financial forecasts, but not including the underlying causes thereof; except, with respect to subclause (A), (B), (C) or (D), to the extent that the effects of such change are disproportionately adverse to the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole, as compared to similar companies in the industry in which such party and its Subsidiaries operate); or (ii) the ability of such party to timely consummate the transactions contemplated hereby. (Page 20) **Answer**: B
Task name: `maud_changes_in_gaap_or_other_accounting_principles__subject_to_disproportionate_impact_modifier` **Question**: Do changes in GAAP or other accounting principles that have disproportionate impact qualify for Material Adverse Effect (MAE)? **Options**: A: No; B: Yes **Example**: "Company Material Adverse Effect" shall mean any state of facts, circumstance, condition, event, change, development, occurrence, result, effect, action or omission (each, an "Effect") that, individually or in the aggregate with any one or more other Effects, (i) results in a material adverse effect on the business, financial condition or results of operations of the Company and its Subsidiaries, taken as a whole or (ii) prevents, materially impairs, materially impedes or materially delays the consummation of the Merger and the other transactions contemplated hereby on or before the End Date; provided, however, that with respect to clause (i) only, no Effect to the extent resulting or arising from any of the following, shall, to such extent, be deemed to constitute, or be taken into account in determining the occurrence of, a Company Material Adverse Effect: (A) general economic, political, business, financial or market conditions; (B) any outbreak, continuation or escalation of any military conflict, declared or undeclared war, armed hostilities, or acts of foreign or domestic terrorism; (C) any pandemic (including the SARS-CoV-2 virus and COVID-19 disease), epidemic, plague, or other outbreak of illness or public health event, hurricane, flood, tornado, earthquake or other natural disaster or act of God; (D) any failure by the Company or any of its Subsidiaries to meet any internal or external projections or forecasts or any decline in the price or trading volume of Company Common Stock (but excluding, in each case, the underlying causes of such failure or decline, as applicable, which may themselves constitute or be taken into account in determining whether there has been, or would be, a Company Material Adverse Effect); (E) the public announcement or pendency of the Merger and the other transactions contemplated hereby; (F) changes in applicable Legal Requirements; (G) changes in GAAP or any other applicable accounting standards; or (H) any action expressly required to be taken by the Company pursuant to the terms of the Agreement or at the express written direction or consent of Parent; provided, further, that any Effect relating to or arising out of or resulting from any change or event referred to in clause (A), (B), (C), (F) or (G) above may constitute, and be taken into account in determining the occurrence of, a Company Material Adverse Effect to the extent that such change or event has a disproportionate impact (but solely to the extent of such disproportionate impact) on the Company and its Subsidiaries as compared to other participants that operate in the industry in which the Company and its Subsidiaries operate. (Pages 87-88) **Answer**: B
Task name: `maud_pandemic_or_other_public_health_event_specific_reference_to_pandemic-related_governmental_responses_or_measures` **Question**: Is there specific reference to pandemic-related governmental responses or measures in the clause that qualifies pandemics or other public health events for Material Adverse Effect (MAE)? **Options**: A: No; B: Yes

Table 40 – continued from previous page

Task
Example: "Company Material Adverse Effect" shall mean any state of facts, circumstance, condition, event, change, development, occurrence, result, effect, action or omission (each, an "Effect") that, individually or in the aggregate with any one or more other Effects, (i) results in a material adverse effect on the business, financial condition or results of operations of the Company and its Subsidiaries, taken as a whole or (ii) prevents, materially impairs, materially impedes or materially delays the consummation of the Merger and the other transactions contemplated hereby on or before the End Date; provided, however, that with respect to clause (i) only, no Effect to the extent resulting or arising from any of the following, shall, to such extent, be deemed to constitute, or be taken into account in determining the occurrence of, a Company Material Adverse Effect: (A) general economic, political, business, financial or market conditions; (B) any outbreak, continuation or escalation of any military conflict, declared or undeclared war, armed hostilities, or acts of foreign or domestic terrorism; (C) any pandemic (including the SARS-CoV-2 virus and COVID-19 disease), epidemic, plague, or other outbreak of illness or public health event, hurricane, flood, tornado, earthquake or other natural disaster or act of God; (D) any failure by the Company or any of its Subsidiaries to meet any internal or external projections or forecasts or any decline in the price or trading volume of Company Common Stock (but excluding, in each case, the underlying causes of such failure or decline, as applicable, which may themselves constitute or be taken into account in determining whether there has been, or would be, a Company Material Adverse Effect); (E) the public announcement or pendency of the Merger and the other transactions contemplated hereby; (F) changes in applicable Legal Requirements; (G) changes in GAAP or any other applicable accounting standards; or (H) any action expressly required to be taken by the Company pursuant to the terms of the Agreement or at the express written direction or consent of Parent; provided, further, that any Effect relating to or arising out of or resulting from any change or event referred to in clause (A), (B), (C), (F) or (G) above may constitute, and be taken into account in determining the occurrence of, a Company Material Adverse Effect to the extent that such change or event has a disproportionate impact (but solely to the extent of such disproportionate impact) on the Company and its Subsidiaries as compared to other participants that operate in the industry in which the Company and its Subsidiaries operate. (Pages 87-88) **Answer**: A
Task name: `maud_pandemic_or_other_public_health_event__subject_to_disproportionate_impact_modifier` **Question**: Do pandemics or other public health events have to have disproportionate impact to qualify for Material Adverse Effect (MAE)? **Options**: A: No; B: Yes **Example**: "Material Adverse Effect" means, with respect to BancorpSouth, Cadence or the Surviving Entity, as the case may be, any effect, change, event, circumstance, condition, occurrence or development that, either individually or in the aggregate, has had or would reasonably be expected to have a material adverse effect on (i) the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries taken as a whole (provided, however, that, with respect to this clause (i), Material Adverse Effect shall not be deemed to include the impact of (A) changes, after the date hereof, in U.S. generally accepted accounting principles ("GAAP") or applicable regulatory accounting requirements, (B) changes, after the date hereof, in laws, rules or regulations (including the Pandemic Measures) of general applicability to companies in the industries in which such party and its Subsidiaries operate, or interpretations thereof by courts or Governmental Entities (as defined below), (C) changes, after the date hereof, in global, national or regional political conditions (including the outbreak of war or acts of terrorism) or in economic or market (including equity, credit and debt markets, as well as changes in interest rates) conditions affecting the financial services industry generally and not specifically relating to such party or its Subsidiaries (including any such changes arising out of a Pandemic or any Pandemic Measures), (D) changes, after the date hereof, resulting from hurricanes, earthquakes, tornados, floods or other natural disasters or from any outbreak of any disease or other public health event (including a Pandemic), (E) public disclosure of the execution of this Agreement, public disclosure or consummation of the transactions contemplated hereby (including any effect on a party's relationships with its customers or employees) or actions expressly required by this Agreement or that are taken with the prior written consent of the other party in contemplation of the transactions contemplated hereby, or (F) a decline in the trading price of a party's common stock or the failure, in and of itself, to meet earnings projections or internal financial forecasts (it being understood that the underlying causes of such decline or failure may be taken into account in determining whether a Material Adverse Effect has occurred), except to the extent otherwise excepted by this proviso); except, with respect to subclauses (A), (B), (C), or (D) to the extent that the effects of such change are materially disproportionately adverse to the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole, as compared to other companies in the industry in which such party and its Subsidiaries operate), or (ii) the ability of such party to timely consummate the transactions contemplated hereby. (Page 19) **Answer**: B
Task name: `maud_relational_language_(mae)_applies_to` **Question**: What carveouts pertaining to Material Adverse Effect (MAE) does the relational language apply to? **Options**: A: All MAE carveouts; B: No; C: Some MAE carveouts

Table 40 – continued from previous page

Task
Example: “Material Adverse Effect” means, with respect to Huntington, TCF or the Surviving Corporation, as the case may be, any effect, change, event, circumstance, condition, occurrence or development that, either individually or in the aggregate, has had or would reasonably be likely to have a material adverse effect on (i) the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole (provided, however, that, with respect to this clause (i), Material Adverse Effect shall not be deemed to include the impact of (A) changes, after the date hereof, in U.S. generally accepted accounting principles (“GAAP”) or applicable regulatory accounting requirements, (B) changes, after the date hereof, in laws, rules or regulations (including the Pandemic Measures) of general applicability to companies in the industries in which such party and its Subsidiaries operate, or interpretations thereof by courts or Governmental Entities, (C) changes, after the date hereof, in global, national or regional political conditions (including the outbreak of war or acts of terrorism) or in economic or market (including equity, credit and debt markets, as well as changes in interest rates) conditions affecting the financial services industry generally and not specifically relating to such party or its Subsidiaries (including any such changes arising out of the Pandemic or any Pandemic Measures), (D) changes, after the date hereof, resulting from hurricanes, earthquakes, tornados, floods or other natural disasters or from any outbreak of any disease or other public health event (including the Pandemic), (E) public disclosure of the execution of this Agreement, public disclosure or consummation of the transactions contemplated hereby (including any effect on a party’s relationships with its customers or employees) (it being understood that the foregoing shall not apply for purposes of the representations and warranties in Sections 3.3(b), 3.4, 4.3(b) or 4.4) or actions expressly required by this Agreement or that are taken with the prior written consent of the other party in contemplation of the transactions contemplated hereby, or (F) a decline in the trading price of a party’s common stock or the failure, in and of itself, to meet earnings projections or internal financial forecasts, but not, in either case, including any underlying causes thereof; except, with respect to subclauses (A), (B), (C) or (D), to the extent that the effects of such change are materially disproportionately adverse to the business, properties, assets, liabilities, results of operations or financial condition of such party and its Subsidiaries, taken as a whole, as compared to other companies in the industry in which such party and its Subsidiaries operate) or (ii) the ability of such party to timely consummate the transactions contemplated hereby. (Page 18) **Answer**: C
Task name: `maud_knowledge_definition` **Question**: What counts as Knowledge? **Options**: A: Actual knowledge; B: Constructive knowledge **Example**: provided, however, that with respect to clause (i) only, no Effect to the extent resulting or arising from any of the following, shall <omitted> be deemed to constitute <omitted> a Company Material Adverse Effect (Pages 87-88) **Answer**: B
Task name: `maud_buyer_consent_requirement_(ordinary_course)` **Question**: In case the Buyer’s consent for the acquired company’s ordinary business operations is required, are there any limitations on the Buyer’s right to condition, withhold, or delay their consent? **Options**: A: Yes. Consent may not be unreasonably withheld, conditioned or delayed.; B: No. **Example**: Section 5.1 Interim Operations of the Company and Parent. (a) From the date of this Agreement and until the Effective Time or the earlier termination of this Agreement in accordance with its terms, except as (v) otherwise expressly contemplated by this Agreement, (w) set forth in the applicable subsection of Section 5.1 of the Company Disclosure Letter (it being agreed that disclosure of any item in any subsection of Section 5.1 of the Company Disclosure Letter shall be deemed disclosure with respect to any other subsection of Section 5.1 of the Company Disclosure Letter only to the extent that the relevance of such item to such subsection is reasonably apparent on its face), (x) required by applicable Law, (y)(A) required to comply with COVID-19 Measures or otherwise taken (or not taken) by the Company or any of its Subsidiaries reasonably and in good faith to respond to COVID-19 or COVID-19 Measures or (B) taken (or not taken) by the Company or any of its Subsidiaries reasonably and in good faith to respond to any other extraordinary event that was not reasonably foreseeable as of the date of this Agreement and occurring after the date of this Agreement that is outside of the control of the Company or its Affiliates and is outside of the ordinary course of business of the Company and its Subsidiaries and Joint Ventures (and is not related to a Company Takeover Proposal); provided that prior to taking any actions in reliance on this clause (y), which would otherwise be prohibited by any provision of this Agreement, the Company will use commercially reasonable efforts to provide advance notice to and consult with Parent (if reasonably practicable) with respect thereto or (z) consented to in writing by Parent (which consent shall not be unreasonably withheld, conditioned or delayed), the Company shall, and shall cause each of its Subsidiaries to, use its commercially reasonable efforts to conduct its business in all material respects in the ordinary course of business consistent with past practice and in compliance in all material respects with all material applicable Laws, and shall, and shall cause each of its Subsidiaries to, use its commercially reasonable efforts to preserve intact its present business organization, keep available the services of its directors, officers and employees and maintain existing relations and goodwill with customers, distributors, lenders, partners (including Joint Venture partners and others with similar relationships), suppliers and others having material business associations with it or its Subsidiaries; (Pages 40-41) **Answer**: A
Task name: `maud_includes_consistent_with_past_practice` **Question**: Does the wording of the Efforts Covenant clause include “consistent with past practice”? **Options**: A: No; B: Yes

Table 40 – continued from previous page

Task
Example: 5.2 Operation of the Acquired Corporations' Business. (a) During the Pre-Closing Period, except (w) as required or otherwise contemplated under this Agreement or as prohibited or required by applicable Legal Requirements, (x) with the written consent of Parent (which consent shall not be unreasonably withheld, delayed or conditioned, and provided that no consent shall be required if the Company reasonably believes after consulting with outside legal counsel that seeking such consent would violate Antitrust Law), (y) for any action required to be or reasonably taken, or omitted to be taken, pursuant to any COVID-19 Measures or which is otherwise required or reasonably taken, or omitted to be taken, in response to COVID-19 or any other pandemic, epidemic or disease outbreak, as determined by the Company in its reasonable discretion, or (z) as set forth in Section 5.2 of the Company Disclosure Schedule, the Company shall, and shall cause each Acquired Corporation to, use commercially reasonable efforts to conduct its business and operations in the ordinary course in all material respects (Page 41) **Answer**: A
Task name: `maud_ordinary_course_efforts_standard` **Question**: What is the efforts standard? **Options**: A: Commercially reasonable efforts; B: Flat covenant (no efforts standard); C: Reasonable best efforts **Example**: "Ordinary Course of Business" means, with respect to any Person, the conduct of such Person's business that is consistent with the past practices of such Person prior to the date of this Agreement and taken in the ordinary course of normal, day-to-day operations of such Person, but excluding any conduct that would reasonably be expected to violate applicable Law in any material respect. <omitted> 7.1. Interim Operations. (a) The Company shall, and shall cause each of its Subsidiaries to, from and after the date of this Agreement until the earlier of the Effective Time and the termination of this Agreement pursuant to Article IX (unless Parent shall otherwise approve in writing (such approval not to be unreasonably withheld, conditioned or delayed), and except as otherwise expressly required by this Agreement or as required by a Governmental Entity or applicable Law and any Material Contract in effect prior to the date of this Agreement), conduct its business in the Ordinary Course of Business (Page 66) **Answer**: B
Task name: `maud_application_of_buyer_consent_requirement_(negative_interim_covenant)` **Question**: What negative covenants does the requirement of Buyer consent apply to? **Options**: A: Applies only to specified negative covenants; B: Applies to all negative covenants **Example**: Except (w) with respect to the Specified Exceptions (other than as applied to Section 5.1(a), Section 5.1(b), or Section 5.1(k)), (x) 25 as otherwise expressly contemplated or permitted by this Agreement, (y) as set forth in Section 5.1 of the Company Disclosure Schedule, or (z) with the Parent's consent (which shall not be unreasonably withheld, conditioned or delayed), during the Pre-Closing Period the Company shall not, and shall not permit any of its Subsidiaries to, directly or indirectly, do any of the following: (Pages 29-30) **Answer**: B
Task name: `maud_fiduciary_exception__board_determination_standard` **Question**: Under what circumstances could the Board take actions on a different acquisition proposal notwithstanding the no-shop provision? **Options**: A: If failure to take actions would lead to "breach" of fiduciary duties; B: If failure to take actions would be "inconsistent" with fiduciary duties; C: If failure to take actions would lead to "reasonably likely/expected breach" of fiduciary duties; D: If failure to take actions would lead to "reasonably likely/expected to be inconsistent" with fiduciary duties; E: If failure to take actions would lead to "reasonably likely/expected violation" of fiduciary duties; F: If taking such actions is "required to comply" with fiduciary duties; G: If failure to take actions would lead to "violation" of fiduciary duties; H: Under no circumstances could the Board do so.; I: Other circumstances **Example**: Section 5.4 No Company Solicitation. <omitted> (b) Notwithstanding anything in Section 5.4(a) to the contrary, until the Company Stockholder Approval is obtained, if the Company receives a bona fide written Alternative Acquisition Proposal made after the date hereof that does not result from a material breach of this Section 5.4, and the Company Board determines in good faith (after consultation with outside legal counsel and a nationally recognized financial advisor) that such Alternative Acquisition Proposal is, or could reasonably be expected to lead to, a Superior Acquisition Proposal, (i) the Company may negotiate and enter into an Acceptable Confidentiality Agreement with the Person making such Alternative Acquisition Proposal; provided, that the Company shall promptly (and in no event later than twenty-four (24) hours after execution thereof) deliver a copy of such Acceptable Confidentiality Agreement to Parent, (ii) following entry into such Acceptable Confidentiality Agreement by the Company, the Company and its Representatives may provide information (including nonpublic information) subject to such executed Acceptable Confidentiality Agreement; provided, that any nonpublic information provided to such Person, including if posted to an electronic data room, shall be provided to Parent prior to or substantially concurrently with the time it is provided to such Person, and (iii) the Company and its Representatives may engage in discussion or negotiations for such Alternative Acquisition Proposal with such Person and its Representatives. (Page 59) **Answer**: H
Task name: `maud_fiduciary_exception_board_determination_trigger_(no_shop)` **Question**: What type of offer could the Board take actions on notwithstanding the no-shop provision? **Options**: A: Acquisition Proposal only; B: Superior Offer, or Acquisition Proposal reasonably likely/expected to result in a Superior Offer

Table 40 – continued from previous page

Task
Example: SECTION 5.02. Acquisition Proposals. <omitted> (c) Information Exchange; Discussions or Negotiation. Notwithstanding anything to the contrary contained in Section 5.02(a), prior to obtaining the Company Requisite Vote, in the event that the Company, any of its Subsidiaries or its or their Representatives receive from any Person, after the date of this Agreement, an unsolicited, bona fide written Acquisition Proposal that did not result from a breach of this Section 5.02, and that the Company Board determines in good faith, after consultation with its financial advisors and outside legal counsel, is, or is reasonably likely to lead to, a Superior Proposal, the Company may (i) furnish or provide information to the Person making such Acquisition Proposal and its Representatives pursuant to an Acceptable Confidentiality Agreement; provided, however, that the Company shall as promptly as is reasonably practicable (and in any event within one (1) Business Day) make available to Parent and Merger Sub any written material non-public information concerning the Company or its Subsidiaries that is provided to any Person pursuant to this Section 5.02(c)(i), to the extent such information was not previously made available to Parent, Merger Sub or their Representatives, and (ii) engage in discussions and negotiations with such Person and its Representatives with respect to such Acquisition Proposal. (Page 35) **Answer**: B
Task name: `maud_cor_permitted_with_board_fiduciary_determination_only` **Question**: Is Change of Recommendation permitted as long as the board determines that such change is required to fulfill its fiduciary obligations? **Options**: A: No; B: Yes **Example**: SECTION 5.3 No Solicitation by the Company; Company Recommendation. <omitted> (d) <omitted> Notwithstanding the foregoing or any other provision of this Agreement to the contrary, prior to the time the Company Stockholder Approval is obtained (but not thereafter), the Company Board or the Company Special 41 Committee may make a Company Adverse Recommendation Change if either (x) in the case of a Company Adverse Recommendation Change made in response to a Company Acquisition Proposal, the Company Board or the Company Special Committee has determined in good faith, after consultation with its outside financial advisors and outside legal counsel, that such Company Acquisition Proposal constitutes a Company Superior Proposal and that failure to take such action would reasonably be expected to be inconsistent with the directors' fiduciary duties under applicable Law or (y) in the case of a Company Adverse Recommendation Change made in response to a Company Intervening Event, the Company Board or the Company Special Committee has determined in good faith, after consultation with its outside financial advisors and outside legal counsel, that, as a result of a Company Intervening Event, the failure to take such action would reasonably be expected to be inconsistent with its fiduciary duties under applicable Law; (Pages 46-47) **Answer**: A
Task name: `maud_cor_standard_(superior_offer)` **Question**: What standard should the board follow when determining whether to change its recommendation in connection with a superior offer? **Options**: A: "Breach" of fiduciary duties; B: "Inconsistent" with fiduciary duties; C: "Reasonably likely/expected breach" of fiduciary duties; D: "Reasonably likely/expected to be inconsistent" with fiduciary duties; E: "Reasonably likely/expected violation" of fiduciary duties; F: "Required to comply" with fiduciary duties; G: "Violation" of fiduciary duties; H: More likely than not violate fiduciary duties; I: None; J: Other specified standard **Example**: Section 5.2. No Solicitation. <omitted> (c) Notwithstanding anything to the contrary contained in this Agreement, at any time prior to obtaining the Company Stockholder Approval, the Company Board may make a Change in Recommendation in response to an unsolicited bona fide written Acquisition Proposal or cause the Company to enter into an Alternative Acquisition Agreement concerning an Acquisition Proposal, in each case only if: (i) such Acquisition Proposal or Superior Proposal did not result from a breach of Section 5.2(a); (ii)the Company Board (or a committee thereof) determines in good faith (A) after consultation with the Company's outside legal counsel and Independent Financial Advisor, that such Acquisition Proposal constitutes a Superior Proposal and (B) after consultation with the Company's outside legal counsel, that in light of such Acquisition Proposal, a failure to make a Change in Recommendation or to cause the Company to enter into such Alternative Acquisition Agreement would be inconsistent with the Company Board's fiduciary obligations to the Company's stockholders under the DGCL; (Page 27) **Answer**: B
Task name: `maud_cor_permitted_in_response_to_intervening_event` **Question**: Is Change of Recommendation permitted in response to an intervening event? **Options**: A: No; B: Yes

Table 40 – continued from previous page

Task
Example: 6.1 No Solicitation. <omitted> Notwithstanding the foregoing or anything to the contrary set forth in this Agreement (including the provisions of this Section 6.1), at any time prior to receipt of the Company Stockholder Approval, the Company Board may effect a Company Board Recommendation Change in response to a Superior Proposal or an Intervening Event if: (i) the Company Board shall have determined in good faith (after consultation with outside counsel and outside financial advisor) that the failure to effect a Company Board Recommendation Change would be reasonably likely to be inconsistent with its fiduciary obligations under applicable law; (ii) so long as the Company and its Subsidiaries are not in material breach of their obligations pursuant to this Section 6.1 with respect to an Acquisition Proposal underlying such Company Board Recommendation Change; (iii) the Company has notified the Parent in writing that it intends to effect a Company Board Recommendation Change, describing in reasonable detail the reasons for such Company Board Recommendation Change (a "Recommendation Change Notice") (it being understood that the Recommendation Change Notice shall not constitute a Company Board Recommendation Change or a Trigger Event for purposes of this Agreement); (iv) if requested by the Parent, the Company shall have made its Representatives available to negotiate (to the extent that Parent desires to so negotiate) with the Parent's Representatives any proposed modifications to the terms and conditions of this Agreement during the three (3) Business Day period following delivery by the Company to the Parent of such Recommendation Change Notice; and (v) if the Parent shall have delivered to the Company a written, binding and irrevocable offer to alter the terms or conditions of this Agreement during such three (3) Business Day period, the Company Board shall have determined in good faith (after consultation with outside counsel), after considering the terms of such offer by the Parent, that the failure to effect a Company Board Recommendation Change would still be reasonably likely to be inconsistent with its fiduciary obligations under applicable law; provided, however, that in the event of any material revisions to an Acquisition Proposal underlying a potential Company Board Recommendation Change, the Company will be required to notify Parent of such revisions and the applicable three (3) Business Day period described above shall be extended until two (2) Business Days after the time Parent receives notification from the Company of such revisions. (Page 34) **Answer**: B
Task name: `maud_cor_standard_(intervening_event)` **Question**: What standard should the board follow when determining whether to change its recommendation in response to an intervening event? **Options**: A: "Breach" of fiduciary duties; B: "Inconsistent" with fiduciary duties; C: "Reasonably likely/expected breach" of fiduciary duties; D: "Reasonably likely/expected to be inconsistent" with fiduciary duties; E: "Reasonably likely/expected violation" of fiduciary duties; F: "Required to comply" with fiduciary duties; G: "Violation" of fiduciary duties; H: More likely than not violate fiduciary duties; I: Other specified standard **Example**: 6.3 Shareholders' Approval and Stockholder Approval. <omitted> (c) <omitted> if the Board of Directors of <omitted> the Company, after receiving the advice of its outside counsel and, with respect to financial matters, its outside financial advisors, determines in good faith that it would more likely than not result in a violation of its fiduciary duties under applicable law to make or continue to make the Parent Board Recommendation or the Company Board Recommendation, as applicable, such Board of Directors may <omitted> submit this Agreement to its shareholders or stockholders, respectively, without recommendation (which, for the avoidance of doubt, shall constitute a Recommendation Change) (Page 57) **Answer**: I
Task name: `maud_initial_matching_rights_period_(cor)` **Question**: How long is the initial matching rights period in case the board changes its recommendation? **Options**: A: 2 business days or less; B: 3 business days; C: 3 calendar days; D: 4 business days; E: 4 calendar days; F: 5 business days; G: Greater than 5 business days **Example**: 6.3 No Solicitation by Golden. <omitted> in response to a <omitted> Golden Competing Proposal <omitted> the Golden Board may effect a Golden Change of Recommendation; provided, however, that such a Golden Change of Recommendation may not be made unless and until: <omitted>; provided that in the event of any material amendment or material modification to any Golden Superior Proposal <omitted> , Golden shall be required to deliver a new written notice to Labrador and to comply with the requirements of this Section 6.3(e)(iv) with respect to such new written notice, except that the advance written notice obligation set forth in this Section 6.3(e)(iv) shall be reduced to two Business Days (Pages 34-35) **Answer**: D
Task name: `maud_additional_matching_rights_period_for_modifications_(cor)` **Question**: How long is the additional matching rights period for modifications in case the board changes its recommendation? **Options**: A: 2 business days or less; B: 3 business days; C: 3 days; D: 4 business days; E: 5 business days; F: > 5 business days; G: None

Table 40 – continued from previous page

Task
Example: Section 5.4 Non-Solicitation. <omitted> (b) <omitted> Notwithstanding the foregoing, at any time prior to obtaining the East Stockholder Approval, and subject to East's compliance in all material respects at all times with the provisions of this Section 5.4 and Section 5.3, in response to a Superior Proposal with respect to East that was not initiated, solicited, knowingly encouraged or knowingly facilitated by East or any of the East Subsidiaries or any of their respective Representatives, the East Board may make an East Adverse Recommendation Change; provided, however, that East shall not be entitled to exercise its right to make an East Adverse Recommendation Change in response to a Superior Proposal with respect to East (x) until three (3) Business Days after East provides written notice to Central (an "East Notice") advising Central that the East Board or a committee thereof has received a Superior Proposal, specifying the material terms and conditions of such Superior Proposal, and identifying the Person or group making such Superior Proposal, (y) if during such three (3) Business Day period, Central proposes any alternative transaction (including any modifications to the terms of this Agreement), unless the East Board determines in good faith (after consultation with East's financial advisors and outside legal counsel, and taking into account all financial, legal, and regulatory terms and conditions of such alternative transaction proposal, including any conditions to and expected timing of consummation, and any risks of non-consummation of such alternative transaction proposal) that such alternative transaction proposal is not at least as favorable to East and its stockholders as the Superior Proposal (it being understood that any change in the financial or other material terms of a Superior Proposal shall require a new East Notice and a new two (2) Business Day period under this Section 5.4(b)) and (z) unless the East Board, after consultation with outside legal counsel, determines that the failure to make an East Adverse Recommendation Change would be inconsistent with its fiduciary duties. (Page 76) **Answer**: A
Task name: `maud_definition_includes_stock_deals` **Question**: What qualifies as a superior offer in terms of stock deals? **Options**: A: "All or substantially all"; B: 50%; C: Greater than 50% but not "all or substantially all"; D: Less than 50% **Example**: 5.4 No Solicitation by the Company; Other Offers. <omitted> the Company shall not be entitled to: (i) make a Change in Company Board Recommendation <omitted> unless: <omitted> the Company shall have first provided prior <omitted> notice to Parent that it is prepared to <omitted> make a Change in Company Board Recommendation (a "Recommendation Change Notice") <omitted> Any material changes with respect to the Intervening Event <omitted> or material changes to the financial terms of such Superior Proposal <omitted> shall require the Company to provide to Parent a new Recommendation Change Notice <omitted> and a new three (3) Business Day period. (Pages 45-46) **Answer**: C
Task name: `maud_definition_includes_asset_deals` **Question**: What qualifies as a superior offer in terms of asset deals? **Options**: A: "All or substantially all"; B: 50%; C: Greater than 50% but not "all or substantially all"; D: Less than 50% **Example**: Section 5.4 Acquisition Proposals. <omitted> (d) <omitted> following receipt of a <omitted> Acquisition Proposal <omitted> that the Company Board determines <omitted> constitutes a Superior Proposal, the Company Board may <omitted> make an Adverse Recommendation Change <omitted> if <omitted> (i) (A) the Company shall have provided to Parent <omitted> notice, <omitted> (it being understood and agreed that any amendment to the financial terms or any other material term or condition of such Superior Proposal shall require a new notice and an additional three Business Day period) (Pages 44-45) **Answer**: B
Task name: `maud_financial_point_of_view_is_the_sole_consideration` **Question**: Is "financial point of view" the sole consideration when determining whether an offer is superior? **Options**: A: No; B: Yes **Example**: 5.4 No Solicitation by the Company; Other Offers. <omitted> the Company shall not be entitled to: (i) make a Change in Company Board Recommendation <omitted> unless: <omitted> the Company shall have first provided prior <omitted> notice to Parent that it is prepared to <omitted> make a Change in Company Board Recommendation (a "Recommendation Change Notice") <omitted> Any material changes with respect to the Intervening Event <omitted> or material changes to the financial terms of such Superior Proposal <omitted> shall require the Company to provide to Parent a new Recommendation Change Notice <omitted> and a new three (3) Business Day period. (Pages 45-46) **Answer**: A
Task name: `maud_definition_contains_knowledge_requirement_-_answer` **Question**: What is the knowledge requirement in the definition of "Intervening Event"? **Options**: A: Known, but consequences unknown or not reasonably foreseeable, at signing; B: Known, but consequences unknown, at signing; C: Not known and not reasonably foreseeable at signing; D: Not known at signing

Table 40 – continued from previous page

Task
Example: “Acquisition Proposal” means any inquiry, proposal or offer from any Person or group of Persons other than Parent or one of its Subsidiaries made after the date of this Agreement relating to (A) a merger, reorganization, consolidation, share purchase, share exchange, business combination, recapitalization, liquidation, dissolution, joint venture, partnership, spin-off, extraordinary dividend or similar transaction involving the Company or any of its Subsidiaries, which is structured to permit such Person or group of Persons to, directly or indirectly, acquire beneficial ownership of 20% or more of the outstanding equity securities of the Company, or 20% or more of the consolidated net revenues, net income or total assets of the Company and its Subsidiaries, taken as a whole or (B) the acquisition in any manner, directly or indirectly, of over 20% of the equity securities or consolidated total assets of the Company and its Subsidiaries, in each case other than the Merger and the other transactions contemplated by this Agreement. <omitted> “Superior Proposal” means any bona fide written Acquisition Proposal (A) on terms which the Company Board determines in good faith, after consultation with its outside legal counsel and financial advisors, to be more favorable from a financial point of view to the holders of Shares than the Merger and the other transactions contemplated by this Agreement, taking into account all the terms and conditions of such proposal and this Agreement and (B) that the Company Board determines in good faith is capable of being completed, taking into account all financial, regulatory, legal and other aspects of such proposal; provided, that for purposes of the definition of “Superior Proposal,” the references to “20%” in the definition of Acquisition Proposal shall be deemed to be references to “50%.” (Page 47) **Answer**: A
Task name: `maud_intervening_event_-_required_to_occur_after_signing_-_answer` **Question**: Is an “Intervening Event” required to occur after signing? **Options**: A: No. It may occur or arise prior to signing.; B: Yes. It must occur or arise after signing. **Example**: “Superior Proposal” shall mean, with respect to a party hereto, any <omitted> Acquisition Proposal with respect to such party made by a third party to acquire, directly or indirectly, pursuant to a tender offer, exchange offer, merger, share exchange, consolidation or other business combination, (A) all or substantially all of the assets of such party and its Subsidiaries, taken as a whole, (Page 120) **Answer**: A
Task name: `maud_initial_matching_rights_period_(ftr)` **Question**: How long is the initial matching rights period in connection with the Fiduciary Termination Right (FTR)? **Options**: A: 2 business days or less; B: 3 business days; C: 3 calendar days; D: 4 business days; E: 4 calendar days; F: 5 business days; G: 5 calendar days; H: Greater than 5 business days **Example**: SECTION 5.3 No Solicitation by the Company; Company Recommendation. <omitted> (d) <omitted> provided, however, that the Company Board and the Company Special Committee shall not, and shall cause the Company not to, make a Company Adverse Recommendation Change in connection with a Company Superior Proposal unless (I) the Company has given Parent at least four (4) Business Days’ prior written notice of its intention to take such action (which notice shall reasonably describe the material terms of the Company Superior Proposal or attach the agreement and all material related documentation providing for such Company Superior Proposal), (II) the Company has negotiated, and has caused its Representatives to negotiate, in good faith with Parent during such notice period, to the extent Parent wishes to negotiate, to enable Parent to propose in writing a binding offer to effect revisions to the terms of this Agreement such that it would cause such Company Superior Proposal to no longer constitute a Company Superior Proposal, (III) following the end of such notice period, the Company Board or the Company Special Committee shall have considered in good faith any such binding offer from Parent, and shall have determined that the Company Superior Proposal would continue to constitute a Company Superior Proposal if the revisions proposed in such binding offer were to be given effect and (IV) in the event of any material change to the material terms of such Company Superior Proposal, the Company shall, in each case, have delivered to Parent an additional notice consistent with that described in clause (I) above and the notice period shall have recommenced, except that the notice period shall be at least two (2) Business Days (rather than the four (4) Business Days otherwise contemplated by clause (I) above); (Page 47) **Answer**: D
Task name: `maud_tail_period_length` **Question**: How long is the Tail Period? **Options**: A: 12 months or longer; B: Other; C: within 12 months; D: within 6 months; E: within 9 months **Example**: Section 7.3 Termination Fees. <omitted> (b) <omitted> if <omitted> Parent or the Company terminates this Agreement <omitted> (iii) <omitted> the Company shall have consummated an Alternative Acquisition Proposal or entered into an Alternative Acquisition Agreement for any Alternative Acquisition Proposal <omitted> which Alternative Acquisition Proposal is ultimately consummated (Page 80) **Answer**: C
Task name: `maud_specific_performance` **Question**: What is the wording of the Specific Performance clause regarding the parties’ entitlement in the event of a contractual breach? **Options**: A: "entitled to seek" specific performance; B: "entitled to" specific performance

Table 40 – continued from previous page

Task
Example: Section 9.10 Specific Performance. The parties hereto hereby agree that irreparable damage would occur in the event that any provision of this Agreement were not performed in accordance with its specific terms or were otherwise breached, and that money damages or other legal remedies would not be an adequate remedy for any such damages. Accordingly, the parties acknowledge and agree that each party shall be entitled to, in accordance with the provisions of this Agreement, an injunction or injunctions, specific performance or other equitable relief to prevent breaches of this Agreement and/or to enforce specifically the terms and provisions hereof in any court, in addition to any other remedy to which they are entitled at law or in equity (Page 73) **Answer**: B

Significance and value Reading comprehension is a particularly challenging part of contract review, both to human and to machine. The MAUD tasks evaluate an LLM's ability to understand and categorize a wide spectrum of legal clauses in the context of merger agreements.

F.17 New York State Judicial Ethics

In LEGALBENCH, the New York State Judicial Ethics task is denoted as `nys_judicial_ethics`.

Background The New York State Advisory Committee on Judicial Ethics posts rulings on real ethical scenarios. The Committee, established in 1987, offers guidance to roughly New York State judges and justices, as well as other judicial personnel and candidates in the state. By interpreting the Rules Governing Judicial Conduct and the Code of Judicial Conduct, the Committee assists these individuals in maintaining high ethical standards. Actions taken by judges in accordance with the Committee's formal opinions are deemed proper, which helps protect them during any future investigations by the New York State Commission on Judicial Conduct.

Task 300 real-world scenarios and fact patterns have been reformulated into yes or no questions to understand whether models understand ethical rules and how they might apply to different judicial situations.

For example in a 2022 decision the Committee noted that: "A judge who previously served as the District Attorney may not preside over a parole recognizance hearing concerning a parolee or releasee who had originally been convicted and sentenced during the judge's former tenure as the District Attorney."

This is converted to a Yes/No question:

Question: Can a judge who previously served as the District Attorney preside over a parole recognizance hearing concerning a parolee or releasee who had originally been convicted and sentenced during the judge's former tenure as the District Attorney?
Answer: No

Question	Answer
Can a judge's law clerk assist the judge with any election-related matters during a general election when the judge is on-call?	No
Can a part-time town justice be employed part-time as a community school liaison with the county sheriff's office simultaneously?	No
Can an appellate judge who successfully sought to vacate a vexatious lien filed by a disgruntled litigant against their real property preside over appeals from other decisions or orders rendered by the lower court judge who granted the petition to vacate?	Yes
Can a part-time town justice serve as a part-time assistant conflict defender in the same county as their court?	Yes

Table 41: Examples for `nys_judicial_ethics`.

Construction process We collect digest statements from the New York State Unified Court System Advisory Committee on Judicial Ethics.[29] We collect samples from 2010, 2021, 2022, and 2023 and then use ChatGPT to reformulate the statements into yes or no questions. To ensure that data is not used for training OpenAI models, we opt out of data use for accounts used for task creation. We leave 2010 and 2021 data for understanding scope of data leakage from opinions being online. 2022 and 2023 data should not have been seen by most models that were trained prior to these years.

Significance and value An important part of legal practice is abiding by ethics rules. As agents become more involved in the legal process it will be important to understand not only whether they can understand and reason about rules for the public, but also whether they can reason about ethical principles and rules governing judges and lawyers.

[29] https://www.nycourts.gov/legacyhtm/ip/judicialethics/opinions/

F.18 OPP-115 Tasks

In LEGALBENCH, the OPP-115 tasks are denoted as `opp115_*`.

Background The OPP-115 Corpus, consisting of 115 online privacy policies, provides a comprehensive collection of privacy statements expressed in natural language [149]. Each of these policies has been meticulously read and annotated by a team of three law graduate students. The annotations present in the text specifically outline various data practices.

These privacy policies are classified into ten distinct categories:

1. First Party Collection/Use: This describes how and why a service provider collects user information.
2. Third Party Sharing/Collection: This explains how user information may be shared with or collected by third parties.
3. User Choice/Control: This delineates the choices and control options available to users.
4. User Access, Edit, & Deletion: This describes if and how users may access, edit, or delete their information.
5. Data Retention: This states how long user information is stored.
6. Data Security: This communicates how user information is protected.
7. Policy Change: This explains if and how users will be informed about changes to the privacy policy.
8. Do Not Track: This discusses if and how Do Not Track signals for online tracking and advertising are honored.
9. International & Specific Audiences: This focuses on practices that pertain only to a specific group of users (e.g., children)
10. Other: This encompasses additional sub-labels for introductory or general text, contact information, and practices not covered by the other categories.

Task and construction process A separate binary classification task has been created for each category, with negative samples drawn from the rest of the text. To ensure consistency, any text with less than 10 words has been eliminated. The 'other' category was not included in the categorization because it was deemed too broad and wouldn't provide much value in terms of specific classification.

Significance and value The classification task associated with the OPP-115 Corpus serves as a significant measure of an LLM's logical reasoning ability. By assigning privacy policy segments to the right categories, LLMs demonstrate their understanding and interpretation of the language and nuances within these privacy policies. Although the task may be seen as "simple" from a human legal practitioner's viewpoint, it provides an invaluable and objective gauge of an LLM's progress in logical reasoning and language comprehension.

Task/category	**Example of clause**
opp115_data_retention	The name of the domain from which you access the Internet (for example, gmail.com, if you are connecting from a Google account);
opp115_data_security	However, no system can be 100% secure and human errors occur, so there is the possibility that there could be unauthorized access to your information. By using our services, you assume this risk.
opp115_do_not_track	Do Not Track Signals Our websites do not treat browsers that send a do not track signal differently from browsers that do not send one.
opp115_first_party_collection_use	Send-a-friend: In the case of send-a-friend email or card, we only collect
opp115_international_and_specific_audiences	CalOPPA is the first state law in the nation to require commercial websites and online services to post a privacy policy. The law's reach stretches well beyond California to require a person or company in the United States (and conceivably the world) that operates websites collecting personally identifiable information from California consumers to post a conspicuous privacy policy on its website stating exactly the information being collected and those individuals with whom it is being shared, and to comply with this policy. - See more at: http://consumercal.org/california-online-privacy-protection-act-caloppa/sthash.0FdRbT51.dpuf
opp115_policy_change	If we make a significant or material change in the way we use your personal information, t
opp115_third_party_sharing_collection	We use third-party payment service providers such as Amazon.com (Privacy Policy), Stripe.com (Privacy Policy), and PayPal (Privacy Policy)
opp115_user_access,_edit_and_deletion	you can access your personal information by contacting ABITA.COM as described at the bottom of this statement, or through alternative means of access described by the service.
opp115_user_choice_control	do not wish to receive any additional marketing material, you can indicate your preference on our store partners order form.

Table 42: Examples for OPP-115 tasks.

F.19 Purpose of Oral Argument Questions

In LEGALBENCH, the Purpose of Oral Argument Questions task is denoted as `oral_argument_question_purpose`.

Background Before a court decides a case, it typically calls before it the attorneys for the parties to the lawsuit to orally present their arguments for why the case should be resolved in favor of their clients and to answer any questions that the judge or judges have of them. In modern times, however, oral argument is not lawyers' primary avenue for explaining their positions. Instead, parties submit their arguments in written form ("briefs") and use their oral argument time primarily to supplement those submissions by reiterating their key positions, clarifying areas of ambiguity, and seeking to persuade judges who are uncertain how the case should be resolved.

Although there is no universally accepted listing, judges questions at oral argument tend to fall into a few categories:[30]

- Background: A question seeking factual or procedural information that is missing or not clear in the briefing
- Clarification: A question seeking to get an advocate to clarify her position or the scope of the rule being advocated.
- Implications: A question about the limits of a rule or its implications for future cases.
- Support: A question offering support for the advocate's position.
- Criticism: A question criticizing an advocate's position.
- Communicate: A question designed primarily to communicate with one or more other judges on the court.
- Humor: A question designed to interject humor into the argument and relieve tension.

A lawyer presenting her case before a court at oral argument must be able to quickly and accurately determine why the judge has asked a particular question so as to answer it on behalf of her client in the most persuasive way possible. It is also a difficult task. Under the pressure of persistent and difficult questioning it is easy for a lawyer to misread a question and offer an unresponsive or misguided answer. Skillful lawyers learn to quickly understand not only judges' questions but the reasons they are asked.

Task The Purpose of Oral Argument Questions task requires an LLM to determine–given a question from an oral argument transcript–for which of the seven purposes above the judge asked the question.

Construction process We created a dataset of questions from U.S. Supreme Court oral argument transcripts, classified into one of the seven functions above. Questions were taken from cases argued in the 2022-23 Supreme Court term, in reverse chronological order. A question was defined as the totality of a judge's words prior to an advocate's response, regardless whether the words constituted a true interrogatory sentence. To include a sufficient number of questions of each type in the dataset, questions were not drawn at random. Instead rarer question types (e.g. humor and communication) were targeted for inclusion, questions of more common types (e.g. clarification) were frequently omitted.

Significance and value Young attorneys, and even many experienced ones, struggle with oral advocacy. It requires comfort in the courtroom, quick thinking, and careful demeanor to assess from the tone and content of a question why the judge is asking it and how most effectively to respond. In one regard, LLMs will be superior. They do not suffer from human nervousness. However, given only a text prompt rather than an audible question from which to infer tone, and given only a judge's question and not the full case context, this task would be a challenge even for a seasoned lawyer. Whether LLMs can succeed will be an extremely interesting measure of progress in legal analysis.

[30] This categorization scheme is from [150], employed and discussed along with other possible classification schemes at [46].

Question	Purpose
May I ask you a question about standing? So it's the case, isn't it, that if any party in either of these two cases has standing, then it would be permissible for us to reach the merits of the issue?	Background
I guess I don't understand that answer. In other words, is it simply adding for religious reasons to the label that would change whether it could be regulated or not?	Clarification
And we have amicus brief from different stakeholders, some saying it may not apply in parody, but it could apply in movie titles, it might apply in something else and not this, in novels, et cetera. Why should we rule broadly? And if we rule narrowly, on what basis? You heard earlier at least three alliterations, one, the – Justice Kagan's, one Justice Jackson, one me, limit this just to parodies, because parodies really do rely on is this a joke that people are going to get.	Communicate
Mr. Martinez, I think one of the problems that you have, as evidenced by a lot of the questions that you've been getting, is with the derivative works protection, you know, which, in, you know, 106(2), actually talks about transforming any other form in which a work may be recast, transformed, or adapted. And it seems to me like your test, this meaning or message test, risks stretching the concept of transformation so broadly that it kind of eviscerates Factor 1 and puts all of the emphasis on Factor 4. I mean, when you've been asked about book to movie and – and – and, you know, songs, you keep flipping to Factor 4. So, if a work is derivative, like Lord of the Rings, you know, book to movie, is your answer just like, well, sure, that's a new meaning or message, it's transformative, so all that matters is 4?	Criticism
What are your two ones that you're like killers?	Humor
So what's the limiting line of yours – of yours? Justice Kagan asked you about another website designer. But how about people who don't believe in interracial marriage or about people who don't believe that What's – where's the line? I choose to serve whom I want. If I disagree with their personal characteristics, like race or disability, I can choose not to sell to those peopledisabled people should get married?	Implications
There were several questions earlier about the justification for granting preference for foster or adoptive parents who are members of an entirely different tribe. Could you speak to that?	Support

Table 43: Examples for `oral_argument_question_purpose`

F.20 Overruling

In LEGALBENCH, the Overruling task is denoted as `overruling`.

Background A hallmark of the common-law legal system is that courts will *overrule* previous judicial decisions. The act of overruling is significant, and it indicates that the overruled decision was either in accurate in its articulation/application of a particular law, or that overruling court wishes to announce a substantive change in law.

Task In this task, an LLM is required—given an excerpt of judicial text—to determine if the text overrules a previous decision.

Construction process This task is taken from [157], which previously studied the capacity for finetuned BERT models to perform this task. Please refer to [157] for more information on the construction process.

Significance and value Identifying when judicial text overrules another case is a basic but essential lawyering skill. From a practical standpoint, the capacity for LLMs to correctly classify overruling sentences could have practical applications for the design and construction of legal opinion databases. When using or citing a case in legal arguments, lawyers must ensure that the case hasn't been an overruled, and is still "good law." Tools which automatically parse legal databases and extract cases which have been overruled would thus be helpful for constructing legal arguments.

Sentence	Overruling sentence?
brockett v. spokane arcades, inc., 472 u.s. 491, 501 (1985) (citations omitted).	No
we overrule so much of kerwin as holds that a criminal defendant is not entitled to inspect and make an analysis of the seized controlled substance.	Yes

Table 44: Examples for `overruling`.

F.21 Personal Jurisdiction

In LEGALBENCH, the Personal Jurisdiction task is denoted as `personal_jurisdiction`.

Background Personal jurisdiction refers to the ability of a particular court (e.g. a court in the Northern District of California) to preside over a dispute between a specific plaintiff and defendant. A court (sitting in a particular forum) has personal jurisdiction over a defendant only when that defendant has a relationship with the forum. We focus on a simplified version of the rule for federal personal jurisdiction, using the rule:

> There is personal jurisdiction over a defendant in the state where the defendant is domiciled, or when (1) the defendant has sufficient contacts with the state, such that they have availed themselves of the privileges of the state and (2) the claim arises out of the nexus of the defendant's contacts with the state.

Under this rule, there are two paths for a court have jurisdiction over a defendant: through domicile or through contacts.

- Domicile: A defendant is domiciled in a state if they are a citizen of the state (i.e. they live in the state). Changing residency affects a change in citizenship.
- Contacts: Alternatively, a court may exercise jurisdiction over a defendant when that defendant has *sufficient contacts* with the court's forum, and the legal claims asserted arise from the *nexus* of the defendant's contacts with the state. In evaluating whether a set of contacts are sufficient, lawyers look at the extent to which the defendant interacted with the forum, and availed themselves of the benefits and privileges of the state's laws. Behavior which usually indicates sufficient contacts include: marketing in the forum or selling/shipping products into the forum. In assessing nexus, lawyers ask if the claims brought against the defendant arise from their contacts with the forum. In short: is the conduct being litigated involve the forum or its citizens in some capacity?

Task The personal jurisdiction task requires an LLM to determine—given a fact pattern describing the events leading up to a legal claim—whether a particular court has personal jurisdiction over the defendant.

Construction process We manually construct a dataset to test application of the personal jurisdiction rule, drawing inspiration from exercises found online and in legal casebooks. Each sample in our dataset describes a "fact pattern," and asks if a court located in particular state (**A**) can exercise personal jurisdiction over an individual (**B**) named in the fact pattern. In designing the dataset, we use 5 base fact patterns, and create 4 slices, where each slice evaluates a different aspect of the personal jurisdiction rule:

- Domicile: Fact patterns where **B** is domiciled in **A**. Hence, personal jurisdiction exists.
- No contacts: Fact patterns where **B** has insufficient contacts with **A**. Hence there is no personal jurisdiction.
- Yes contacts, no nexus: Fact patterns where **B** has sufficient contacts with **A**, but the claims against **B** do not arise from those contacts. Hence, personal jurisdiction does not exist.
- Yes contacts, yes nexus: Fact patterns where **B** has sufficient contacts with **A**, and the claims against **B** arise from those contacts. Hence, there is personal jurisdiction.

Caveat. Personal jurisdiction is a rich and complex doctrine. Our dataset focuses on a narrow class of fact patterns, related to jurisdiction over individuals. We don't consider, for instance, more complex questions related to adjudicating citizenship (e.g. the *Hertz* test) or the classic stream-of-commerce problems. We leave this to future work.

Significance and value Identifying when personal jurisdiction exists is a skill that law students learn in their first-year civil procedure course. The personal jurisdiction task is interesting because applying even the simplified version of the rule requires reasoning over the degree of connection between a defendant and the forum state.

Facts	Personal jurisdiction?
Dustin is a repairman who lives in Arizona and repairs computers in California, Oregon, and Washington. Dustin is an avid skier, so his favorite place to go on vacation is Colorado. While travelling to repair a computer in Washington, Dustin is involved in a car crash in Oregon with Laura, a citizen of Oregon. After the accident, Dustin returns to Arizona. Laura sues him in Colorado.	No
David is a citizen of California. He flies to New York for a vacation, where he meets Maggie, who is also visiting from Rhode Island. While they chat, Dave fraudulently tricks Maggie into giving him her savings. David then continues his vacation and visits Texas, Oregon, Florida, and New Mexico. After he returns home, Maggie sues David for fraud in Oregon.	No
Ana is a lawyer who resides in Texas. While visiting Louisiana, she meets David, who runs a bike shop. David's bike shop is famous, and he frequently advertises his bikes in Texas newspapers. Ana buys a bike from David and rides it back home. Right after she crosses the border, the bike seat explodes, injuring Ana. Ana sues David in Texas.	Yes
Tony (from Texas) is a regional manager for a cookbook company, Tasty Eats Books (incorporated and principal place of business in Delaware). Tony's job requires him to travel from city to city to show new cookbooks to chefs. In January 2022, he was scheduled to visit restaurants in Illinois, Indiana, and Michigan. While in Michigan, Tony goes to Lake Erie to blow off some steam. He ends up getting into a fight with Arthur, a lawyer from Detroit, Michigan. Tony and Arthur each blame the other for starting the fight. Arthur sues Tony in Texas.	Yes

Table 45: Examples for `personal_jurisdiction`.

F.22 Privacy Policy Entailment

In LEGALBENCH, the Privacy Policy Entailment task is denoted as `privacy_policy_entailment`.

Background The Privacy Policy Entailment task is created from the APP-350 corpus [161], which consists of 350 Android app privacy policies annotated with different privacy practices. In this corpus, individual clauses are annotated based on whether they do or do not perform a certain practice (e.g., "We access your location information").

Task Given a clause from a privacy policy and a description of the practice, the LLM must determine if the clause describes the performance of that practice. This is analagous to an entailment task, where the premise is the clause, and the hypothesis is the practice description.

Construction process For each practice coded in the APP-350 corpus, we derive a natural language description of that practice, which serves as our "hypothesis." Each instance of this task corresponds to a triple containing a clause, a practice description, and a binary classification (Yes/No) based on whether the clause performs the practice. Across the dataset there are 57 unique policy descriptions.

Clause	Description	Performed?
We may collect and record information through the SN Service in accordance with the policies and terms of that SN Service. The information we collect when you connect your user account to an SN Service may include: (1) your name, (2) your SN Service user identification number and/or user name, (3) locale, city, state and country, (4) sex, (5) birth date, (6) email address, (7) profile picture or its URL, and (8) the SN Service user identification numbers for your friends that are also connected to Supercell's game(s).	The policy describes receiving data from an unspecified single sign on service	Yes
Your e-mail address will not be stored.	The policy describes collection of the user's e-mail by a party to the contract.	No

Table 46: Example clause-description-label pairs for Privacy Policy Entailment task.

Significance and value The privacy policy entailment task is similar to ContractNLI, in that it evaluate a LLM's capacity to perform entailment-style reasoning over formal legal language. From a lawyerly perspective, understanding whether a policy performs certain functions or empowers one of the parties to pursue practices is an essential element of legal comprehension. From a practical perspective, the ability for LLMs to perform this task could empower researchers to conduct broader studies of privacy agreements. As [161] observes, annotation cost limitations often restrict the scope of empirical studies of privacy agreements.

F.23 Privacy Policy QA

In name, the Privacy Policy QA task is denoted as `privacy_policy_qa`.

Background The Privacy Policy QA task is derived from [112], which annotated clauses in mobile application privacy policies based on whether they contain the answer to a question.

Task Given an excerpt from a privacy policy and a question, the LLM must determine whether the excerpt is relevant to answering the question or not.

Construction process We used the snippet annotations available in [112] to construct this task, removing all snippets with fewer than 10 words. Examples of excerpt/question/relevant tuples are shown in the table below. 5449 instances correspond to a relevant question-clause pair, and 5474 instances correspond to an irrelevant question-clause pair.

Excerpt	Question	Class
We also use cookies, tags, web beacons, local shared objects, files, tools and programs to keep records, store your preferences, improve our advertising, and collectNon-Identifying Information, includingDevice Dataand information about your interaction with the Site and ourBusiness Partners'web sites.	is my search and purchase history shared with advertisers?	Relevant
We collect information about the value added services you are using over Viber and/or apps (such as games) you have downloaded through Viber.	does viber sell my information to advertisers and marketers?	Irrelevant

Table 47: Examples for Privacy Policy QA

Significance and value Determining when a particular legal excerpt is relevant to answering a question is essential to interpreting legal documents. This task allows us to evaluate LLMs for this capability. From a more practical standpoint, the Privacy Policy QA task is a helpful evaluation task when developing LLM systems which involve decompositions over long documents. A common approach—in order to account for the fact that many long documents exceed ordinary context windows—is to chunk documents into smaller segments, and apply a LLM independently to filter out irrelevant segments. For QA tasks involving long policies, this task allows practitioners to measure performance for the filtering step.

F.24 Private Right of Action (PROA)

In LEGALBENCH, the Private Right of Action task is denoted as `proa`.

Background A private right of action (PROA) exists when a statute empowers an ordinary individual (i.e. a private person) to legally enforce their rights by bringing an action in court. In short, a PROA creates the ability for an individual to sue someone in order to recover damages or halt some offending conduct. PROAs are ubiquitous in antitrust law (in which individuals harmed by anti-competitive behavior can sue offending firms for compensation) and environmental law (in which individuals can sue entities which release hazardous substances for damages) [51].

Task In the PROA task, an LLM must determine if a statutory clause contains a private right of action.

Construction process We construct a dataset of PROAs by hand, drawing inspiration from clauses found in different state codes. We construct 50 clauses which do contain a PROA, and 50 clauses which do not. Clauses which do not contain a private right may either create no cause of action, or empower a non-private individual (e.g., an attorney general) to bring a claim. 5 randomly sampled clauses constitute the training set, and the remaining 95 form the test set.

Input	Answer
The attorney general or an attorney representing the state may bring an action for an injunction to prohibit a person from violating this chapter or a rule adopted under this chapter.	No
The administrator may bring an action in a court of competent jurisdiction to enforce this chapter.	No
The sheriff or the sheriff's designee shall maintain a permanent personnel file oneach department employee.	No
If any laborer, without good cause, shall abandon his or her employer before the expiration of his or her contract, he or she shall be liable to his or her employer for the full amount of any account he or she may owe his or her employer.	Yes
No employer may discharge any employee by reason of the fact that earnings have been subjected to garnishment or execution. If an employer discharges an employee in violation of this section, the employee may within ninety days of discharge bring a civil action for recovery of twice the wages lost as a result of the violation and for an order requiring reinstatement.	Yes
In addition to all other penalties, rights, or remedies provided by law, an individual or entity that uses or attempts to use its official authority or influence for the purpose of interfering with the right of a legislative employee to make a protected disclosure is liable in a civil action for damages brought by a legislative employee.	Yes

Table 48: Examples for proa

Significance and value The PROA task evaluates an LLM's ability to perform a two-step reasoning test: (1) does the statute allow a party to bring a claim in court, and (2) is that party private? Law students and legal professionals should be capable of performing this task at near-perfect accuracy.

The PROA task derives additional significance from a recent movement towards studying state statutory language [155]. Legal scholars have long been unable to conduct large scale empirical studies of state statutory language, given the sheer volume of state statutes. The ability for LLMs to accurately classify or annotate statutes could thus empower new empirical studies of state statues.

F.25 Rule QA

In LEGALBENCH, the Rule QA task is denoted as `rule_qa`.

Background Lawyers are regularly required to recall specifical legal *rules* that are drawn from cases, statutes, or other sources. Rules can take many shapes and forms. For instance, the rule pertaining to the federal requirements for a class (in a class action lawsuit) are codified in Rule 23(a) of the Federal Rules of Civil Procedure and are simply known as the need for "numerosity, commonality, typicality, and adequacy."

Task The Rule QA task evaluates a LLM's ability to answer questions on different legal rules. The rules are drawn from subjects typically studied in the first year of law school (e.g., civil procedure, constitutional law, etc.). This is an open-generation task.

Construction process We manually wrote 50 question-answer pairs, focusing on the types of rules which are regularly tested in law school courses on civil procedure, evidence, and intellectual property. The questions ask the LLM to either (1) restate a rule, (2) identify where a rule is codified, or (3) list the factors employed in a particular rule. Several questions explicitly narrow their scope to a jurisdiction (e.g., California state evidence law), in order to avoid bias towards merely federal law.

Question	Answer	Area of law
What are the four categories of patentable subject matter?	"process, machine, manufacture, or composition of matter."	IP
What are the requirements for diversity jurisdiction?	Diversity jurisdiction exists when the amount in controversy exceeds \$75,000 and the plaintiffs and defendants are completely diverse (i.e. no plaintiff shares a state of citizenship with any defendant)	Civil Procedure
Under which statute are patentable subject matter requirements codified?	35 USC 101	IP
What are the factors of the Mathews balancing test?	A three-part test that determines whether an individual has received due process under the Constitution. The test balances (1) the importance of the interest at stake; (2) the risk of an erroneous deprivation of the interest because of the procedures used, and the probable value of additional procedural safeguards; and (3) the government's interest.	Constitutional law

Table 49: Examples for Rule QA

Significance and value The Rule QA task is an initial effort to evaluate the propensity for legal hallucination in LLMs. The questions asked are exceedingly basic, and law students taking the relevant course would be expected to answer them nearly perfectly.

F.26 SARA Tasks

In LEGALBENCH, the SARA tasks are denoted as `sara_*`.

Background An important skill for lawyers is to determine, given the facts of a case, whether a given law applies and what it prescribes. For example, *does the payment received by the defendant on August 21st , 2017 qualify as wages under §3306(b) of the US Tax Code?* This task has been introduced by [78] as statutory reasoning. [65] further introduce the Statutory Reasoning Assessment dataset (SARA). SARA contains (1) a set of 9 sections, taken from US federal tax law statutes, pruned and simplified; and (2) hand-crafted cases that test the understanding of those 9 sections. In this context, a case is a paragraph of text stating facts in plain language. Each case comes either with an entailment prompt — a statement about the statutes and the case that may be true or false — or a question — asking how much tax one of the case's protagonists owes. The SARA dataset is a simplified version of real-world cases, that retains many of the features of statutory reasoning for tax law. It poses, however, a significant challenge to NLP models [12].

Tasks There are two SARA tasks. The first, `sara_entailment`, corresponds to the entailment cases. The entailment cases state that a given law applies to a given case, and require the LLM to produce a binary answer — akin to Recognising Textual Entailment [40]. This is an approximation of real-world statutory reasoning, where the answer is usually not strictly binary.

The second task, `sara_numeric`, consists of the numeric cases. Here, the goal is to compute the amount of tax owed. We frame this as a floating point number. To measure numerical accuracy, we use the metric introduced by [65], which includes a tolerance for inaccurate predictions.

Construction process We framed the SARA dataset for the paradigm of language modeling. Due to dependencies across sections in the statutes, the entirety of the statutes are generally relevant to determine the answer to any of the cases. However, all 9 sections do not fit into the LLM's context window, and must be pruned. In entailment cases, the entailment prompt specifies which law from the statutes to apply. We automatically extract the text of that law, and use it as the language model prompt. For numerical cases, the entirety of the statutes are relevant, and we use that as the language model prompt. Pruning is left to the LLM's pre-processing.

Significance and value Statutory reasoning is an important skill for lawyers, that is used within many other legal tasks. It is a fundamental task for legal AI, probing whether a computational model can understand and reason with legal rules expressed in natural language. The types of reasoning involved in SARA are diverse — defeasible, temporal, numerical reasoning, inter alia — and relevant beyond the legal domain. Statutory reasoning combines natural language understanding and logical reasoning, a major goal for AI.

If statutory reasoning were solved, it could serve as a basis for more complex legal tasks. For example, it could be used to automate the computation of taxes and benefits, without the need for coding the expert systems in use in many parts of the world [101]. A system that can do statutory reasoning would also be a step towards machine reading models that can analyze legislation and anticipate its effects, coming up with possible application scenarios [11]. As a final example, a statutory reasoning agent could be used for basic legal advice, increasing access to justice.

Task name: `sara_entailment`
Statute: (2) an individual legally separated from his spouse under a decree of divorce or of separate maintenance shall not be considered as married.
Description: Alice and Bob got married on April 5th, 2012. Alice and Bob were legally separated under a decree of divorce on September 16th, 2017.
Statement: Section 7703(a)(2) applies to Alice for the year 2018.
Answer: Entailment

Task name: `sara_numeric`
Statute: §3301. Rate of tax

 There is hereby imposed on every employer (as defined in section 3306(a)) for each calendar year an excise tax, with respect to having individuals....**[Ommitted from clarity]**...This section shall not apply to any taxable year beginning after December 31, 2017, and before January 1, 2026.

Description: Bob is Charlie and Dorothy's son, born on April 15th, 2015. Alice married Charlie on August 8th, 2018. Alice's and Charlie's gross incomes in 2018 were $324311 and $414231 respectively. Alice, Bob and Charlie have the same principal place of abode in 2018. Alice and Charlie file jointly in 2018, and take the standard deduction.
Question: How much tax does Alice have to pay in 2018?
Answer: $259487

Table 50: Example of each SARA task. For `sara_numeric`, we ommit part of the statute for brevity.

F.27 SCALR

In LEGALBENCH, the SCALR task is denoted as `scalr`.

Background Each case decided by the Supreme Court addresses a specific *question presented for review*. Both the questions and the Court's opinions are published on the Supreme Court's website.

Many of the Court's opinions are briefly described by other judges who recount the holdings of the Court in their own writing. For example, consider the following passage from State of South Carolina v. Key, 27971 (S.C. 2020; emphasis added):

> The United States Supreme Court has addressed the constitutionality of warrantless blood draws in several DUI cases. See Schmerber, 384 U.S. at 770-71 (**holding** the warrantless blood draw of a DUI suspect was valid because the law enforcement officer, dealing with a car accident, could "reasonably have believed that he was confronted with an emergency, in which the delay necessary to obtain a warrant, under the circumstances, threatened 'the destruction of evidence'")...

We refer to these brief descriptions as 'holding statements' or 'holding parentheticals,' since they are often enclosed by parentheses. Identifying the holding parenthetical that corresponds to a question presented for review requires a notion of 'responsiveness' or relevance between questions and answers as well as an understanding of the kinds of legal issues that could be implicated by a specific question presented for review.

Task The SCALR benchmark is a collection of 571 multiple choice questions designed to assess the legal reasoning and reading comprehension ability of large language models. Each multiple-choice task gives the question presented for review in a particular Supreme Court case. The solver must determine which holding parenthetical describes the Court's ruling in response to the question presented. Here is an example from *AT&T Mobility LLC v. Concepcion*, 563 U.S. 333 (2011) with the correct response emphasized:

> Question: Whether the Federal Arbitration Act preempts States from conditioning the enforcement of an arbitration agreement on the availability of particular procedures–here, class -wide arbitration–when those procedures are not necessary to ensure that the parties to the arbitration agreement are able to vindicate their claims.
>
> A: holding that when the parties in court proceedings include claims that are subject to an arbitration agreement, the FAA requires that agreement to be enforced even if a state statute or common-law rule would otherwise exclude that claim from arbitration
> B: holding that the Arbitration Act "leaves no place for the exercise of discretion by a district court, but instead mandates that district courts shall direct the parties to proceed to arbitration on issues as to which an arbitration agreement has been signed"
> C: holding that class arbitration "changes the nature of arbitration to such a degree that it cannot be presumed the parties consented to it by simply agreeing to submit their disputes to an arbitrator"
> **D: holding that a California law requiring classwide arbitration was preempted by the FAA because it "stands as an obstacle to the accomplishment and execution of the full purposes and objectives of Congress," which is to promote arbitration and enforce arbitration agreements according to their terms**
> E: holding that under the Federal Arbitration Act, a challenge to an arbitration provision is for the courts to decide, while a challenge to an entire contract which includes an arbitration provision is an issue for the arbitrator

Construction The data used to create this task comes from two sources:

1. Questions presented were gathered from the Supreme Court of the United States' website, which hosts PDFs of questions granted for review in each case dating back to the 2001 Term.
2. Holding statements that comprise the "choices" for each question were compiled from (a) CourtListener's collection of parenthetical descriptions and (b) extraction of parenthetical descriptions from Courtlistener's and the Caselaw Access Project's collections of court decisions using Eyecite.

Because questions presented for review in Supreme Court cases are not easily available prior to 2001, the benchmark is limited to questions from cases decided in the 2001 Term and later. To ensure that "holding" statements would address the particular question presented, we limited the set of cases to those in which exactly one question was granted for review. We also perform some manual curation to exclude questions which are not answerable without specific knowledge of a case. For example, we eliminated a case that presented this question: "Whether this Court's decision in Harris v. United States, 536 U.S. 545 (2002), should be overruled."

To create choices for each question presented, we first filter our set of parenthetical descriptions as follows:

1. We limited our parenthetical descriptions to only those that begin with "holding that...", as these are most likely to describe the core holding of the case, rather than some peripheral issue.

2. We use only parentheticals that describe Supreme Court cases. This avoids the creation of impossible questions that ask the solver to distinguish between “holding" statements dealing with exactly the same issue at different stages of appellate review.
3. We then select for each case the longest parenthetical meeting the above criteria. We use the longest parenthetical because it is most likely to be descriptive enough to make the question answerable.

We then create a task for each case which has both a question presented and a “holding" statement meeting the above requirements. (While question-correct holding pairs are only for cases decided after 2001, we allow the use of parentheticals describing any Supreme Court case as alternative answer choices.) We then need to select the four alternative answer choices for each question in a manner that makes the task challenging. To select choices that are at least facially plausible, we find the four “holding" statements from the remaining pool that are most TF-IDF similar to the question presented. The inclusion of difficult alternative choices requires the solver to draw nuanced distinctions between legal issues that share overlap in terminology.

Significance and value This task is significant because it tracks the useful and challenging skill of identifying a passage as relevant or responsive to a given query. LLMs that are able to perform well at this task have the potential to be more useful for complex legal question-answering and retrieval. The poor performance of simpler models on this task demonstrates that it is a challenging one that requires a level of understanding beyond the word/synonym level.

F.28 Securities Complaint Extraction

In LEGALBENCH, the Securities Complaint Extraction tasks are denoted as `ssla_*`.

Background Securities Class Actions (SCAs) are lawsuits filed by, and on behalf of, investors alleging economic injury as a result of material misstatements or omissions in public disclosures made by corporate directors and officers. These actions allege violations of the Securities Act of 1933 and Securities Exchange Act of 1934 and are predominately filed in federal court, though in 2018 the United States Supreme Court determined actions brought under the '33 Act were permitted in state court.

"Plaintiff(s)" is the legal term to describe the individual, company, or organization bringing forth a lawsuit. Under the class action system, one or more plaintiffs are appointed "Lead Plaintiff" by the court to represent the interests of a larger group of "similarly situated" parties. In securities class actions, investors that suffered the greatest financial loss, often public pensions or unions, are appointed lead plaintiff.

"Defendant(s)" is the legal term to describe the individual, company, or organization that must defend themself against the alleged violations or misconduct outlined in the lawsuit. There is always at least one defendant. The majority of securities class actions name the company, its CEO and its CFO. Many name additional C-suite level officers, members of the Board of Directors and additional third-parties such as the company's independent auditor and the underwriters of public offerings.

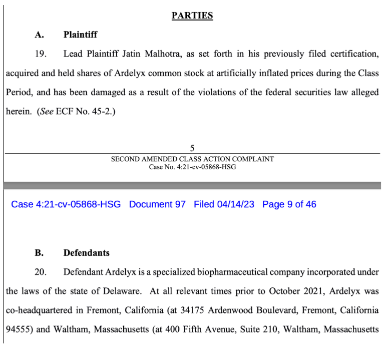
Each designated lead plaintiff, and all named defendants, are explicitly identified under the "Parties" section of the class action complaint.

Tasks There are three extraction tasks.

- The plaintiff task requires an LLM to extract the named plaintiffs within a text.
- The individual defendants tasks require an LLM to extract named defendants who are individuals from within a text.
- The company defendants tasks require an LLM to extract named defendants who are corporations/companies from within a text.

For certain samples, the complaint excerpt may not exactly contain the answer. For example, the correct answer may be "Strongbridge Biopharma PLC," while the complaint only mentions "Strongbridge." In order to maintain fidelity to the workflow used by SSLA, we evaluate an LLM's ability to generate the official name of the entity, as represented in the answer. This requires the LLM to possess some background knowledge regarding official corporation names. We find that larger models are generally able to account for this.

Sometimes, the provided text will not explicitly name the plaintiff, an individual defendant, or a corporate defendant. In these cases, the LLM is expected to return "Not named".



PARTIES

A. Plaintiff

19. Lead Plaintiff Jatin Malhotra, as set forth in his previously filed certification, acquired and held shares of Ardelyx common stock at artificially inflated prices during the Class Period, and has been damaged as a result of the violations of the federal securities law alleged herein. (*See* ECF No. 45-2.)

5

SECOND AMENDED CLASS ACTION COMPLAINT
Case No. 4:21-cv-05868-HSG

Case 4:21-cv-05868-HSG Document 97 Filed 04/14/23 Page 9 of 46

B. Defendants

20. Defendant Ardelyx is a specialized biopharmaceutical company incorporated under the laws of the state of Delaware. At all relevant times prior to October 2021, Ardelyx was co-headquartered in Fremont, California (at 34175 Ardenwood Boulevard, Fremont, California 94555) and Waltham, Massachusetts (at 400 Fifth Avenue, Suite 210, Waltham, Massachusetts

Figure 5: Example of the typical SCA structure (Case 4:21-cv-05868-HSG)

Construction process Stanford Securities Litigation Analytics (SSLA) identifies, tracks, and aggregates data on the several hundred private shareholder lawsuits and public SEC/DOJ enforcements filed each year. SSLA fellows manually extract and analyze information including plaintiffs, defendants, judges, mediators, plaintiff and defense firms, key litigation events, real-time case statuses, settlement timing, settlement dollar amounts, attorneys' fees and expenses, and imposed SEC / DOJ penalties. There is no ambiguity regarding the answers for this task given its nature.

This dataset is an extract from the corpus of texts of securities class action complaints in the SSLA database. Given the typical structure and headings for these types of cases, this dataset represents text extracted from the complaint, between the sections titled "Parties" and "Substantive Allegations". For cases where the second heading was not found, texts fragments were limited to 2,000 characters. Cases with both headings were then filtered to include only those with texts up to 2,000 characters, which excluded cases with longer "Parties" sections. Thus, all observations in this dataset are 2,000 characters or less.

Text was scraped from complaints using python's PyPDF2 library and left unformatted and uncleaned. This training set includes several observations where the text does not include all or some of the named entities due to the method of text collection and variation in case structure. In several of these cases, plaintiff names are not present in the selected text because the plaintiff had been named earlier in the complaint.

Table 51: Examples from Securities Complaint Tasks

Task
Task name: `ssla_company_defendants` **Excerpt**: 6. Plaintiff Don L. Gross, as set forth in the accompanying certification and incorporated by reference herein, purchased the common stock of KCS during the Cla ss Period and has been damaged thereby. 7. Defendant KCS, headquartered in Kansas C ity, Missouri, operates railroads in the Midwest and Mexico that run north to south, unlike most other U.S. ra ilroads that run east to west. The Company's stock traded on the NYSE, an ef ficient market, during th e Class Period under the ticker symbol "KSU." As of October 11, 2013, there were more than 110 million shares issued and outstanding. 8. Defendant David L. Starling ("Starling"), at all relevant times, served as KCS's President and Chief Executive Officer ("CEO"). Case 4:14-cv-00345-BCW Document 1 Filed 04/15/14 Page 2 of 293 9. Defendant David R. Ebbrecht ("Ebbrecht"), at all relevant times, served as KCS's Executive Vice President and Chief Operating Officer ("COO"). 10. Defendant Patrick J. Ottensmeyer ("Ottensmeye r"), at all relevant times, served as KCS's Executive Vice President Sales & Marketing. 11. Defendant Michael W. Upchurch ("Upchurch "), at all relevant times, served as KCS's Executive Vice President and Chief Financial Officer ("CFO"). 12. Defendants Starling, Ebbrecht, Ottensmeyer an d Upchurch are collec tively referred to herein as the "Individual Defendants." 13. During the Class Period, the Individual Defe ndants, as senior executive officers and/or directors of KCS, were privy to confidential and proprie tary information concerning KCS, its operations, finances, financial cond ition and present and future business prospects. The Individual Defendants also had access to material adverse non-public information concerning KCS, as discussed in detail below. Because of their positions with KCS, the Individual Defendants had access to non-public information about its busine ss, finances, products, markets and present and future business prospects via interna **Answer**: Kansas City Southern
Task name: `ssla_individual_defendants` **Excerpt**: 6. Plaintiff Don L. Gross, as set forth in the accompanying certification and incorporated by reference herein, purchased the common stock of KCS during the Cla ss Period and has been damaged thereby. 7. Defendant KCS, headquartered in Kansas C ity, Missouri, operates railroads in the Midwest and Mexico that run north to south, unlike most other U.S. ra ilroads that run east to west. The Company's stock traded on the NYSE, an ef ficient market, during th e Class Period under the ticker symbol "KSU." As of October 11, 2013, there were more than 110 million shares issued and outstanding. 8. Defendant David L. Starling ("Starling"), at all relevant times, served as KCS's President and Chief Executive Officer ("CEO"). Case 4:14-cv-00345-BCW Document 1 Filed 04/15/14 Page 2 of 293 9. Defendant David R. Ebbrecht ("Ebbrecht"), at all relevant times, served as KCS's Executive Vice President and Chief Operating Officer ("COO"). 10. Defendant Patrick J. Ottensmeyer ("Ottensmeye r"), at all relevant times, served as KCS's Executive Vice President Sales & Marketing. 11. Defendant Michael W. Upchurch ("Upchurch "), at all relevant times, served as KCS's Executive Vice President and Chief Financial Officer ("CFO"). 12. Defendants Starling, Ebbrecht, Ottensmeyer an d Upchurch are collec tively referred to herein as the "Individual Defendants." 13. During the Class Period, the Individual Defe ndants, as senior executive officers and/or directors of KCS, were privy to confidential and proprie tary information concerning KCS, its operations, finances, financial cond ition and present and future business prospects. The Individual Defendants also had access to material adverse non-public information concerning KCS, as discussed in detail below. Because of their positions with KCS, the Individual Defendants had access to non-public information about its busine ss, finances, products, markets and present and future business prospects via interna **Answer**: David L Starling, David R Ebbrecht, Patrick J. Ottensmeyer, Michael W. Upchurch
Task name: ssla_plaintiff **Excerpt**: 11. Plaintiff, as set forth in th e attached certification, purchas ed Catalyst securities at artificially inflated prices dur ing the Class Period and has b een damaged upon the revelation of the alleged corrective disclosures. 12. Defendant Catalyst is a Coral Gates, Florida headquartered company located at 355 Alhambra Circle Suite 1500 Coral Gates, FL 33134. The common stock is traded on the NASDAQ Stock Market ("NASDAQ") unde r the ticker symbol "CPRX." 13. Defendant Patrick J. McEnany ("McEna ny") is the Company's co-founder, CEO and President. 14. Defendant Dr. Hubert E. Huckel M.D. ("Huckel") is the Company's co-founder and one of its directors. 15. Defendant Steven R. Miller Ph. D. ("M iller") is the company's COO and CSO. 16. The defendants referenced above in ¶¶ 13- 15 are sometimes referred to herein as the "Individual Defendants." DEFENDANTS' WRONGDOING Background Case 1:13-cv-23878-UU Document 1 Entered on FLSD Docket 10/25/2013 Page 4 of 20 5 17. Catalyst is a specialty pharmaceutical company which develops and commercializes drugs treating orphan (rare) neuromuscular an d neurological diseases. 18. Lambert-Eaton Myasthenic Syndrome ("LEM S") is an extremely serious disase which is also extremely rare, afflicting about 3.4 persons per million, and about one to two thousand patients in the United States. 19. FDA rules permit so-called "compassionate use" – use of a drug that has not been approved by the FDA outside of clinical trials. A patient may be given drugs under a compassionate use program if the patient may benefit from the treatment, the therapy can be given safely outside the clinical trial setting, no other alternative therapy is available, and the drug developer agrees to provide access to the drug. 20. Jacobus is a tiny privat e pharmaceutical company in New Jersey, with only dozens of employees, and only 35 as of 2009. Jacobus has b een manufacturing 3,4 DAP and providing it to patients through a **Answer**: Not named

Significance and value Extracting data from legal documents is an extraordinarily resource- and time-intensive effort prone to human error. As a result, there are no known databases of non-securities class action litigation, despite the obvious public policy implications of the class action system. Automation of identification tasks coupled with human approval would improve efficiency and reduce collection costs and data errors. This task may be useful to other legal researchers and industry practitioners extracting structured data from complex texts. Identification is a very simple task that can be done by those with an understanding and familiarity with the underlying legal documents and legal system, but an LLM's ability to accurately and precisely identify entities is a useful metric to assess.

F.29 Successor Liability

In LEGALBENCH, the Successor Liability task is denoted as `successor_liability`.

Background When one company sells its assets to another company, the purchaser is generally not liable for the seller's debts and liabilities. Successor liability is a common law exception to this general rule. In order to spot a successor liability issue, lawyers must understand how courts apply the doctrine.

The doctrine holds purchasers of all, or substantially all, of a seller's assets liable for the debts and liabilities of the seller if:

1. the purchaser expressly agrees to be held liable;
2. the assets are fraudulently conveyed to the purchaser in order to avoid liability;
3. there is a de facto merger between the purchaser and seller; or
4. the purchaser is a mere continuation of the seller.

Express agreement is governed by standard contract law rules. In practice, if a purchase agreement contains a provision to assume liabilities, litigation will rarely arise. Courts, however, sometimes interpret an implied agreement in the absence of a written provision.

Assets are fraudulently conveyed when the seller intends to escape liability through a sale or knows that liability will be avoided through a sale.

De facto merger is a multifactor test that consists of (1) continuity of ownership; (2) cessation of ordinary business and dissolution of the acquired corporation as soon as possible; (3) assumption by the purchaser of the liabilities ordinarily necessary for the uninterrupted continuation of the business of the acquired corporation; and (4) continuity of management, personnel, physical location, assets, and general business operation. Some jurisdictions require a showing of all four elements. Others do not, and simply emphasize that the substance of the asset sale is one of a merger, regardless of its form.

Mere continuation typically requires a showing that after the asset sale, only one corporation remains and there is an overlap of stock, stockholders, and directors between the two corporations. There are two variations of the mere continuation exception. The first variation is the "continuity of enterprise" exception. In order to find continuity of enterprise, and thus liability for the purchaser of assets, courts engage in a multifactor analysis. Factors include: (1) retention of the same employees; (2) retention of the same supervisory personnel; (3) retention of the same production facilities in the same physical location; (4) production of the same product; (5) retention of the same name; (6) continuity of assets; (7) continuity of general business operations; and (8) whether the successor holds itself out as the continuation of the previous enterprise. The second variation is the product line exception. This exception imposes liability on asset purchasers who continue manufacturing products of a seller's product line. This exception generally requires that defendants show that the purchaser of assets is able to assume the risk spreading role of the original manufacturer, and that imposing liability is fair because the purchaser enjoys the continued goodwill of the original manufacturer.

Scholars have noted that fraud, de facto merger, and mere continuation (and its variants) overlap. They share the common thread of inadequate consideration, that is, the consideration given in exchange for the assets is unable to fund the liabilities that underwrite those assets. Because of the overlap, different courts might apply different doctrines to identical sets of facts, but arrive at the same policy [50].

Successor liability doctrine is commonly taught in a course on corporate law or business associations in law school. Sometimes it is reserved for upper level courses in corporate finance or mergers and acquisitions. Students are expected to spot successor liability issues and understand how to determine if a successor will be held liable.

Task The Successor Liability task requires an LLM to spot a successor liability issue and identify its relevant exception to no liability. If more than one exception is relevant, the LLM is required to state the additional exception(s). The task does not include identification of the two variations to the mere continuation exception (continuity of enterprise and product line).

Facts	Issue	Relevant exception
Large Incarceration Services purchased a substantial amount of Small Prison's assets last year. The asset purchase agreement expressly disclaimed Small Prison's potential liability for an employment discrimination claim arising out of its prison service activities. Small conveyed its assets to Large because Small's owners were concerned that the liability arising from the lawsuit would lead to bankruptcy. Several months following the asset purchase, Small Prison lost the discrimination lawsuit. The plaintiffs now seek relief from Large Incarceration Services.	successor liability	fraudulent conveyance, mere continuation
Large Incarceration Services purchased a substantial amount of Small Prison's assets last year. The asset purchase agreement expressly assumed any liability arising out of its prison service activities. Several months following the asset purchase, Small Prison lost a number of discrimination lawsuits. The plaintiffs now seek relief from Large Incarceration Services.	successor liability	express agreement
Big Pharma purchases substantially all of DW I's assets. The purchase agreement expressly provides for assumption of only those liabilities necessary for continuing operations of DW I. DW I had developed a successful drug that regulated oxygen levels in the blood. After the purchase of DW I's assets, DW I dissolves. DW I's shareholders maintain ownership in Big Pharma equivalent to their ownership in DW I. In addition, there is some overlap between the two companies' management teams. Moreover, Big Pharma continues to employ seventy percent of DW I's workforce. Past users of DW I's drug bring a mass tort claim against Big Pharma alleging that DW I's drug incorrectly measured oxygen levels in the blood leading to harm.	successor liability	de facto merger, mere continuation
Big Pharma purchases substantially all of DW I's assets. The purchase agreement expressly provides for assumption of only those liabilities necessary for continuing operations of DW I. DW I had developed a successful drug that regulated oxygen levels in the blood. After the purchase of DW I's assets, DW I dissolves. DW I's shareholders do not own any stock in Big Pharma. However, there is some overlap between the two companies' management teams. In addition, Big Pharma continues to employ seventy percent of DW I's workforce. Past users of DW I's drug bring a mass tort claim against Big Pharma alleging that DW I's drug incorrectly measured oxygen levels in the blood leading to harm.	successor liability	mere continuation

Table 52: Examples for successor liability.

F.30 Supply Chain Disclosure Tasks

In LEGALBENCH, the Supply Chain Disclosure Tasks are denoted as `supply_chain_disclosure_*`.

Background Corporations are frequently legally required to disclose information that may be relevant to investors, regulators, or members of the public. One example of this kind of disclosure requirement is laws that require corporations doing business in particular jurisdictions to provide detailed information on their supply chains, which is intended to ensure that the company's business practices are not supporting things like human trafficking or human rights violations. One example of these kind of disclosure requirements is the California Transparency in Supply Chains Act (CTSCA). The CTSCA applies to corporations that are a: "[1] retail seller and manufacturer [2] doing business in this state [of California] and [3] having annual worldwide gross receipts that exceed one hundred million dollars ($100,000,000)."[31] If a corporation meets these criteria, they are required to post information on five topics:

- **Verification**: "[A]t a minimum, disclose to what extent, if any, that the retail seller or manufacturer . . . [e]ngages in verification of product supply chains to evaluate and address risks of human trafficking and slavery. The disclosure shall specify if the verification was not conducted by a third party."[32]
- **Audits**: "[A]t a minimum, disclose to what extent, if any, that the retail seller or manufacturer . . . [c]onducts audits of suppliers to evaluate supplier compliance with company standards for trafficking and slavery in supply chains. The disclosure shall specify if the verification was not an independent, unannounced audit."[33]
- **Certification**: "[A]t a minimum, disclose to what extent, if any, that the retail seller or manufacturer . . . [r]equires direct suppliers to certify that materials incorporated into the product comply with the laws regarding slavery and human trafficking of the country or countries in which they are doing business."[34]
- **Accountability**: "[A]t a minimum, disclose to what extent, if any, that the retail seller or manufacturer . . . [m]aintains internal accountability standards and procedures for employees or contractors failing to meet company standards regarding slavery and trafficking."[35]
- **Training**: "[A]t a minimum, disclose to what extent, if any, that the retail seller or manufacturer . . . [p]rovides company employees and management, who have direct responsibility for supply chain management, training on human trafficking and slavery, particularly with respect to mitigating risks within the supply chains of products."[36]

In addition to requiring corporations that meet the specified criteria to post disclosures that provide this information, the California Attorney General's office has also posted a guide informing firms of "Best Practices" for what specific information to provide on each of these five topics.[37]

However, prior research has suggested that companies do not always post disclosures that cover each of these topics; and, even when they do, the disclosures are not always consistent with the recommended best practices [27].

Construction process We constructed this task based on an existing dataset of supply chain disclosures. In the summer of 2015, we, with the help of research assistants, we searched the websites of corporations that had previously been identified by an organization called "KnowTheChain" as being required to post supply chain disclosures to be compliant with the California Supply Chain Transparency Act. Through this process, we found disclosures for roughly 400 firms out of roughly 500 firms for which KnowTheChain suggested were required to post disclosures.

For each of these roughly 400 firms, we saved copies of their supply chain disclosures. We then had research assistants read the disclosures and code whether they included each of the five required topics for disclosure and, if so, whether the disclosures on those five topics were consistent with the best practices outlined by the California Attorney General's office.

We convert each of these 10 coded variables into a distinct binary classificationt task, producing 10 tasks. Table 53 lists each task, along with the precise question used to code the disclosure.

Significance and value Corporate disclosure requirements are a commonly used regulatory tool, but evidence suggests that firms do not always fully comply with these disclosure requirements. The Supply Chain Disclosure task evaluates whether LLMs may be able to determine whether corporations are complying with those disclosure requirements. Because these disclosures are often formatted very differently, written in complex language, and may be designed to obfuscate relevant information, this task provides a useful measure of whether LLMs can parse the content covered in legal documents.

[31] *See* California Transparency in Supply Chains Act, CAL. CIV. CODE § 1714.43(a)(1) (West 2012).
[32] CAL. CIV. CODE § 1714.43(c)(1).
[33] CAL. CIV. CODE § 1714.43(c)(2).
[34] CAL. CIV. CODE § 1714.43(c)(3).
[35] CAL. CIV. CODE § 1714.43(c)(4).
[36] CAL. CIV. CODE § 1714.43(c)(5).
[37] CAL. DEP'T OF JUSTICE, THE CALIFORNIA TRANSPARENCY IN SUPPLY CHAINS ACT: A RESOURCE GUIDE (2015), `https://oag.ca.gov/sites/all/files/agweb/pdfs/sb657/resource-guide.pdf`.

Task	**Question**
`disclosed_verification`	Does the statement disclose to what extent, if any, that the retail seller or manufacturer engages in verification of product supply chains to evaluate and address risks of human trafficking and slavery? If the company conducts verification], the disclosure shall specify if the verification was not conducted by a third party.
`disclosed_audits`	Does the statement disclose to what extent, if any, that the retail seller or manufacturer conducts audits of suppliers to evaluate supplier compliance with company standards for trafficking and slavery in supply chains? The disclosure shall specify if the verification was not an independent, unannounced audit.
`disclosed_certification`	Does the statement disclose to what extent, if any, that the retail seller or manufacturer requires direct suppliers to certify that materials incorporated into the product comply with the laws regarding slavery and human trafficking of the country or countries in which they are doing business?
`disclosed_accountability`	Does the statement disclose to what extent, if any, that the retail seller or manufacturer maintains internal accountability standards and procedures for employees or contractors failing to meet company standards regarding slavery and trafficking?
`disclosed_training`	Does the statement disclose to what extent, if any, that the retail seller or manufacturer provides company employees and management, who have direct responsibility for supply chain management, training on human trafficking and slavery, particularly with respect to mitigating risks within the supply chains of products?
`best_practice_verification`	Does the statement disclose whether the retail seller or manufacturer engages in verification and auditing as one practice, expresses that it may conduct an audit, or expressess that it is assessing supplier risks through a review of the US Dept. of Labor's List?
`best_practice_audits`	Does the statement disclose whether the retail seller or manufacturer performs any type of audit, or reserves the right to audit?
`best_practice_certification`	Does the statement disclose whether the retail seller or manufacturer requires direct suppliers to certify that they comply with labor and anti-trafficking laws?
`best_practice_accountability`	Does the statement disclose whether the retail seller or manufacturer maintains internal compliance procedures on company standards regarding human trafficking and slavery? This includes any type of internal accountability mechanism. Requiring independently of the supply to comply with laws does not qualify or asking for documentary evidence of compliance does not count either.
`best_practice_training`	Does the statement disclose whether the retail seller or manufacturer provides training to employees on human trafficking and slavery? Broad policies such as ongoing dialogue on mitigating risks of human trafficking and slavery or increasing managers and purchasers knowledge about health, safety and labor practices qualify as training. Providing training to contractors who failed to comply with human trafficking laws counts as training.

Table 53: Supply Chain Disclosure Tasks

F.31 Telemarketing Sales Rule

In LEGALBENCH, the Telemarketing Sales Rule task is denoted as `telemarketing_sales_rule`.

Background The Telemarketing Sales Rule (16 C.F.R. Part 310) is a set of regulations promulgated by the Federal Trade Commission to implement the Telemarketing and Consumer Fraud and Abuse Prevention Act. Its purpose is to protect consumers from specified deceptive and abusive telemarketing practices. This task focuses on 16 C.F.R. § 310.3(a)(1) and 16 C.F.R. § 310.3(a)(2), which outline a series of specific telemarketing practices prohibited as "deceptive." 16 C.F.R. § 310.3(a)(1) lists information that must be disclosed to a consumer before a sale is made, and 16 C.F.R. § 310.3(a)(2) lists categories of information that a telemarketer is prohibited from misrepresenting. 16 C.F.R. § 310.2 provides definitions relevant to both of these subsections.

The Telemarketing Sales Rule (TSR) is not commonly taught in law school as its own topic, but may be used as examples in courses on consumer protection law, administrative law, telecommunications law, and the like. Because of its simplicity, it has also been used in beginner-level legal research exercises tasking students with finding the TSR in the Code of Federal Regulations and applying it to a set of facts.

Applying the TSR would require an LLM to classify a set of facts as either falling within or outside of the specific prohibitions outlined in the rule. For example, the TSR requires that telemarketers disclose certain material information before a sale is made, such as the total cost of the goods or services, the quantities of goods or services being purchased, and exchange and return restrictions. It also forbids telemarketers from making material misrepresentations as to cost, quantity, quality, endorsement or sponsorship, and the like. In many real-life situations, it would be ambiguous whether certain telemarketer behavior would violate the TSR; for example, it could be contentious whether a given misrepresentation fits the definition of “material.” However, this task is limited to clear, unambiguous violations or non-violations, such as if a telemarketer told a consumer that they were selling four apples for four dollars, when in fact they were selling four apples for six dollars.

The following subsections 16 C.F.R. § 310.3(a)(1) and 16 C.F.R. § 310.3(a)(2) were ignored in the task, given their complexity or their reference to other statutes and regulations:

- 16 C.F.R. § 310.3(a)(2)(vi)
- 16 C.F.R. § 310.3(a)(2)(viii)

Task The TSR task is meant to test whether an LLM can classify simple sets of facts as describing a violation of the TSR, or not describing a violation of the TSR.

Construction process We manually created 50 samples, such that examples of at least one violation and at least one non-violation of each relevant subsection of 16 C.F.R. § 310.3(a)(1) and 16 C.F.R. § 310.3(a)(2) were present.

Significance and value Determining whether a simple and unambiguous set of facts falls within the ambit of 16 C.F.R. § 310.3(a)(1) or 16 C.F.R. § 310.3(a)(2) would be an easy task for law students and lawyers, as well as many non-lawyers. However, an LLM that was trained to recognize clear violations of consumer protection laws could help administrative agencies like the Federal Trade Commission inform normal citizens of their rights.

Input	Answer
Acme Toys is a telemarketer subject to the Telemarketing Sales Rule. Acme Toys sold a customer a frisbee. It disclosed the brand of the frisbee, but did not tell the customer the frisbee was manufactured in Portugal. Is this a violation of the Telemarketing Sales Rule?	No
Acme Toys is a telemarketer subject to the Telemarketing Sales Rule. Acme Toys told a customer that it would sell them a handful of frisbees at a very reasonable price, and that shipping would be $5. Then, the customer agreed to the sale. Is this a violation of the Telemarketing Sales Rule?	Yes

Table 54: Examples for `telemarketing_sales_rule`

F.32 Textualism Tasks

In LEGALBENCH, the Textualism tasks are denoted as `textualism_tool_*`.

Background Courts regularly interpret statutes to determine the precise meaning of words contained in the statute. For instance, suppose a statute specifies that "It shall be illegal to park a vehicle inside public parks for longer than thirty minutes." A court may be asked to determine whether the statute prohibits persons from parking bicycles inside public parks. This requires defining the term "vehicle" and determining if a bicycle is a type of vehicle.

To guide the interpretation of ambigous statutory terms, American jurisprudence has developed numerous *principles* of statutory construction or interpretation. These principles—also referred to as tools or canons—are rules which dictate how terms in statutes should be interpreted. For instance, the principle of *ejusdem generis* states that where general words or phrases follow a number of specific words or phrases, the general words are specifically construed as limited and apply only to persons or things of the same kind or class as those expressly mentioned [132].

One approach to statutory interpretation—known as *textualism*—states that only the text of the statute should be considered [133]. In contrast, other approaches to interpreting an ambigous term might call for a court to analyze the purpose of the statute, the history of the statute, or the intent of the legislature.

Task The Textualism tasks ask a LLM to determine if an excerpt of judicial text is applying a specific textual tool when performing statutory interpretation. There are two tasks: dictionaries (textualism_tool_dictionaries) and .

- The first task is plain-meaning (textualism_tool_plain), and it requires an LLM to determine if a court is applying the "plain meaning" rule. The plain meaning rule says that statutory text should be interpreted according to its plain or ordinary meaning.
- The second task is dictionaries (textualism_tool_dictionaries), and it requires an LLM to determine if a court is using dictionaries to define the statutory text.

Construction process For each task we extracted paragraphs from Court of Appeals opinions and manually annotated whether the paragraphs showed the court as "using" the respective tool.

- In order to count as using plain meaning, the paragraph must reference the plain or ordinary meaning of the text. This includes directly saying "plain meaning" or referencing the general logic of the plain meaning rule. There must also be evidence that the court used the tool in its decision. This latter condition is notable because legal scholars often care about whether the court actually used the tool when defending its decision. Common examples of using include stating it as a general rule of decision ("[0]ur obligation is to look to the plain language of the statute to effectuate the intent of congress") or applying it to the facts ("The statute's plain language indicates the 150% fee cap applies if (1) the plaintiff was "a prisoner" at the time he brought the action and (2) he was awarded attorney's fees pursuant to § 1988."). "Using" does not, for example, include paragraphs that criticize the use of the plain meaning rule.
- In order to count as using dictionaries, the paragraph must reference a dictionary. There must also be evidence that the court used a dictionary as part of its rationale. This latter condition is notable because legal scholars often care about whether the court actually used the tool when defending its decision. Common examples of using include stating it as a general rule of decision ("[We use a dictionary to help determine the plain meaning of the statutory text") or applying it to the facts ("According to the Websters dictionary, a vehicle is any *means in or by which someone travels, or something is carried or conveyed*"). "Using" does not, for example, include paragraphs that criticize the use of dictionaries.

Significance and value Recognizing when a court is applying a particular canon of interpretation is a classical skill law students are expected to learn. LLM performance on this task thus offers a heuristic for comparing LLM comprehension of judicial text to that of a law student's. More practically, the capacity for LLMs to detect when certain canons are being applied could make them a valuable tool for empirical legal scholars.

Input	Answer
overcome our prior interpretation of a statute depends, in turn, on whether we regarded the statute as unambiguously compelling our interpretation.	No
the statutory waiver is express, and its range is defined in unmistakable language. to say that a private person, but not the united states, is liable under title vii for interest as an element of an attorneys fee would rob the unambiguous statutory language of its plain meaning. it would defeat the statutory imposition upon the united states of a liability for costs, and the statutory inclusion of a reasonable attorneys fee as part of the costs, identical to that of a private party in similar circumstances. the scope-setting statutory words the same as a private person mark out the united states liability for attorneys fees as well as costs in the traditional sense. our responsibility as judges is to enforce this provision according to its terms.	Yes

Table 55: Examples for `textualism_tool_plain`.

Input	Answer
we pause to note that even if congress sought, through the csra, to regulate the nonuse of interstate channels, it would still be within its constitutional command to do so. the supreme court has often held, in several contexts, that the defendants nonuse of interstate channels alone does not shield him from federal purview under the commerce clause. in heart of atlanta motel, inc. v. united states, 379 u.s. 241, 250, 85 s.ct. 348, 353, 13 l.ed.2d 258 (1964), the court upheld commerce clause jurisdiction over a local motel that failed to engage in interstate commerce when it refused to rent rooms to black guests. the court held that by failing to rent the rooms, the hotel inhibited black travelers from crossing state lines and thus obstructed interstate commerce that otherwise would have occurred. id. at 253, 85 s.ct. at 356. in standard oil co. v. united states, 221 u.s. 1, 68, 31 s.ct. 502, 518, 55 l.ed. 619 (1911), the court upheld the sherman act, 15 u.s.c. 1, 2, as permissible congressional action under the commerce clause. the sherman act prohibits restraints of trade and obstructions of interstate commerce in order to facilitate commerce that otherwise would occur absent the defendants monopolistic behavior. finally, in united states v. green, 350 u.s. 415, 420, 76 s.ct. 522, 525, 100 l.ed. 494 (1956), the court found constitutional the hobbs act, 18 u.s.c. 1951, which punishes interference with interstate commerce by extortion, robbery or physical violence [by] ... outlaw[ing] such interference in any way or degree. to accept baileys nonuse argument would mean, as emphasized by the second circuit, that congress would have no power to prohibit a monopoly so complete as to thwart all other interstate commerce in a line of trade[;] or to punish under the hobbs act someone who successfully prevented interstate trade by extortion and murder. sage, 92 f.3d at 105.	No
our primary area of concern with the district courts determination is its confident assertion that the language of 326(a) is unambiguous. see lan assocs., 237 b.r. at 56-57. in this day and age when we exchange by a keystroke or series of keystrokes what we used to handle only in cash, we do not think that the term moneys is so clear as the district court indicated. in fact, one of the definitions cited by the district court refers to money as a measure of value, see id. at 55-56 (citing websters third new intl dictionary 1458 (1986)), which surely is a concept that evolves along with and is dependent upon changing cultural, social, and economic practices and institutions. for example, in todays society the term money could easily encompass the concept of credit, which increasing numbers of people use as a method of payment. the term money might also encompass property, especially when property is used as a method of payment or a measure of wealth. see websters ii new college dictionary 707 (defining money as [a] medium that can be exchanged for goods and services and is used as a measure of their values on the market and as [p]roperty and assets considered in terms of monetary value); supra note 5 (describing the nabts argument that an exchange of property involves an exchange of value). but see in re brigantine beach hotel corp., 197 f.2d 296, 299 (3d cir.1952) (referring to precode statute governing receiver compensation and stating that [i]t is clear that the word moneys in the clause ... upon all moneys disbursed or turned over ... is not the equivalent of property.). these reasonable interpretations of the term moneys render it ambiguous for purposes of our interpretation of 326(a). see taylor v. continental group change in control severance pay plan, 933 f.2d 1227, 1232 (3d cir.1991) (a term is ambiguous if it is subject to reasonable alternative interpretations.); accord united states v. gibbens, 25 f.3d 28, 34 (1st cir.1994) (a statute is ambiguous if it reasonably can be read in more than one way.).	Yes

Table 56: Examples for `textualism_tool_dictionaries`

F.33 UCC vs Common Law

In LEGALBENCH, the UCC vs Common Law task is denoted as `ucc_v_common_law`.

Background In the United States, contracts are typically governed by one of two different bodies of law depending on the subject matter of the contract. Contracts for the sale of goods (physical, moveable things) are governed by the Uniform Commercial Code (UCC), a uniform set of laws created by the Uniform Law Commission and adopted in all US jurisdictions. Contracts for services and real estate, on the other hand, are governed by state common law. For example, a contract for Alice to sell Bob her bike would be governed by the UCC (sale of a good), but a contract for Bob to repair Alice's bike would be governed by the common law (service).

This distinction is significant because the UCC and the common law diverge on numerous important legal issues such as:

- Offer and acceptance: The common law requires an offeree's acceptance to exactly match the terms of the offeror's offer in order for a contract to be formed (the "mirror image" rule). The UCC, on the other hand, allows for some variation in the terms under UCC Section 2-207.
- Definiteness: For a common law contract to be enforceable, it must be reasonably definite with respect to all material terms. For example, a service contract would not be enforceable without a price term or an adequate description of the service to be provided. The UCC only requires that a goods contract include the good being sold and the quantity. If any other term is missing from the contract (such as price or delivery), it will be filled in by UCC default rules.
- Options: To create an option contract (by which the offeror provides the offeree with a defined period of irrevocability), the common law requires that the offeree give the offeror separate consideration for the option. The UCC allows merchants to make "firm offers" (effectively option contracts) without the offeree providing separate consideration.
- Modification: To modify an existing contract, the common law requires both parties to provide new consideration (the "preexisting duty rule") whereas the UCC only requires that modifications be made in good faith.

Task The UCC vs. Common Law task requires an LLM to determine whether a contract is governed by the UCC or by the common law.

Construction process The dataset was manually created to test an LLM's ability to determine whether a contract is governed by the UCC or by the common law. The dataset is composed of 100 descriptions of simple contracts such as "Alice and Bob enter into a contract for Alice to sell her bike to Bob for $50" (UCC) and "Aria pays Owen $100 to mount a television on the wall of her living room" (common law). Each description is followed the question, "Is this contract governed by the UCC or the common law?"

The dataset does not include "mixed purpose" contracts which incorporate both the sale of a good and a service. For example, a contract in which Alice sells Bob her bike for $100 and agrees to inflate the tires each week for the first month would be a mixed purpose contract. To determine whether a mixed purpose contract is governed by the UCC or the common law, most jurisdictions apply the "predominant purpose" test under which the predominant purpose of the contract (good or service) determines which body of law applies.

Significance and value The UCC vs. Common Law task is significant for a number of reasons. First, it provides a measure of an LLM's legal reasoning ability relative to a human lawyer (who would almost certainly score a 100% on the task). Second, it demonstrates an LLM's ability to determine the subject matter of a legal text, which has implications for the use of LLMs for legal tasks far beyond contract classification. Third, this task could prove useful in the context of contract lifecycle management (CLM) in which a CLM software product could automatically sort contracts by subject matter for review purposes. Fourth, while the sample contracts in the dataset were simple and easily identifiable as either UCC or common law contracts, real-world mixed purpose contracts can be difficult to classify and sometimes generate costly litigation. This task could be used as a starting point for developing a more fine-tuned task that can classify mixed purpose contracts.

G Full results

G.1 Models

HuggingFace links for the studied open-source models in Section 5.2 can be found below.

LLM	HuggingFace URL
Incite-Instruct-3B	togethercomputer/RedPajama-INCITE-Instruct-3B-v1
Incite-Base-7B	togethercomputer/RedPajama-INCITE-Base-7B-v0.1
Incite-Instruct-7B	togethercomputer/RedPajama-INCITE-Instruct-7B-v0.1
BLOOM-3B	bigscience/bloom-3b
BLOOM-7B	bigscience/bloom-7b1
OPT-2.7B	facebook/opt-2.7b
OPT-6.7B	facebook/opt-6.7b
OPT-13B	facebook/opt-13b
Falcon-7B-Instruct	tiiuae/falcon-7b-instruct
MPT-7B-8k-Instruct	mosaicml/mpt-7b-instruct
Vicuna-7B-16k	lmsys/vicuna-7b-v1.5-16k
Vicuna-13B-16k	lmsys/vicuna-13b-v1.5-16k
Flan-T5-XL	google/flan-t5-xl
Flan-T5-XXL	google/flan-t5-xxl
LLaMA-2-7B	meta-llama/LLaMA-2-7b-hf
LLaMA-2-13B	meta-llama/LLaMA-2-13b-hf
WizardLM-13B	WizardLM/WizardLM-13B-V1.2

Table 57: HuggingFace links for open-source models.

G.2 Prompts

Prompts for all LEGALBENCH experiments are available on the Github repository. For experiments reported in Section 5.2, Table 58 provides the number of in-context demonstrations used.

G.3 Results

We provide results for each LLM on each of the tasks. Models are divided into four groups based on type: commercial models, 13B models, 7B models, and 3B models. Model names are abbreviated to the family name to ensure well-formed tables.

Number of in-context demonstrations	Tasks
0	Canada Tax Court Outcomes, Consumer Contracts QA, Corporate Lobbying, International Citizenship Questions, Rule QA, Supply Chain Disclosure Tasks, SARA (Numeric)
1	MAUD Tasks, SCALR
2	Citation Prediction Tasks
3	Securities Complaint Extraction Tasks
4	Legal Reasoning Causality, Personal Jurisdiction, Successor Liability, SARA (Entailment) Telemarketing Sales Rule, Textualism Tools
5	Abercrombie, Hearsay, Insurance Policy Interpretation, Private Right of Action
6	CUAD Tasks, Diversity Tasks, J.Crew Blocker, Learned Hands Tasks, Overruling, UCC v. Common Law
7	Function of Decision Section, Oral Argument Question Purpose
8	Contract NLI Tasks, Contract QA, Definition Tasks, New York State Judicial Ethics, OPP-115 Tasks, Privacy Policy Entailment, Privacy Policy QA
9	Unfair Terms of Service

Table 58: Number of in-context demonstrations used for each type.

Task	GPT-4	GPT-3.5	Claude-1
abercrombie	84.2 / 84.2	48.4 / 48.4	65.2 / 63.1
abercrombie	84.2 / 84.2	48.4 / 48.4	65.2 / 63.1
diversity_1	96.7 / 96.7	86.7 / 86.7	86.7 / 86.7
diversity_2	100.0 / 100.0	53.3 / 50.0	73.3 / 73.3
diversity_3	96.7 / 96.7	83.3 / 66.7	53.3 / 53.3
diversity_4	93.3 / 93.3	70.0 / 66.7	43.3 / 43.3
diversity_5	76.6 / 76.6	66.7 / 66.7	36.7 / 36.7
diversity_6	80.0 / 80.0	6.7 / 6.7	53.3 / 53.3
hearsay	75.5 / 47.9	55.3 / 23.4	68.1 / 41.5
personal_jurisdiction	94.0 / 94.0	68.0 / 10.0	70.0 / 70.0
successor_liability	19.1 / 19.1	15.2 / 15.2	38.3 / 38.3
telemarketing_sales_rule	72.3 / 70.2	48.9 / 42.5	55.3 / 55.3
ucc_v_common_law	97.8 / 97.8	100.0 / 47.8	93.6 / 93.6

Table 59: Performance on rule-application tasks for commercial models. We report correctness/analysis.

Task	GPT-4	GPT-3.5	Claude-1
rule_qa	72.0	46.0	77.0
international_citizenship_questions	59.3	52.7	60.0
nys_judicial_ethics	78.3	74.3	79.1
citation_prediction_classification	71.3	54.6	61.1
citation_prediction_open	15.1	3.8	11.3

Table 60: Commercial models on rule-recall tasks.

Task	Flan	Llama-2	OPT	Vicuna	WizardLM
rule_qa	0.0	22.0	8.0	14.0	20.0
international_citizenship_questions	52.4	50.0	18.0	21.5	49.8
nys_judicial_ethics	69.1	67.6	63.3	61.3	63.8
citation_prediction_classification	56.5	47.2	52.8	50.0	52.8
citation_prediction_open	1.9	1.9	0.0	0.0	3.8

Table 61: 13B models on rule-recall tasks.

Task	Bloom	Falcon	Incite-Base	Incite-Inst.	Llama-2	MPT	OPT	Vicuna
rule_qa	6.0	8.0	18.0	28.5	22.0	20.0	6.0	0.0
international_ citizenship_questions	0.0	10.2	49.5	47.0	27.2	2.4	2.4	3.0
nys_judicial_ethics	62.5	55.8	62.7	54.9	67.5	57.2	59.7	52.3
citation_prediction_ classification	50.0	50.9	50.9	47.2	50.0	50.0	47.2	14.8
citation_prediction_open	1.9	0.0	0.0	0.0	1.9	0.0	0.0	0.0

Table 62: 7B models on rule-recall tasks.

Task	Bloom	Flan	Incite	Opt
rule_qa	0.0	0.0	14.0	2.0
international_citizenship_questions	0.0	50.0	18.9	3.0
nys_judicial_ethics	54.0	58.5	49.9	56.7
citation_prediction_classification	49.1	50.0	51.9	49.1
citation_prediction_open	0.0	0.0	0.0	0.0

Table 63: 3B models on rule-recall tasks.

Task	GPT-4	GPT-3.5	Claude-1
corporate_lobbying	81.7	59.1	75.8
learned_hands_benefits	87.9	62.1	66.7
learned_hands_business	81.6	58.6	47.7
learned_hands_consumer	76.2	59.3	58.1
learned_hands_courts	52.6	54.2	46.9
learned_hands_crime	81.0	62.4	59.0
learned_hands_divorce	84.0	59.3	53.3
learned_hands_domestic_violence	83.9	60.9	62.6
learned_hands_education	91.1	57.1	55.4
learned_hands_employment	69.9	67.7	49.3
learned_hands_estates	96.6	59.0	74.7
learned_hands_family	86.2	57.1	50.9
learned_hands_health	87.2	65.0	49.6
learned_hands_housing	85.0	63.9	58.7
learned_hands_immigration	98.5	79.9	73.1
learned_hands_torts	70.6	60.0	53.2
learned_hands_traffic	95.3	49.8	52.3

Table 64: Commercial models on issue-spotting tasks.

Task	Flan	Llama-2	OPT	Vicuna	WizardLM
corporate_lobbying	55.9	55.9	50.9	50.3	50.2
learned_hands_benefits	68.2	50.0	50.0	4.5	0.0
learned_hands_business	61.5	50.6	48.9	1.1	46.6
learned_hands_consumer	72.8	50.0	45.0	0.0	44.6
learned_hands_courts	58.9	49.5	50.0	50.0	0.0
learned_hands_crime	83.0	50.1	51.5	48.4	26.7
learned_hands_divorce	57.3	49.3	50.0	49.3	0.0
learned_hands_domestic_violence	68.4	48.9	50.0	0.0	0.0
learned_hands_education	89.3	50.0	46.4	50.0	26.8
learned_hands_employment	74.2	49.2	49.9	23.1	0.0
learned_hands_estates	67.4	50.0	68.5	49.4	38.8
learned_hands_family	5.4	50.1	63.6	49.3	0.0
learned_hands_health	66.8	49.6	63.7	50.0	38.1
learned_hands_housing	68.7	49.8	49.5	47.8	0.0
learned_hands_immigration	79.9	51.5	53.0	50.0	53.0
learned_hands_torts	60.6	50.0	50.2	49.8	46.1
learned_hands_traffic	84.2	49.8	58.6	10.8	38.7

Table 65: 13B models on issue-spotting tasks.

Task	Bloom	Falcon	Incite-Base	Incite-Inst.	Llama-2	MPT	OPT	Vicuna
corporate_lobbying	40.3	44.8	50.1	49.7	50.0	49.6	50.9	45.5
learned_hands_benefits	48.5	51.5	50.0	51.5	51.5	56.1	50.0	0.0
learned_hands_business	55.7	60.3	50.0	55.7	50.0	50.0	59.8	0.0
learned_hands_consumer	45.9	45.0	49.7	48.0	50.0	46.6	48.9	7.5
learned_hands_courts	53.6	50.5	50.5	48.4	49.5	57.8	49.5	0.0
learned_hands_crime	50.3	51.6	50.0	54.1	51.0	53.9	51.7	0.0
learned_hands_divorce	48.0	50.0	50.0	62.0	46.7	58.0	49.3	0.0
learned_hands_domestic_violence	48.9	48.9	51.7	59.8	51.1	49.4	50.6	0.0
learned_hands_education	53.6	62.5	50.0	60.7	50.0	53.6	48.2	0.0
learned_hands_employment	50.6	49.7	49.9	54.8	49.4	51.1	49.7	0.0
learned_hands_estates	51.7	51.1	50.0	46.1	50.6	62.4	55.6	0.0
learned_hands_family	55.0	57.3	49.6	60.8	50.2	56.0	53.3	0.0
learned_hands_health	49.1	52.7	50.0	59.3	52.2	57.1	53.5	0.0
learned_hands_housing	49.2	49.6	49.8	47.1	50.5	49.3	50.1	0.0
learned_hands_immigration	53.7	47.0	50.0	64.9	50.0	67.9	62.7	0.0
learned_hands_torts	51.2	48.8	50.0	56.0	50.0	48.4	50.0	0.0
learned_hands_traffic	54.7	50.0	49.8	54.3	50.2	56.5	57.2	13.3

Table 66: 7B models on issue-spotting tasks.

Task	Bloom	Flan	Incite	Opt
corporate_lobbying	26.2	51.9	47.6	51.0
learned_hands_benefits	43.9	43.9	47.0	53.0
learned_hands_business	51.1	49.4	47.7	44.8
learned_hands_consumer	39.7	70.8	50.2	38.1
learned_hands_courts	56.2	46.9	47.4	60.4
learned_hands_crime	49.7	60.6	51.9	51.7
learned_hands_divorce	58.0	50.0	61.3	56.0
learned_hands_domestic_violence	36.2	49.4	52.9	51.7
learned_hands_education	50.0	67.9	51.8	46.4
learned_hands_employment	50.7	46.6	48.0	50.1
learned_hands_estates	42.7	66.3	52.8	62.4
learned_hands_family	52.1	47.6	59.0	64.6
learned_hands_health	46.9	58.4	53.1	52.2
learned_hands_housing	51.2	51.7	47.1	51.8
learned_hands_immigration	59.0	73.1	53.7	56.7
learned_hands_torts	44.4	66.2	51.2	56.7
learned_hands_traffic	46.8	65.3	46.4	65.3

Table 67: 3B models on issue-spotting tasks.

Task	GPT-4	GPT-3.5	Claude-1
abercrombie	85.3	63.2	66.3
diversity_1	100.0	88.4	83.4
diversity_2	99.8	87.3	92.9
diversity_3	97.0	89.4	84.3
diversity_4	100.0	90.1	97.9
diversity_5	93.2	92.6	81.0
diversity_6	90.5	77.3	56.4
hearsay	83.8	69.2	76.4
personal_jurisdiction	91.4	63.3	81.9
successor_liability	57.1	52.3	72.7
telemarketing_sales_rule	82.4	63.1	71.9
ucc_v_common_law	98.8	100.0	88.8

Table 68: Commercial models on rule-conclusion tasks.

Task	Flan	Llama-2	OPT	Vicuna	WizardLM
abercrombie	42.1	40.0	0.0	22.1	44.2
diversity_1	76.0	50.0	55.0	25.5	60.5
diversity_2	59.4	62.1	53.9	48.4	57.1
diversity_3	78.6	65.8	50.0	52.6	77.8
diversity_4	53.2	68.3	49.6	55.4	82.5
diversity_5	53.5	57.7	52.4	52.7	53.2
diversity_6	50.0	50.0	48.1	50.0	54.3
hearsay	64.0	56.5	52.3	48.7	66.3
personal_jurisdiction	62.6	57.9	51.3	0.0	64.0
successor_liability	52.7	51.2	26.4	0.0	39.1
telemarketing_sales_rule	69.8	74.3	50.0	63.1	72.7
ucc_v_common_law	98.1	77.9	52.5	0.0	79.8

Table 69: 13B models on rule-conclusion tasks.

Task	Bloom	Falcon	Incite-Base	Incite-Inst.	Llama-2	MPT	OPT	Vicuna
abercrombie	17.9	24.2	27.4	34.7	32.6	34.7	17.9	2.1
diversity_1	47.6	51.7	56.8	61.5	73.4	54.9	58.0	55.3
diversity_2	50.0	56.3	50.5	65.7	50.0	50.0	49.8	49.1
diversity_3	51.9	53.6	50.0	50.8	62.8	50.0	49.9	49.3
diversity_4	58.3	68.7	50.0	67.9	72.9	50.0	49.7	49.7
diversity_5	57.9	50.0	50.4	49.4	51.3	50.0	50.0	46.8
diversity_6	50.0	50.0	50.0	44.7	50.0	50.0	49.7	50.0
hearsay	51.7	53.7	50.0	73.2	64.1	61.7	36.4	30.7
personal_jurisdiction	51.8	54.6	40.0	46.1	52.6	43.8	50.6	50.0
successor_liability	21.7	39.1	37.2	32.6	43.4	35.7	38.8	0.0
telemarketing_sales_rule	57.4	55.3	51.8	58.6	61.3	55.8	54.5	43.6
ucc_v_common_law	50.0	50.0	50.0	50.0	56.6	50.0	50.0	0.0

Table 70: 7B models on rule-conclusion tasks.

Task	Bloom	Flan	Incite	Opt
abercrombie	20.0	31.6	25.3	26.3
diversity_1	50.0	50.0	50.3	51.5
diversity_2	50.0	50.0	49.3	52.6
diversity_3	50.0	50.0	46.2	50.0
diversity_4	50.0	50.0	62.5	51.4
diversity_5	51.8	50.0	52.8	50.0
diversity_6	50.5	50.0	51.5	50.0
hearsay	51.2	57.5	57.0	49.7
personal_jurisdiction	50.0	51.1	46.1	50.0
successor_liability	21.7	44.6	24.8	14.4
telemarketing_sales_rule	44.5	51.4	53.5	55.9
ucc_v_common_law	50.0	88.8	50.0	50.0

Table 71: 3B models on rule-conclusion tasks.

Task	GPT-4	GPT-3.5	Claude-1
canada_tax_court_outcomes	98.9	80.0	76.9
definition_classification	96.6	80.2	87.9
definition_extraction	81.8	85.0	82.7
function_of_decision_section	43.3	35.2	37.6
legal_reasoning_causality	84.5	72.1	66.9
oral_argument_question_purpose	37.4	28.4	35.1
overruling	95.2	88.9	95.4
scalr	77.9	58.8	64.7
textualism_tool_dictionaries	93.9	65.1	71.2
textualism_tool_plain	84.7	73.2	70.2

Table 72: Commercial models on rhetorical-understanding tasks.

Task	Flan	Llama-2	OPT	Vicuna	WizardLM
canada_tax_court_outcomes	71.7	34.4	1.8	2.9	75.1
definition_classification	80.8	51.2	57.4	50.0	81.5
definition_extraction	80.5	85.0	80.1	62.4	82.0
function_of_decision_section	34.6	18.0	20.8	14.2	13.2
legal_reasoning_causality	78.6	52.8	52.3	48.4	46.2
oral_argument_question_purpose	24.5	20.0	15.0	16.1	29.3
overruling	94.2	92.3	78.3	52.7	89.5
scalr	66.5	56.7	20.2	5.0	45.6
textualism_tool_dictionaries	93.9	66.5	54.9	5.6	61.1
textualism_tool_plain	81.5	72.6	51.7	43.6	74.6

Table 73: 13B models on rhetorical-understanding tasks.

Task	Bloom	Falcon	Incite-Base	Incite-Inst.	Llama-2	MPT	OPT	Vicuna
canada_tax_court_ outcomes	0.0	0.0	0.0	5.2	35.4	0.0	28.5	0.0
definition_classification	54.4	80.2	50.1	65.4	50.0	57.5	58.3	42.5
definition_extraction	77.4	77.7	80.6	73.4	84.1	80.9	73.5	4.2
function_of_decision_ section	14.2	2.3	10.2	10.6	16.7	13.9	22.3	0.0
legal_reasoning_ causality	47.4	59.4	57.6	56.0	53.2	55.5	50.0	38.7
oral_argument_question_ purpose	13.7	15.5	20.2	27.0	14.5	17.2	14.4	1.2
overruling	82.3	67.9	50.3	75.2	92.2	72.4	53.3	28.7
scalr	18.6	20.7	22.0	23.0	33.7	25.7	19.3	0.0
textualism_tool_ dictionaries	48.5	60.9	55.7	59.6	47.4	59.5	51.0	19.8
textualism_tool_plain	50.0	57.2	61.6	55.9	50.0	60.5	51.0	4.6

Table 74: 7B models on rhetorical-understanding tasks.

Task	Bloom	Flan	Incite	Opt
canada_tax_court_outcomes	16.0	62.3	11.7	0.3
definition_classification	51.0	78.5	51.3	57.2
definition_extraction	59.1	77.6	59.4	70.6
function_of_decision_section	9.9	34.8	26.7	14.1
legal_reasoning_causality	50.1	67.5	49.5	55.7
oral_argument_question_purpose	5.3	19.9	21.5	14.3
overruling	63.4	93.6	54.4	54.8
scalr	17.7	64.5	21.5	20.0
textualism_tool_dictionaries	39.7	92.9	55.6	54.8
textualism_tool_plain	51.5	82.2	50.7	55.9

Table 75: 3B models on rhetorical-understanding tasks.

Table 76: Commercial models on interpretation tasks.

Task	**GPT-4**	**GPT-3.5**	**Claude-1**
consumer_contracts_qa	93.6	85.9	90.3
contract_nli_confidentiality_of_agreement	96.3	96.3	92.7
contract_nli_explicit_identification	82.4	81.1	65.0
contract_nli_inclusion_of_verbally_conveyed_information	90.7	83.0	83.4
contract_nli_limited_use	86.6	85.4	80.8
contract_nli_no_licensing	92.5	76.7	79.2
contract_nli_notice_on_compelled_disclosure	97.2	97.2	96.5
contract_nli_permissible_acquirement_of_similar_information	96.1	96.6	93.8
contract_nli_permissible_copy	80.4	77.7	72.8
contract_nli_permissible_development_of_similar_information	98.5	99.3	96.3
contract_nli_permissible_post-agreement_possession	94.6	89.3	92.2
contract_nli_return_of_confidential_information	95.6	92.5	89.4
contract_nli_sharing_with_employees	94.6	94.8	95.9
contract_nli_sharing_with_third-parties	93.3	75.0	86.6
contract_nli_survival_of_obligations	94.0	74.5	78.3
contract_qa	96.2	93.6	98.7
cuad_affiliate_license-licensee	90.9	90.9	85.9
cuad_affiliate_license-licensor	92.0	95.5	89.8
cuad_anti-assignment	91.4	89.1	92.4
cuad_audit_rights	97.9	89.5	92.7
cuad_cap_on_liability	95.6	94.1	92.9
cuad_change_of_control	88.9	89.7	89.2
cuad_competitive_restriction_exception	84.1	80.0	71.4
cuad_covenant_not_to_sue	95.8	88.0	89.3
cuad_effective_date	92.8	75.0	74.2
cuad_exclusivity	92.9	89.0	87.3
cuad_expiration_date	82.0	87.0	78.8
cuad_governing_law	99.3	98.3	98.7
cuad_insurance	99.2	95.3	94.9
cuad_ip_ownership_assignment	91.7	91.0	89.2
cuad_irrevocable_or_perpetual_license	97.5	95.4	88.6
cuad_joint_ip_ownership	94.3	91.1	85.4
cuad_license_grant	94.0	90.3	91.0
cuad_liquidated_damages	96.4	86.4	90.9
cuad_minimum_commitment	89.1	86.1	88.6
cuad_most_favored_nation	96.9	95.3	96.9
cuad_no-solicit_of_customers	100.0	98.8	92.9
cuad_no-solicit_of_employees	100.0	97.9	96.5
cuad_non-compete	93.0	91.0	90.5
cuad_non-disparagement	97.0	95.0	87.0

Table 76 – continued from previous page

Task	GPT-4	GPT-3.5	Claude-1
cuad_non-transferable_license	90.2	82.1	87.6
cuad_notice_period_to_terminate_renewal	95.9	97.7	96.4
cuad_post-termination_services	94.6	89.0	77.8
cuad_price_restrictions	95.7	87.0	89.1
cuad_renewal_term	96.1	95.9	95.6
cuad_revenue-profit_sharing	95.3	91.2	88.9
cuad_rofr-rofo-rofn	88.6	81.9	86.2
cuad_source_code_escrow	96.6	91.5	94.1
cuad_termination_for_convenience	96.7	94.2	95.8
cuad_third_party_beneficiary	89.7	83.8	82.4
cuad_uncapped_liability	85.4	70.4	74.1
cuad_unlimited-all-you-can-eat-license	93.8	93.8	87.5
cuad_volume_restriction	80.7	68.6	80.1
cuad_warranty_duration	77.8	81.2	78.1
insurance_policy_interpretation	69.6	55.0	64.0
jcrew_blocker	100.0	88.9	55.6
maud_ability_to_consummate_concept_is_subject_to_mae_carveouts	50.0	50.0	31.5
maud_financial_point_of_view_is_the_sole_consideration	50.0	38.8	50.0
maud_accuracy_of_fundamental_target_rws_bringdown_standard	29.3	33.3	10.4
maud_accuracy_of_target_general_rw_bringdown_timing_answer	63.6	51.0	46.9
maud_accuracy_of_target_capitalization_rw_(outstanding_shares)_bringdown_standard_answer	20.7	16.2	13.4
maud_additional_matching_rights_period_for_modifications_(cor)	57.4	43.3	24.5
maud_application_of_buyer_consent_requirement_(negative_interim_covenant)	63.7	68.8	35.3
maud_buyer_consent_requirement_(ordinary_course)	50.0	60.8	28.9
maud_change_in_law__subject_to_disproportionate_impact_modifier	53.0	48.3	65.2
maud_changes_in_gaap_or_other_accounting_principles__subject_to_disproportionate_impact_modifier	51.7	47.4	53.8
maud_cor_permitted_in_response_to_intervening_event	50.0	52.5	50.1
maud_cor_permitted_with_board_fiduciary_determination_only	21.4	50.0	50.6
maud_cor_standard_(intervening_event)	0.5	36.5	0.0
maud_cor_standard_(superior_offer)	40.5	45.5	0.0
maud_definition_contains_knowledge_requirement_-_answer	25.0	33.7	20.7
maud_definition_includes_asset_deals	33.3	30.5	0.5
maud_definition_includes_stock_deals	33.3	37.5	27.4
maud_fiduciary_exception__board_determination_standard	40.1	27.5	0.0
maud_fiduciary_exception_board_determination_trigger_(no_shop)	50.0	48.8	50.0
maud_fls_(mae)_standard	25.0	44.6	22.5
maud_general_economic_and_financial_conditions_subject_to_disproportionate_impact_modifier	54.2	56.0	53.6
maud_includes_consistent_with_past_practice	54.2	55.3	84.2

Table 76 – continued from previous page

Task	GPT-4	GPT-3.5	Claude-1
maud_initial_matching_rights_period_(cor)	15.4	31.9	20.7
maud_initial_matching_rights_period_(ftr)	49.4	32.8	14.0
maud_intervening_event_-_required_to_occur_after_signing_-_answer	51.9	51.4	12.5
maud_knowledge_definition	51.1	49.1	38.5
maud_liability_standard_for_no-shop_breach_by_target_non-do_representatives	44.2	51.9	49.4
maud_ordinary_course_efforts_standard	91.1	70.3	53.6
maud_pandemic_or_other_public_health_event__subject_to_disproportionate_impact_modifier	48.7	50.0	51.3
maud_pandemic_or_other_public_health_event_specific_reference_to_pandemic-related_governmental_responses_or_measures	79.5	70.9	60.8
maud_relational_language_(mae)_applies_to	57.9	47.2	26.0
maud_specific_performance	51.5	90.6	73.7
maud_tail_period_length	68.1	39.5	60.4
maud_type_of_consideration	99.5	82.7	75.1
opp115_data_retention	67.0	70.5	55.7
opp115_data_security	87.5	84.2	55.6
opp115_do_not_track	99.1	93.6	90.0
opp115_first_party_collection_use	76.7	80.6	63.0
opp115_international_and_specific_audiences	92.3	82.6	79.4
opp115_policy_change	91.9	89.3	83.8
opp115_third_party_sharing_collection	80.1	77.0	71.0
opp115_user_access,_edit_and_deletion	90.2	87.7	79.3
opp115_user_choice_control	82.9	79.3	71.3
privacy_policy_entailment	85.5	78.8	89.6
privacy_policy_qa	71.3	65.5	63.0
proa	99.0	90.6	88.5
ssla_company_defendants	65.0	65.3	16.5
ssla_individual_defendants	29.6	25.8	11.1
ssla_plaintiff	92.0	86.5	86.7
sara_entailment	86.8	68.4	67.6
sara_numeric	8.3	4.2	6.2
supply_chain_disclosure_best_practice_accountability	71.5	69.5	74.6
supply_chain_disclosure_best_practice_audits	74.4	76.6	75.5
supply_chain_disclosure_best_practice_certification	76.6	77.7	77.4
supply_chain_disclosure_best_practice_training	83.3	87.1	85.3
supply_chain_disclosure_best_practice_verification	68.3	59.4	64.3
supply_chain_disclosure_disclosed_accountability	77.0	80.4	75.5
supply_chain_disclosure_disclosed_audits	81.6	83.7	80.0
supply_chain_disclosure_disclosed_certification	71.2	67.3	75.8
supply_chain_disclosure_disclosed_training	89.1	83.0	75.6
supply_chain_disclosure_disclosed_verification	56.6	62.0	67.6

Table 76 – continued from previous page

Task	GPT-4	GPT-3.5	Claude-1
unfair_tos	9.1	13.7	5.5

Table 77: 13B models on interpretation tasks.

Task	Flan	LLaMA-2	OPT	Vicuna	WizardLM
consumer_contracts_qa	92.6	68.1	24.8	37.9	67.8
contract_nli_confidentiality_of_agreement	85.4	50.0	54.9	48.8	76.8
contract_nli_explicit_identification	81.4	49.4	51.6	50.0	67.0
contract_nli_inclusion_of_verbally_conveyed_information	56.2	50.0	53.0	50.0	58.2
contract_nli_limited_use	71.1	44.5	60.6	49.5	57.1
contract_nli_no_licensing	54.2	53.3	47.0	51.2	58.4
contract_nli_notice_on_compelled_disclosure	90.8	52.1	61.3	55.6	73.2
contract_nli_permissible_acquirement_of_similar_information	89.3	50.0	45.5	52.2	61.2
contract_nli_permissible_copy	84.3	47.8	47.8	50.0	56.2
contract_nli_permissible_development_of_similar_information	98.5	50.0	55.1	54.4	86.0
contract_nli_permissible_post-agreement_possession	93.4	48.8	44.0	50.0	52.5
contract_nli_return_of_confidential_information	89.4	50.0	44.3	51.3	66.2
contract_nli_sharing_with_employees	92.2	48.2	70.5	50.0	65.7
contract_nli_sharing_with_third-parties	86.7	49.5	50.9	50.5	64.4
contract_nli_survival_of_obligations	74.6	50.0	44.8	50.6	59.6
contract_qa	96.3	82.7	56.8	73.1	35.9
cuad_affiliate_license-licensee	83.8	58.1	49.0	50.0	65.7
cuad_affiliate_license-licensor	90.9	72.7	53.4	50.0	54.5
cuad_anti-assignment	85.0	52.2	48.1	50.2	76.8
cuad_audit_rights	87.1	51.6	71.7	50.7	58.6
cuad_cap_on_liability	81.5	74.0	0.6	50.0	57.5
cuad_change_of_control	75.5	56.2	52.9	49.0	74.8
cuad_competitive_restriction_exception	81.8	51.8	38.2	50.0	50.0
cuad_covenant_not_to_sue	86.0	66.6	49.4	50.0	56.5
cuad_effective_date	92.8	50.0	50.4	48.7	58.1
cuad_exclusivity	84.0	86.1	61.7	49.6	58.7
cuad_expiration_date	60.5	51.6	53.1	51.0	73.2
cuad_governing_law	99.5	90.6	71.7	59.5	54.3
cuad_insurance	90.2	55.5	66.3	55.0	55.7
cuad_ip_ownership_assignment	89.8	74.0	49.8	49.7	59.9
cuad_irrevocable_or_perpetual_license	95.7	85.4	67.9	50.0	73.9
cuad_joint_ip_ownership	76.0	53.6	50.5	50.0	60.4
cuad_license_grant	92.8	62.2	50.8	49.9	68.7
cuad_liquidated_damages	73.6	50.9	48.6	49.5	71.8
cuad_minimum_commitment	62.3	54.5	51.6	49.9	47.0
cuad_most_favored_nation	75.0	56.2	43.8	50.0	57.8

Table 77 – continued from previous page

Task	Flan	LLaMA-2	OPT	Vicuna	WizardLM
cuad_no-solicit_of_customers	97.6	57.1	56.0	50.0	85.7
cuad_no-solicit_of_employees	95.8	97.9	69.0	50.0	75.4
cuad_non-compete	91.6	52.7	46.8	49.3	55.2
cuad_non-disparagement	83.0	73.0	67.0	49.0	74.0
cuad_non-transferable_license	69.6	55.0	54.2	49.4	75.6
cuad_notice_period_to_terminate_renewal	92.8	50.0	54.5	50.0	83.3
cuad_post-termination_services	88.7	50.7	51.4	49.6	62.5
cuad_price_restrictions	76.1	71.7	50.0	47.8	50.0
cuad_renewal_term	89.1	50.5	50.3	57.5	83.9
cuad_revenue-profit_sharing	72.1	58.4	54.5	49.7	52.7
cuad_rofr-rofo-rofn	65.1	50.4	54.1	50.0	59.6
cuad_source_code_escrow	66.1	58.5	67.8	50.0	52.5
cuad_termination_for_convenience	91.4	50.9	49.8	55.8	84.2
cuad_third_party_beneficiary	85.3	54.4	63.2	47.1	69.1
cuad_uncapped_liability	50.3	55.4	46.6	51.0	74.5
cuad_unlimited-all-you-can-eat-license	83.3	52.1	64.6	47.9	54.2
cuad_volume_restriction	55.6	50.0	55.3	50.0	54.0
cuad_warranty_duration	74.7	57.5	55.9	50.6	63.1
insurance_policy_interpretation	44.8	46.6	38.0	13.2	51.9
jcrew_blocker	86.7	66.7	51.1	50.0	8.9
maud_ability_to_consummate_concept_is_subject_to_mae_carveouts	50.0	47.3	51.8	50.0	8.2
maud_financial_point_of_view_is_the_sole_consideration	53.6	51.0	50.0	53.1	49.5
maud_accuracy_of_fundamental_target_rws_bringdown_standard	12.5	33.3	33.3	31.7	33.3
maud_accuracy_of_target_general_rw_bringdown_timing_answer	46.6	50.0	49.1	49.8	50.0
maud_accuracy_of_target_capitalization_rw_(outstanding_shares)_ bringdown_standard_answer	6.1	25.8	26.9	20.9	24.8
maud_additional_matching_rights_period_for_modifications_(cor)	0.0	25.0	19.8	18.4	12.9
maud_application_of_buyer_consent_requirement_(negative_ interim_covenant)	63.4	47.8	45.0	3.1	49.4
maud_buyer_consent_requirement_(ordinary_course)	34.4	56.5	44.4	45.1	50.0
maud_change_in_law__subject_to_disproportionate_impact_ modifier	50.0	55.9	46.6	21.3	0.0
maud_changes_in_gaap_or_other_accounting_principles__subject_ to_disproportionate_impact_modifier	50.0	57.5	47.1	20.6	0.0
maud_cor_permitted_in_response_to_intervening_event	50.0	57.6	58.8	26.8	47.0
maud_cor_permitted_with_board_fiduciary_determination_only	50.0	50.0	51.2	46.7	42.0
maud_cor_standard_(intervening_event)	0.0	24.0	16.7	10.0	23.3
maud_cor_standard_(superior_offer)	11.9	16.8	3.0	0.0	26.9
maud_definition_contains_knowledge_requirement_-_answer	0.0	28.9	25.0	24.5	24.1
maud_definition_includes_asset_deals	2.8	35.4	32.7	0.9	34.2
maud_definition_includes_stock_deals	4.6	31.8	24.7	30.1	17.3
maud_fiduciary_exception__board_determination_standard	1.2	6.5	12.5	0.5	14.1

Table 77 – continued from previous page

Task	Flan	LLaMA-2	OPT	Vicuna	WizardLM
maud_fiduciary_exception_board_determination_trigger_(no_shop)	50.0	59.2	44.7	8.8	48.8
maud_fls_(mae)_standard	17.1	5.0	24.5	25.0	4.6
maud_general_economic_and_financial_conditions_subject_to_ disproportionate_impact_modifier	50.0	53.6	52.4	6.0	0.0
maud_includes_consistent_with_past_practice	62.3	50.0	52.6	53.2	61.0
maud_initial_matching_rights_period_(cor)	0.0	8.3	20.8	0.9	14.2
maud_initial_matching_rights_period_(ftr)	0.0	13.3	23.2	19.1	18.8
maud_intervening_event_-_required_to_occur_after_signing_-_ answer	39.2	46.2	48.9	47.1	47.1
maud_knowledge_definition	50.6	46.0	48.4	0.0	51.2
maud_liability_standard_for_no-shop_breach_by_target_non-do_ representatives	48.7	48.7	49.4	0.6	50.0
maud_ordinary_course_efforts_standard	81.1	57.8	33.7	0.0	67.3
maud_pandemic_or_other_public_health_event__subject_to_ disproportionate_impact_modifier	48.1	52.2	46.2	7.6	31.0
maud_pandemic_or_other_public_health_event_specific_reference_ to_pandemic-related_governmental_responses_or_measures	50.0	50.0	51.9	50.7	37.4
maud_relational_language_(mae)_applies_to	51.0	44.2	1.4	51.7	5.1
maud_specific_performance	94.9	52.1	50.0	0.0	56.7
maud_tail_period_length	25.8	51.5	52.5	13.5	34.6
maud_type_of_consideration	73.6	39.1	30.2	27.8	27.2
opp115_data_retention	55.7	51.1	45.5	50.0	63.6
opp115_data_security	75.1	49.8	51.2	55.5	59.4
opp115_do_not_track	79.1	50.0	42.7	51.8	88.2
opp115_first_party_collection_use	75.0	67.0	68.5	52.3	55.2
opp115_international_and_specific_audiences	80.1	59.5	18.2	50.4	66.2
opp115_policy_change	70.5	64.9	60.9	52.8	56.5
opp115_third_party_sharing_collection	71.4	54.9	58.8	52.4	60.9
opp115_user_access,_edit_and_deletion	75.4	54.3	60.9	49.1	59.4
opp115_user_choice_control	80.9	53.7	50.0	47.5	58.0
privacy_policy_entailment	58.9	56.2	50.1	0.6	65.9
privacy_policy_qa	52.5	50.5	50.9	0.0	55.6
proa	94.8	76.0	52.1	50.0	80.1
ssla_company_defendants	34.4	63.4	56.6	3.1	2.8
ssla_individual_defendants	21.9	21.6	16.2	0.0	0.0
ssla_plaintiff	88.8	26.4	31.6	0.0	0.0
sara_entailment	35.3	58.1	48.9	15.4	50.0
sara_numeric	1.0	0.0	0.0	0.0	0.0
supply_chain_disclosure_best_practice_accountability	73.6	49.4	48.8	67.2	52.1
supply_chain_disclosure_best_practice_audits	69.2	49.2	66.6	46.5	65.7
supply_chain_disclosure_best_practice_certification	69.9	51.6	56.1	67.2	64.8
supply_chain_disclosure_best_practice_training	81.0	49.7	49.4	71.5	64.9

Table 77 – continued from previous page

Task	Flan	LLaMA-2	OPT	Vicuna	WizardLM
supply_chain_disclosure_best_practice_verification	70.5	49.1	50.8	51.6	52.5
supply_chain_disclosure_disclosed_accountability	80.7	49.4	45.8	59.1	58.8
supply_chain_disclosure_disclosed_audits	83.3	49.3	48.9	64.1	64.7
supply_chain_disclosure_disclosed_certification	68.7	49.4	53.8	61.6	59.3
supply_chain_disclosure_disclosed_training	87.0	49.2	48.0	58.1	62.8
supply_chain_disclosure_disclosed_verification	64.3	49.3	48.6	58.0	50.4
unfair_tos	10.0	12.8	10.0	8.3	13.8

Table 78: 7B models on interpretation tasks.

Task	BLOOM	Falcon	Incite-Base	Incite-Inst.	LLaMA-2	MPT	OPT	Vicuna
consumer_contracts_qa	0.6	57.9	42.4	49.0	63.5	40.5	14.7	34.0
contract_nli_confidentiality_of_agreement	53.7	64.6	65.9	59.8	50.0	52.4	57.3	45.1
contract_nli_explicit_identification	49.4	66.1	61.6	68.9	50.0	50.0	59.9	34.8
contract_nli_inclusion_of_verbally_conveyed_information	50.0	67.5	66.7	60.9	50.0	49.3	55.5	33.1
contract_nli_limited_use	56.2	46.0	59.0	66.3	51.1	61.8	61.7	24.8
contract_nli_no_licensing	49.4	44.6	51.9	56.3	48.2	45.9	49.5	15.9
contract_nli_notice_on_compelled_disclosure	51.4	51.4	64.1	62.7	50.7	65.5	68.3	27.5
contract_nli_permissible_acquirement_of_similar_information	49.4	31.5	26.4	47.8	50.0	53.4	50.0	39.9
contract_nli_permissible_copy	49.9	36.8	46.4	57.6	58.9	55.0	49.6	33.2
contract_nli_permissible_development_of_similar_information	49.3	43.4	59.6	44.1	50.0	54.4	53.7	41.9
contract_nli_permissible_post-agreement_possession	48.2	43.1	53.2	55.6	50.0	55.3	44.5	2.4
contract_nli_return_of_confidential_information	50.0	57.7	50.2	70.0	50.0	56.5	75.6	5.9
contract_nli_sharing_with_employees	55.6	61.6	48.2	70.3	50.0	48.1	58.9	6.1
contract_nli_sharing_with_third-parties	49.5	48.1	39.3	53.7	50.0	48.4	48.7	11.9
contract_nli_survival_of_obligations	49.5	55.1	41.2	43.2	50.0	41.7	49.3	1.8
contract_qa	14.8	7.7	11.4	87.7	31.9	9.0	39.0	50.3
cuad_affiliate_license-licensee	53.0	50.5	68.7	71.7	50.0	69.2	66.2	34.3
cuad_affiliate_license-licensor	45.5	70.5	64.8	85.2	50.0	76.1	52.3	38.6
cuad_anti-assignment	47.7	33.2	60.4	76.2	58.8	56.5	55.1	37.8
cuad_audit_rights	67.4	59.7	80.0	71.5	50.7	52.5	72.8	28.2

Table 78 – continued from previous page

Task	BLOOM	Falcon	Incite-Base	Incite-Inst.	LLaMA-2	MPT	OPT	Vicuna
cuad_cap_on_liability	41.8	42.1	40.4	57.4	50.2	70.3	37.0	16.8
cuad_change_of_control	55.0	49.8	57.0	67.8	50.0	56.0	55.5	39.2
cuad_competitive_restriction_exception	37.3	30.9	50.0	47.7	48.6	46.8	36.4	16.4
cuad_covenant_not_to_sue	47.7	48.1	67.9	78.2	57.5	71.8	50.0	0.0
cuad_effective_date	56.8	60.2	53.0	46.6	50.0	44.5	65.7	6.4
cuad_exclusivity	62.9	57.1	66.4	65.4	63.4	56.7	60.0	14.4
cuad_expiration_date	75.9	62.7	55.5	67.6	50.1	52.9	85.0	10.5
cuad_governing_law	79.3	77.3	28.5	68.2	55.8	66.3	73.1	17.6
cuad_insurance	83.2	66.8	65.4	72.0	50.1	72.7	77.2	39.1
cuad_ip_ownership_assignment	53.0	65.8	61.1	73.8	51.6	60.4	70.8	14.4
cuad_irrevocable_or_perpetual_license	57.5	58.6	72.1	83.2	52.9	59.6	80.0	46.1
cuad_joint_ip_ownership	55.2	58.3	60.9	74.5	50.0	55.2	72.4	30.2
cuad_license_grant	62.7	60.3	65.5	77.0	63.0	39.0	67.9	20.8
cuad_liquidated_damages	54.1	65.0	65.0	70.5	50.0	51.4	57.7	36.4
cuad_minimum_commitment	51.4	58.4	55.3	53.8	49.9	55.3	59.2	31.3
cuad_most_favored_nation	51.6	48.4	60.9	59.4	51.6	57.8	48.4	23.4
cuad_no-solicit_of_customers	47.6	38.1	61.9	77.4	50.0	51.2	27.4	22.6
cuad_no-solicit_of_employees	39.4	35.9	54.9	80.3	69.0	70.4	42.3	19.7
cuad_non-compete	32.8	42.5	55.2	67.4	63.3	52.3	29.0	42.3
cuad_non-disparagement	41.0	50.0	64.0	70.0	64.0	51.0	47.0	34.0
cuad_non-transferable_license	67.2	65.9	56.5	78.0	50.0	48.5	68.5	27.9
cuad_notice_period_to_terminate_renewal	50.0	58.6	50.5	76.6	50.0	50.9	81.5	35.6
cuad_post-termination_services	48.1	45.7	60.0	57.2	50.0	64.7	57.7	20.0
cuad_price_restrictions	58.7	43.5	45.7	58.7	50.0	54.3	52.2	41.3
cuad_renewal_term	50.3	57.3	45.1	75.6	50.0	53.1	82.1	35.2
cuad_revenue-profit_sharing	54.3	57.9	50.6	65.2	52.3	50.0	62.4	11.1
cuad_rofr-rofo-rofn	54.6	43.3	55.9	57.0	50.0	59.9	53.0	20.6
cuad_source_code_escrow	75.4	59.3	62.7	65.3	52.5	51.7	79.7	24.6
cuad_termination_for_convenience	70.5	67.7	47.9	83.7	50.0	49.8	64.4	42.8
cuad_third_party_beneficiary	70.6	50.0	69.1	79.4	50.0	67.6	70.6	29.4
cuad_uncapped_liability	46.9	53.4	53.4	61.2	51.0	77.9	36.4	33.0
cuad_unlimited-all-you-can-eat-license	72.9	62.5	75.0	79.2	62.5	60.4	72.9	41.7
cuad_volume_restriction	52.8	47.5	55.0	54.0	50.9	52.5	63.0	42.2
cuad_warranty_duration	67.5	59.4	57.8	64.1	51.9	54.4	65.3	5.3
insurance_policy_interpretation	34.5	42.8	35.4	36.9	43.2	34.9	33.7	29.1

Table 78 – continued from previous page

Task	BLOOM	Falcon	Incite-Base	Incite-Inst.	LLaMA-2	MPT	OPT	Vicuna
jcrew_blocker	45.6	46.7	47.8	51.1	58.9	57.8	56.7	11.1
maud_ability_to_consummate_concept_is_subject_to_mae_carveouts	42.3	48.2	30.9	50.0	50.0	0.9	50.5	45.5
maud_financial_point_of_view_is_the_sole_consideration	48.5	56.6	45.9	43.9	48.5	4.1	63.3	50.0
maud_accuracy_of_fundamental_target_rws_bringdown_standard	33.3	34.5	31.7	35.9	34.4	31.6	37.5	33.3
maud_accuracy_of_target_general_rw_bringdown_timing_answer	50.0	45.8	53.7	48.3	51.7	50.8	53.9	50.0
maud_accuracy_of_target_capitalization_rw_(outstanding_shares)_bringdown_standard_answer	30.7	30.2	25.1	20.2	22.1	22.3	24.7	18.1
maud_additional_matching_rights_period_for_modifications_(cor)	19.5	18.3	21.1	16.3	22.0	8.4	20.3	20.0
maud_application_of_buyer_consent_requirement_(negative_interim_covenant)	47.8	40.0	50.0	55.6	50.9	50.0	47.5	43.1
maud_buyer_consent_requirement_(ordinary_course)	50.6	54.1	49.7	50.0	50.0	7.4	57.5	50.0
maud_change_in_law__subject_to_disproportionate_impact_modifier	36.4	45.8	11.2	49.4	50.6	12.2	50.0	50.0
maud_changes_in_gaap_or_other_accounting_principles__subject_to_disproportionate_impact_modifier	35.9	48.4	10.0	53.3	50.6	7.1	50.0	50.0
maud_cor_permitted_in_response_to_intervening_event	47.6	62.5	47.5	49.4	58.6	7.3	65.6	50.7
maud_cor_permitted_with_board_fiduciary_determination_only	45.8	42.1	37.2	48.8	50.0	13.5	49.4	49.4
maud_cor_standard_(intervening_event)	26.7	9.3	10.9	16.7	21.6	13.8	16.7	16.7
maud_cor_standard_(superior_offer)	16.8	7.8	7.0	15.8	17.8	9.7	11.5	6.4
maud_definition_contains_knowledge_requirement_-_answer	25.3	26.9	20.8	23.1	22.8	28.0	25.3	25.0
maud_definition_includes_asset_deals	23.9	29.4	18.2	28.9	32.8	6.8	35.1	28.6
maud_definition_includes_stock_deals	27.7	16.4	5.0	24.5	18.4	7.6	17.2	22.6
maud_fiduciary_exception__board_determination_standard	4.3	9.6	6.9	12.7	14.6	9.9	13.7	1.0

Table 78 – continued from previous page

Task	BLOOM	Falcon	Incite-Base	Incite-Inst.	LLaMA-2	MPT	OPT	Vicuna
maud_fiduciary_exception_board_determination_trigger_(no_shop)	48.8	44.7	43.3	49.3	55.8	20.1	57.0	42.1
maud_fls_(mae)_standard	38.3	16.0	13.0	28.6	1.8	1.8	24.6	0.0
maud_general_economic_and_financial_conditions_subject_to_disproportionate_impact_modifier	54.8	62.5	38.1	48.8	50.6	14.3	50.0	50.0
maud_includes_consistent_with_past_practice	50.0	48.2	61.6	49.6	52.5	5.4	52.8	50.0
maud_initial_matching_rights_period_(cor)	20.7	11.2	21.5	23.4	27.2	10.8	21.0	18.9
maud_initial_matching_rights_period_(ftr)	20.6	14.5	16.4	11.1	22.8	7.7	20.0	16.9
maud_intervening_event_-_required_to_occur_after_signing_-_answer	50.2	51.5	43.4	50.0	42.6	43.5	47.5	50.0
maud_knowledge_definition	47.1	46.8	40.0	46.7	51.8	49.3	50.7	45.5
maud_liability_standard_for_no-shop_breach_by_target_non-do_representatives	59.6	50.0	50.0	53.8	50.0	46.8	58.3	50.0
maud_ordinary_course_efforts_standard	34.2	32.8	41.9	32.9	77.4	32.7	34.2	33.8
maud_pandemic_or_other_public_health_event__subject_to_disproportionate_impact_modifier	48.2	10.8	49.4	48.9	51.3	8.5	47.5	49.3
maud_pandemic_or_other_public_health_event_specific_reference_to_pandemic-related_governmental_responses_or_measures	41.7	51.9	46.0	48.1	50.0	4.2	50.8	50.0
maud_relational_language_(mae)_applies_to	37.0	47.1	18.8	50.0	7.1	16.5	28.3	46.4
maud_specific_performance	60.4	50.0	49.2	43.5	58.5	0.6	39.4	50.0
maud_tail_period_length	6.9	32.0	25.3	31.4	63.9	14.2	8.6	3.1
maud_type_of_consideration	25.0	28.5	27.1	26.4	25.3	39.3	24.3	25.0
opp115_data_retention	48.9	37.5	46.6	50.0	50.0	50.0	51.1	42.0
opp115_data_security	53.4	60.3	53.8	63.6	50.2	49.8	63.7	45.9
opp115_do_not_track	43.6	34.5	47.3	69.1	50.0	49.1	31.8	45.5
opp115_first_party_collection_use	63.9	59.3	69.5	69.9	61.3	59.2	70.4	46.4
opp115_international_and_specific_audiences	58.1	60.3	50.1	64.9	51.1	52.7	57.5	37.4
opp115_policy_change	68.2	66.2	55.6	55.9	50.0	50.0	74.5	44.0
opp115_third_party_sharing_collection	53.8	57.7	50.9	68.4	50.4	50.3	54.1	36.5
opp115_user_access,_edit_and_deletion	60.2	51.5	56.1	64.5	54.7	50.8	59.0	41.7

Table 78 – continued from previous page

Task	BLOOM	Falcon	Incite-Base	Incite-Inst.	LLaMA-2	MPT	OPT	Vicuna
opp115_user_choice_control	63.6	40.0	46.5	64.5	50.2	48.6	47.4	44.2
privacy_policy_entailment	51.4	49.3	56.8	58.1	57.7	64.1	52.8	26.6
privacy_policy_qa	50.0	49.7	52.0	56.3	50.2	50.0	50.4	0.1
proa	52.1	56.2	50.0	71.6	52.1	58.3	51.0	47.9
ssla_company_defendants	35.5	44.9	54.8	54.1	60.7	51.0	50.9	0.0
ssla_individual_defendants	14.3	17.4	20.8	19.6	23.1	20.8	13.9	0.0
ssla_plaintiff	26.1	9.3	52.3	64.4	60.5	76.0	59.8	0.0
sara_entailment	50.4	50.0	50.0	51.1	50.0	50.0	50.0	0.0
sara_numeric	2.1	1.0	2.1	0.0	0.0	2.1	1.0	0.0
supply_chain_disclosure_ best_practice_accountability	6.2	55.8	50.0	58.4	50.4	56.8	32.7	31.5
supply_chain_disclosure_ best_practice_audits	1.5	71.6	55.9	63.0	55.8	42.2	31.1	8.7
supply_chain_disclosure_ best_practice_certification	2.8	67.1	52.1	57.6	52.3	40.8	51.5	11.3
supply_chain_disclosure_ best_practice_training	6.6	52.3	50.9	55.4	49.3	56.2	50.5	30.7
supply_chain_disclosure_ best_practice_verification	4.5	55.8	49.8	54.3	49.1	36.0	38.1	13.9
supply_chain_disclosure_ disclosed_accountability	2.1	48.2	49.7	48.7	49.0	24.6	20.4	15.7
supply_chain_disclosure_ disclosed_audits	3.6	50.2	48.7	49.6	51.5	52.0	49.7	16.3
supply_chain_disclosure_ disclosed_certification	3.2	50.5	52.0	52.2	54.9	45.8	32.3	20.1
supply_chain_disclosure_ disclosed_training	5.6	48.6	49.1	49.2	49.3	44.0	39.5	10.1
supply_chain_disclosure_ disclosed_verification	5.3	48.9	49.6	47.6	48.6	48.5	49.3	23.1
unfair_tos	8.3	12.3	12.7	8.7	10.6	11.0	10.5	0.0

Table 79: 3B models on interpretation tasks.

Task	BLOOM	Flan	Incite	Opt
consumer_contracts_qa	0.8	93.2	46.1	31.7
contract_nli_confidentiality_of_agreement	57.3	86.6	56.1	46.3
contract_nli_explicit_identification	57.2	81.2	63.6	53.3
contract_nli_inclusion_of_verbally_conveyed_information	50.1	82.2	67.8	60.9
contract_nli_limited_use	63.1	72.8	60.1	63.0
contract_nli_no_licensing	50.0	65.1	43.5	38.5
contract_nli_notice_on_compelled_disclosure	57.0	66.2	65.5	71.8
contract_nli_permissible_acquirement_of_similar_information	59.6	69.1	46.6	54.5
contract_nli_permissible_copy	51.3	83.7	55.3	56.2
contract_nli_permissible_development_of_similar_information	60.3	75.0	49.3	59.6

Table 79 – continued from previous page

Task	BLOOM	Flan	Incite	Opt
contract_nli_permissible_post-agreement_possession	43.8	66.3	43.4	42.4
contract_nli_return_of_confidential_information	51.6	82.3	61.3	51.9
contract_nli_sharing_with_employees	53.3	69.7	47.1	52.8
contract_nli_sharing_with_third-parties	50.0	80.5	51.4	49.6
contract_nli_survival_of_obligations	50.0	72.9	49.4	44.1
contract_qa	35.9	0.0	82.9	44.6
cuad_affiliate_license-licensee	59.6	66.7	60.1	68.7
cuad_affiliate_license-licensor	48.9	77.3	71.6	50.0
cuad_anti-assignment	50.0	74.7	51.5	33.7
cuad_audit_rights	50.0	56.3	56.2	64.6
cuad_cap_on_liability	49.2	50.9	49.4	26.8
cuad_change_of_control	51.4	59.6	60.8	47.4
cuad_competitive_restriction_exception	46.4	73.2	45.5	40.5
cuad_covenant_not_to_sue	50.0	59.1	49.4	44.8
cuad_effective_date	51.7	92.4	47.0	65.7
cuad_exclusivity	64.6	55.2	53.9	60.9
cuad_expiration_date	50.2	67.7	52.1	69.3
cuad_governing_law	56.8	84.4	50.5	74.8
cuad_insurance	53.5	53.9	57.5	70.3
cuad_ip_ownership_assignment	49.0	57.6	57.3	65.1
cuad_irrevocable_or_perpetual_license	51.8	76.1	53.6	71.1
cuad_joint_ip_ownership	48.4	65.1	64.6	68.8
cuad_license_grant	50.1	69.1	51.4	69.2
cuad_liquidated_damages	50.0	52.3	57.3	34.5
cuad_minimum_commitment	52.7	54.1	50.6	57.0
cuad_most_favored_nation	50.0	57.8	56.2	43.8
cuad_no-solicit_of_customers	52.4	47.6	59.5	32.1
cuad_no-solicit_of_employees	48.6	63.4	66.9	42.3
cuad_non-compete	48.0	63.1	56.1	30.3
cuad_non-disparagement	49.0	57.0	60.0	44.0
cuad_non-transferable_license	56.8	58.7	50.9	62.0
cuad_notice_period_to_terminate_renewal	59.5	61.3	54.5	68.5
cuad_post-termination_services	50.0	64.0	50.7	55.9
cuad_price_restrictions	56.5	50.0	58.7	47.8
cuad_renewal_term	50.3	70.7	59.8	73.3
cuad_revenue-profit_sharing	54.1	60.9	52.5	64.1
cuad_rofr-rofo-rofn	50.3	49.4	51.6	55.4
cuad_source_code_escrow	55.9	42.4	70.3	56.8
cuad_termination_for_convenience	50.0	64.4	58.1	38.4
cuad_third_party_beneficiary	58.8	79.4	50.0	64.7

Table 79 – continued from previous page

Task	BLOOM	Flan	Incite	Opt
cuad_uncapped_liability	50.0	53.4	49.7	28.9
cuad_unlimited-all-you-can-eat-license	66.7	81.2	54.2	70.8
cuad_volume_restriction	50.3	49.4	57.5	59.6
cuad_warranty_duration	53.1	61.9	51.6	54.4
insurance_policy_interpretation	32.2	38.8	37.1	32.9
jcrew_blocker	55.6	56.7	48.9	53.3
maud_ability_to_consummate_concept_is_subject_to_mae_carveouts	47.3	13.4	50.0	49.1
maud_financial_point_of_view_is_the_sole_consideration	53.1	49.0	50.0	43.4
maud_accuracy_of_fundamental_target_rws_bringdown_standard	32.6	0.0	32.6	34.8
maud_accuracy_of_target_general_rw_bringdown_timing_answer	50.0	0.0	48.2	53.2
maud_accuracy_of_target_capitalization_rw_(outstanding_shares)_bringdown_standard_answer	26.1	0.0	23.4	28.2
maud_additional_matching_rights_period_for_modifications_(cor)	20.0	0.0	18.8	20.3
maud_application_of_buyer_consent_requirement_(negative_interim_covenant)	49.1	62.5	50.3	50.0
maud_buyer_consent_requirement_(ordinary_course)	49.4	50.0	50.0	49.7
maud_change_in_law__subject_to_disproportionate_impact_modifier	44.3	6.3	44.0	51.1
maud_changes_in_gaap_or_other_accounting_principles__subject_to_disproportionate_impact_modifier	41.9	5.9	51.0	47.9
maud_cor_permitted_in_response_to_intervening_event	51.2	48.8	71.7	50.6
maud_cor_permitted_with_board_fiduciary_determination_only	49.4	36.3	50.0	45.9
maud_cor_standard_(intervening_event)	16.7	1.0	16.7	16.7
maud_cor_standard_(superior_offer)	10.8	4.1	8.8	5.4
maud_definition_contains_knowledge_requirement_-_answer	24.7	0.0	26.2	25.0
maud_definition_includes_asset_deals	34.4	0.0	22.5	33.3
maud_definition_includes_stock_deals	40.4	0.0	9.9	0.0
maud_fiduciary_exception__board_determination_standard	6.3	0.5	12.9	13.3
maud_fiduciary_exception_board_determination_trigger_(no_shop)	56.7	50.0	50.5	50.0
maud_fls_(mae)_standard	23.9	10.5	24.2	1.2
maud_general_economic_and_financial_conditions_subject_to_disproportionate_impact_modifier	50.0	3.0	66.1	53.0
maud_includes_consistent_with_past_practice	50.4	89.7	50.0	43.3
maud_initial_matching_rights_period_(cor)	13.5	0.0	18.0	17.0
maud_initial_matching_rights_period_(ftr)	18.9	0.0	19.6	19.4
maud_intervening_event_-_required_to_occur_after_signing_-_answer	48.5	3.3	50.0	50.0
maud_knowledge_definition	46.0	50.0	47.5	49.4
maud_liability_standard_for_no-shop_breach_by_target_non-do_representatives	50.0	54.5	50.0	50.0
maud_ordinary_course_efforts_standard	33.3	58.3	32.1	33.3
maud_pandemic_or_other_public_health_event__subject_to_disproportionate_impact_modifier	2.6	62.3	60.4	47.4
maud_pandemic_or_other_public_health_event_specific_reference_to_pandemic-related_governmental_responses_or_measures	50.0	59.9	50.0	50.0
maud_relational_language_(mae)_applies_to	32.6	0.0	50.0	49.3

Table 79 – continued from previous page

Task	BLOOM	Flan	Incite	Opt
maud_specific_performance	50.0	89.3	48.0	50.0
maud_tail_period_length	1.5	0.0	35.6	25.0
maud_type_of_consideration	24.4	2.1	27.2	25.2
opp115_data_retention	45.5	52.3	64.8	39.8
opp115_data_security	60.7	69.7	55.9	53.6
opp115_do_not_track	45.5	55.5	42.7	37.3
opp115_first_party_collection_use	60.9	52.3	69.5	56.1
opp115_international_and_specific_audiences	59.9	70.5	51.8	57.0
opp115_policy_change	52.1	60.4	61.2	57.4
opp115_third_party_sharing_collection	55.6	63.3	64.8	47.0
opp115_user_access,_edit_and_deletion	49.2	82.3	66.4	53.5
opp115_user_choice_control	48.5	58.5	59.0	47.5
privacy_policy_entailment	50.0	53.6	66.1	54.1
privacy_policy_qa	50.3	61.0	55.1	50.5
proa	54.1	80.1	64.5	53.1
ssla_company_defendants	36.2	13.1	51.1	47.0
ssla_individual_defendants	11.9	20.0	20.5	17.3
ssla_plaintiff	40.1	86.4	36.8	42.8
sara_entailment	48.5	0.0	50.7	48.9
sara_numeric	3.1	1.0	0.0	1.0
supply_chain_disclosure_best_practice_accountability	43.0	74.5	55.1	37.4
supply_chain_disclosure_best_practice_audits	42.7	74.3	61.0	54.6
supply_chain_disclosure_best_practice_certification	46.8	77.4	52.0	46.6
supply_chain_disclosure_best_practice_training	44.9	83.4	64.2	41.0
supply_chain_disclosure_best_practice_verification	47.7	54.9	53.1	21.8
supply_chain_disclosure_disclosed_accountability	46.9	77.7	60.6	14.6
supply_chain_disclosure_disclosed_audits	46.5	73.0	57.3	2.6
supply_chain_disclosure_disclosed_certification	45.3	77.3	52.0	12.6
supply_chain_disclosure_disclosed_training	47.9	87.2	58.8	16.9
supply_chain_disclosure_disclosed_verification	44.1	66.1	55.3	0.2
unfair_tos	9.0	8.2	4.0	15.3

LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain

Joel Niklaus [1,2,6*] **Veton Matoshi** [2*] **Pooja Rani** [3]

Andrea Galassi [4] **Matthias Stürmer** [1,2] **Ilias Chalkidis** [5]

[1]University of Bern [2]Bern University of Applied Sciences [3]University of Zurich
[4]University of Bologna [5]University of Copenhagen [6]Stanford University

Abstract

Lately, propelled by the phenomenal advances around the transformer architecture, the legal NLP field has enjoyed spectacular growth. To measure progress, well curated and challenging benchmarks are crucial. However, most benchmarks are English only and in legal NLP specifically there is no multilingual benchmark available yet. Additionally, many benchmarks are saturated, with the best models clearly outperforming the best humans and achieving near perfect scores. We survey the legal NLP literature and select 11 datasets covering 24 languages, creating LEXTREME. To provide a fair comparison, we propose two aggregate scores, one based on the datasets and one on the languages. The best baseline (XLM-R large) achieves both a dataset aggregate score a language aggregate score of 61.3. This indicates that LEXTREME is still very challenging and leaves ample room for improvement. To make it easy for researchers and practitioners to use, we release LEXTREME on huggingface together with all the code required to evaluate models and a public Weights and Biases project with all the runs.

1 Introduction

In the last decade, the discipline of Natural Language Processing (NLP) has become more and more relevant for Legal Artificial Intelligence, leading to a shift from symbolic to subsymbolic techniques (Villata et al., 2022). Such a change can be motivated partially by the nature of legal resources, which appear mostly in a textual format (legislation, legal proceedings, contracts, etc.).

Following closely the advances in the development of NLP technologies, the legal NLP literature (Zhong et al., 2020; Aletras et al., 2022; Katz et al., 2023) is flourishing with the release of many new resources, including large legal corpora (Henderson et al., 2022), task-specific datasets (Chalkidis et al., 2021a; Shen et al., 2022), and pre-trained legal-oriented language models (PLMs) (Chalkidis et al., 2020; Zheng et al., 2021; Xiao et al., 2021; Niklaus and Giofré, 2022).

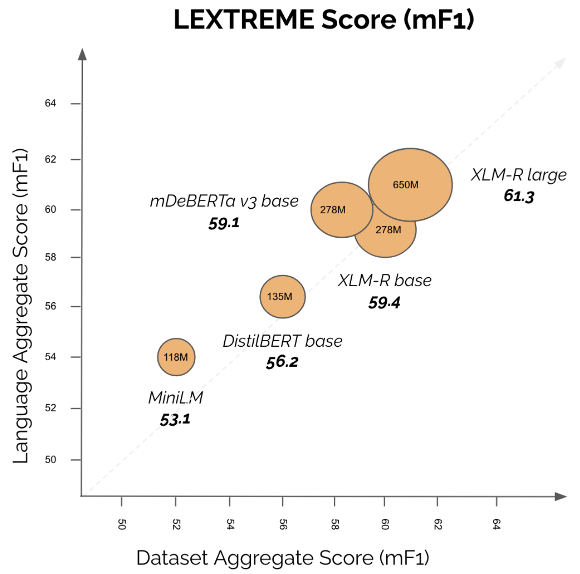


Figure 1: Overview of the multilingual models on the LEXTREME benchmark. The bubble size and text inside indicate the parameter count.

In particular, the development and spread of the so-called Foundation Models (Bommasani et al., 2022), large neural networks trained on vast corpora, led to massive performance improvements on popular benchmarks such as GLUE (Wang et al., 2019b) or SuperGLUE (Wang et al., 2019a). This exemplifies the need for more challenging benchmarks to continually measure progress. Legal benchmark suites (Chalkidis et al., 2022a; Hwang et al., 2022) to evaluate the performance of PLMs in a more systematic way have been also developed, showcasing the superiority of legal-oriented PLMs over generic ones on downstream tasks.

However, general-purpose models, trained on resources such as Wikipedia, may be insufficient to address tasks in the legal domain. Indeed, such a domain is strongly characterized both by its lexicon and by specific knowledge typically not available

[*] Equal contribution.

outside of specialized domain resources. Laypeople even sometimes call the language used in legal documents "legalese" or "legal jargon", emphasizing its complexity. It is therefore necessary to develop specialized Legal Language Models, to be trained on large collections of legal documents, and to be evaluated on proper legal benchmarks.

Existing benchmarks, such as GLUE, often tackle linguistic tasks, such as semantic textual similarity or natural language inference, with no direct application in mind. There is a need for benchmarks that tackle use cases as close as possible to the real world to align model development with practical deployment needs.

The rising need to build NLP systems for languages different from English, the scarcity of textual resources for those languages and the spread of code-switching in many cultures (Torres Cacoullos, 2020) has pushed researchers to design new multilingual learning approaches. This, in turn, has brought the necessity to develop proper multilingual benchmarks to evaluate multilingual language models (Conneau et al., 2020). This is of paramount importance for legal NLP, especially in case of inherently multinational (European Union, Council of Europe), or multilingual (Canada, Switzerland) legal systems.

In this work, we propose a challenging multilingual benchmark for the legal domain containing datasets with valuable use cases, calling it LEXTREME. We survey the literature and select 11 datasets out of 108 papers based on our exclusion and inclusion criteria. We evaluate five popular multilingual encoder-based language models and find that model size correlates well with performance on LEXTREME. For easy evaluation, we release the aggregate dataset on the huggingface hub[1] and the code to run experiments on GitHub.[2]

Contributions

The contributions of this paper are two-fold:

1. We review the literature for suitable legal datasets and compile a multilingual legal benchmark of 11 datasets in 24 languages.

2. We evaluate various baselines on LEXTREME to provide a reference point for researchers and practitioners to compare to.

[1] https://huggingface.co/datasets/joelito/lextreme

[2] https://github.com/JoelNiklaus/LEXTREME

2 Related Work

2.1 Benchmarks for Language Models

GLUE (Wang et al., 2019b) is one of the first benchmarks for the evaluation of general-purpose neural language models. It is a set of supervised sentence understanding predictive tasks in the English language that was created through aggregation and curation of already existing datasets. GLUE became quickly obsolete with the advent of advanced contextual language models such as BERT (Devlin et al., 2019), which performed extremely well on most of them. SUPERGLUE (Wang et al., 2019a) was later proposed as an updated version of GLUE, including new predictive tasks that are solvable by humans but are difficult for machines. Both benchmarks proposed an evaluation computed as the aggregation of the scores obtained by the same model on each task. They are also agnostic regarding the pre-training of the model, and do not provide a specific corpus for it. Following this trend, many other benchmarks have been proposed, Table 1 provides an overview of the most popular ones.

MMLU (Hendrycks et al., 2021) is specifically designed to evaluate the knowledge acquired during pre-training of the model by including only zero-shot and few-shot learning tasks. It contains about 16K multiple-choice questions divided into 57 subtasks, covering subjects in the humanities, social sciences, hard sciences, and other areas.

SUPERB (Yang et al., 2021) and SUPERB-SG (Tsai et al., 2022) were proposed for speech data, unifying popular datasets. They mainly differ in SUPERB-SG not including only predictive tasks but also generative ones, a characteristic that makes it different from all the other benchmarks discussed in this section. Another important difference is that SUPERB-SG includes tasks such as speech translation and cross-lingual automatic speech recognition, for which knowledge of languages other than English is beneficial. Neither of the two proposes an aggregated score.

XTREME (Hu et al., 2020) is a benchmark specifically designed to evaluate the ability of cross-lingual generalization of models. It includes 6 cross-lingual predictive tasks over 10 datasets of miscellaneous texts, covering a total of 40 languages. While some original datasets were already designed for cross-lingual tasks, others were extended by translating part of the data through human professionals and automatic methods.

Name	Source	Domain	Tasks	Datasets	Languages	Agg. Score
GLUE	(Wang et al., 2019b)	Misc. Texts	7	9	English	Yes
SUPERGLUE	(Wang et al., 2019a)	Misc. Texts	8	8	English	Yes
CLUE	(Xu et al., 2020)	Misc. Texts	9	9	Chinese	Yes
XTREME	(Hu et al., 2020)	Misc. Texts	6	9	40	Yes
BLUE	(Peng et al., 2019)	Biomedical Texts	5	10	English	Yes
CBLUE	(Zhang et al., 2022)	Biomedical Texts	9	9	Chinese	Yes
MMLU	(Hendrycks et al., 2021)	Misc. Texts	1	57	English	Yes
LexGLUE	(Chalkidis et al., 2022b)	Legal Texts	7	6	English	Yes
LBOX	(Hwang et al., 2022)	Legal Texts	5	5	Korean	Yes
LEXTREME	(our work)	Legal Texts	18	11	25	Yes
SUPERB	(Yang et al., 2021)	Speech	10	10	English	No
SUPERB-SG	(Tsai et al., 2022)	Speech	5	5	English	No
TAPE	(Rao et al., 2019)	Proteins	5	5	n/a	No

Table 1: Characteristics of popular existing NLP benchmarks.

2.2 Benchmarks for Legal Language Models

LEXGLUE (Chalkidis et al., 2022b) is the first benchmark for the legal domain and covers 6 predictive tasks over 5 datasets made of textual documents in English from the US, EU, and CoE. While some tasks may not require specific legal knowledge to be solved, others would probably need, or at least benefit from, information regarding the EU or US legislation on the specific topic. Among the main limitations of their benchmark, Chalkidis et al. highlight its monolingual nature and remark that "*there is an increasing need for developing models for other languages*". Our work is strongly inspired by LEXGLUE and our purpose is to propose a benchmark that, we hope, will help the development of multilingual models for the legal domain.

In a similar direction, Hwang et al. (2022) released the LBOX benchmark. It covers 3 downstream tasks: two legal judgement prediction (LJP) tasks, and one summarization task in Korean.

The LEGALBENCH initiative (Guha et al., 2022) aims to create an open and collaborative legal reasoning benchmark where legal practitioners and other domain experts can contribute by submitting tasks that will be addressed using language models. At its creation, the authors have already added 44 lightweight tasks. While most tasks require legal reasoning based on the common law system, there is also a clause classification task.

Concerning language models specifically trained for the legal domain, many have been proposed for specific languages but, to the best of our knowledge, no multilingual model has been proposed yet. Legal language models have been proposed for English (Chalkidis et al., 2020; Ying and Habernal, 2022), French (Douka et al., 2021), Romanian (Masala et al., 2021), Italian (Tagarelli and Simeri, 2022; Licari and Comandè, 2022), Chinese (Xiao et al., 2021), Arabic (Al-Qurishi et al., 2022), Korean (Hwang et al., 2022), and Portuguese (Ciurlino, 2021). For an overview of the many tasks related to the automatic analysis of legal texts, we suggest reading the works of Chalkidis et al. (2022b) and Zhong et al. (2020).

3 LEXTREME Tasks and Datasets

3.1 LEXTREME Dataset Selection

To select the datasets for the LEXTREME benchmark, we formulate various criteria. We first systematically explore the literature via the ACL anthologyto find relevant datasets for the legal domain. We identify various venues, such as ACL, EACL, NAACL, EMNLP, LREC, ICAIL, and the NLLP workshop. We search the literature of these venues for the years 2010 to 2022. We search for some common keywords (case insensitive) that are related to legal datasets, e.g., *criminal*, *judicial*, *judgment*, *jurisdictions*, *law*, *legal*, *legislation*, *dataset*, and *corpus*. These keywords help to select potentially relevant papers, i.e., 108 papers. Then, three authors analyze these papers based on the inclusion and exclusion criteria given below to ensure that they indeed propose a legal dataset.

Inclusion criteria

I1: It is about legal text (e.g., patents are not considered part of legal text),

- I2: It performs legal tasks (e.g., judgment prediction) and not other linguistic tasks such as Part-of-Speech (POS) tagging,
- I3: It performs NLU tasks (e.g., information retrieval tasks are not considered due to their evaluation complexity),
- I4: The tasks are in one of the European languages (e.g., China has its own large legal Natural Language Processing (NLP) community and likely would not benefit much from multilingual models), and
- I5: The dataset is annotated by humans directly or indirectly (e.g., judgement labels are extracted with regexes)

Exclusion criteria

- E1: The dataset is not publicly available,
- E2: The dataset does not contain a public license,
- E3: The dataset contains labels that are generated with ML systems.
- E4: It is not a peer-reviewed paper

Task	**# Examples**	**# Labels**
BCD-J	3234 / 404 / 405	3 / 3 / 3
BCD-U	1715 / 211 / 204	2 / 2 / 2
GAM	19271 / 2726 / 3078	4 / 4 / 4
GLC-V	28536 / 9511 / 9516	47 / 47 / 47
GLC-C	28536 / 9511 / 9516	386 / 377 / 374
GLC-S	28536 / 9511 / 9516	2143 / 1679 / 1685
SJP	59709 / 8208 / 17357	2 / 2 / 2
OTS-UL	2074 / 191 / 417	3 / 3 / 3
OTS-CT	19942 / 1690 / 4297	9 / 8 / 9
C19	3312 / 418 / 418	8 / 8 / 8
MEU-1	817239 / 112500 / 115000	21 / 21 / 21
MEU-2	817239 / 112500 / 115000	127 / 126 / 127
MEU-3	817239 / 112500 / 115000	500 / 454 / 465
GLN	17699 / 4909 / 4017	17 / 17 / 17
LNR	7552 / 966 / 907	11 / 9 / 11
LNB	7828 / 1177 / 1390	13 / 13 / 13
MAP-C	27823 / 3354 / 10590	13 / 11 / 11
MAP-F	27823 / 3354 / 10590	44 / 26 / 34

Table 2: Overview of datasets and their tasks. The fields *# Examples* and *# Labels* provide the values for the splits train, validation, test. For a detailed overview of for the language-specific subsets of each multilingual task, see Table 7 and 8.

After applying the above criteria, we reduce from 108 to 11 datasets. We provide the list of all these datasets in the online repository.[3]

[3] https://github.com/JoelNiklaus/LEXTREME

Dataset	**Jurisdiction**	**Languages**
BCD	BR	pt
GAM	DE	de
GLC	GR	el
SJP	CH	de, fr, it
OTS	EU	de, en, it, pl
C19	BE, FR, HU, IT, NL, PL, UK	en, fr, hu, it, nb, nl, pl
MEU	EU	24 EU langs
GLN	GR	el
LNR	RO	ro
LNB	BR	pt
MAP	EU	24 EU langs

Table 3: Overview of datasets and the jurisdiction as well as the languages that they cover. The 24 EU languages are: bg, cs, da, de, el, en, es, et, fi, fr, ga, hu, it, lt, lv, mt, nl, pt, ro, sk, sv

3.2 LEXTREME Tasks

LEXTREME constist of three classification task types: Single Label Text Classification (SLTC), Multi Label Text Classification (MLTC), and Named Entity Recognition (NER). We use the existing train, validation, and test splits if present. In the other cases we split the data ourselves (80% train, 10% validation and test each). In the following, we briefly describe the selected datasets. For more information about the number of examples and label classes per split for each task, see Table 2, 7 and 8. For a detailed overview of the jurisdictions as well as the number of languages covered by each dataset, see Table 3.

3.3 LEXTREME Datasets

Each dataset can be either monolingual or multilingual and can have several configurations or (fine-tuning) tasks, which are the basis of our analyses, i.e., the pretrained models have always been fine-tuned on a single task.

Brazilian Court Decisions (BCD) Legal systems are often huge and complex, and the information is scattered across various sources. Thus, predicting case outcomes from multiple vast volumes of litigation is a difficult task. Lage-Freitas et al. (2022) propose an approach to predict Brazilian legal decisions to support legal practitioners. We use their dataset from the State Supreme Court of Alagoas (Brazil). The input to the models is always the case description. We perform two SLTC tasks: One (BCD-J) is to predict the approval or dismissal of the case or appeal with three labels *no, partial, yes*, and another (BCD-U) is to predict the unanimity on the decision alongside two labels *unanimity, not-unanimity*.

German Argument Mining (GAM) Identifying arguments in court decisions is an important and challenging task for legal practitioners. Urchs. et al. (2021) compiled a dataset of 200 German court decisions for classifying sentences according to their argumentative function. We use their dataset to perform an MLTC task. The input to the models is a sentence and the output is labeled according to four categories: *conclusion, definition, subsumption, other*.

Greek Legal Code (GLC) Legal documents can cover a wide variety of topics, which makes accurate topic classification all the more important. Papaloukas et al. (2021) compiled a dataset for topic classification of Greek legislation documents. The documents cover 47 main thematic topics which are called *volumes*. Each of them is divided into thematic sub categories which are called *chapters* and subsequently, each chapter breaks down to *subjects*. Therefore, the dataset is used to perform three different SLTC tasks along volume level (GLC-V), chapter level (GLC-C), and subject level (GLC-S). The input to the models is the entire document, and the output is one of the several topic categories.

Swiss Judgment Prediction (SJP) Niklaus et al. (2021, 2022b), focus on predicting the judgment outcome of the cases from the Swiss Federal Supreme Court (FSCS). We use their dataset of 85k cases. The input to the models is the appeal description, and the output is whether the appeal is approved or dismissed. It is also a SLTC task.

Online Terms of Service (OTS) While the benefits of multilingualism in the EU legal world are well known, creating an official version of every legal act in 24 languages raises interpretative challenges. Drawzeski et al. (2021), attempt to automatically detect unfair clauses in Terms of Service. We use their dataset of 100 contracts to perform a SLTC and MLTC task. In the SLTC task (OTS-UL), the input to the models is a sentence, and the output presents the sentence classified into three levels of unfairness. In the MLTC task (OTS-CT), the model identifies the sentence for various clause topics.

COVID19 Emergency Event (C19) The COVID-19 pandemic showed various exceptional measures governments around the world have taken to contain the virus. Tziafas et al. (2021), presented a dataset, also known as EXCEPTIUS, that contains legal documents with sentence-level annotation from several European countries to automatically identify the measures. We use their dataset to perform only one task, i.e., the MLTC task of identifying the type of measure described in a sentence. The input to the models are the sentences, and the output is neither or at least one of the measurement types.

MultiEURLEX (MEU) Multilingual transfer learning has gained significant attention recently due to its increasing applications in NLP tasks. Chalkidis et al. (2021b), explored the cross-lingual transfer for legal NLP and presented a corpus of 65K EU laws. They annotated each law document with multiple labels from the EUROVOC taxonomy. We perform a MLTC task to identify labels (given in the taxonomy) for each document. Since the taxonomy exists on multiple levels, we prepare configurations according to three levels (MEU-1, MEU-2, MEU-2).

Greek Legal NER (GLN) Identifying various named entities from natural language text plays an important role for Natural Language Understanding (NLU). Papaloukas et al. (2021) compiled an annotated dataset for NER in Greek legal documents. The source material are 254 daily issues of the Greek Government Gazette over the period 2000-2017. In *all* NER tasks of LEXTREME the input to the models is the list of tokens, and the output is an entity label for each token.

LegalNERo (LNR) Similar to GLN, Pais et al. (2021) manually annotated Romanian legal documents for various named entities. The dataset is derived from 370 documents from the larger MARCELL Romanian legislative subcorpus[4].

LeNER BR (LNB) Luz de Araujo et al. (2018) compiled a dataset for NER for Brazilian legal documents. To compose the dataset, 66 legal documents from several Brazilian Courts were collected. Additionally, four legislation documents were collected, resulting a total of 70 documents that were annotated for named entities.

[4] https://marcell-project.eu/deliverables.html

Model	Source	Params	Vocab	Specs	Corpora	# Langs
MiniLM	Wang et al. (2020)	118M	250K	1M steps / BS 256	2.5T CC100 data	100
DistilBert	Sanh et al. (2019)	135M	120K	BS up to 4000	Wikipedia	104
mDeberta-v3	He et al. (2020, 2021)	278M	128K	500K steps / BS 8192	2.5T CC100 data	100
XLM-R base	Conneau et al. (2020)	278M	250K	1.5M steps / BS 8192	2.5T CC100 data	100
XLM-R large	Conneau et al. (2020)	560M	250K	1.5M steps / BS 8192	2.5T CC100 data	100

Table 4: Multilingual Models: All models can process up to 512 tokens. BS is short for batch size. Params is the total number of parameters (including the embedding layer).

MAPA (MAP) de Gibert et al. (2022), built a multilingual corpus based on EUR-Lex (Baisa et al., 2016) for NER. The dataset comes in two configurations, i.e., two NER tasks, as it has been annotated at a coarse-grained (MAP-C) and fine-grained (MAP-F) level. The structure of the dataset is the same as the other datasets for NER.

4 Models Considered

Since our benchmark only contains NLU tasks, we consider encoder only models for simplicity.

MiniLM MiniLM (Wang et al., 2020) is the result of a novel task-agnostic compression technique, also called distillation, in which a compact model — the so-called student — is trained to reproduce the behaviour of a larger pre-trained model — the so-called teacher. This is achieved by deep self-attention distillation, i.e. only the self-attention module of the last Transformer layer of the teacher, which stores a lot of contextual information (Jawahar et al., 2019), is distilled. The student is trained by closely imitating the teacher's final Transformer layer's self-attention behavior. To aid the learner in developing a better imitation, (Wang et al., 2020) also introduce the self-attention value-relation transfer in addition to the self-attention distributions. The addition of a teacher assistant results in further improvements. For the training of multilingual MiniLM, XLM-R_{BASE} was used.

DistilBERT DistilBERT (Sanh et al., 2019) is a more compressed version of BERT (Devlin et al., 2019) using teacher-student learning, similar to MiniLM. DistilBERT is distilled from BERT, thus both share a similar overall architecture. The pooler and token-type embeddings are eliminated, and the number of layers is decreased by a factor of 2 in DistilBERT. DistilBERT is distilled in very large batches while utilizing gradient accumulation and dynamic masking, but without the next sentence prediction objective. DistilBERT was trained on the same corpus as the original BERT.

mDEBERTa He et al. (2020) suggest a new model architecture called DeBERTa (Decoding-enhanced BERT with disentangled attention), which employs two novel methods to improve the BERT and RoBERTa models. The first is the disentangled attention mechanism, in which each word is represented by two vectors that encode its content and position, respectively, and the attention weights between words are calculated using disentangled matrices on their respective contents and relative positions. To predict the masked tokens during pre-training, an enhanced mask decoder is utilized, which incorporates absolute positions in the decoding layer. Additionally, the generalization of models is enhanced through fine-tuning using a new virtual adversarial training technique. He et al. (2021) introduce mDEBERTa-v3 by further improving the efficiency of pre-training by replacing Masked-Language Modeling (MLM) in DeBERTa with the task of replaced token detection (RTD) where the model is trained to predict whether a token in the corrupted input is either original or replaced by agenerator. Further improvements are achieved via *gradient-disentangled embedding sharing* (GDES).

XLM-RoBERTa XLM-R (Conneau et al., 2020) is a multilingual language model which has the same pretraining objectives as RoBERTa (Liu et al., 2019), such as dynamic masking, but not next sentence prediction. It is pre-trained on a large corpus comprising 100 languages. The authors report a significant performance gain over multilingual BERT (mBERT) in a variety of tasks with results competitive with state-of-the-art monolingual models (Conneau et al., 2020).

4.1 Hierarchical Variants

A significant part of the datasets consists of very long documents, the best examples being all variants of MultiEURLEX, cf. Figure 12. However, Transformer-based models usually allow a maximum input length of 512 tokens. It is possible to use the models without further ado for documents that exceed this length by far. However, this can only be achieved by a massive truncation of the original document. This procedure has the consequence that only the first section of a document is available for classification tasks. This is the reason why we used hierarchical variants of pretraining models for finetuning on data sets with particularly long documents (cf. histograms).

The hierarchical variants used in the study are broadly equivalent to those in (Chalkidis et al., 2021c; Niklaus et al., 2022a). First, we convert each document into a list of equal-length paragraphs. Afterward, we use a pre-trained Transformer-based model to encode each of these paragraphs separately and to obtain the [CLS] embedding of each paragraph which can be used as a context-unaware paragraph representation. In order to make them context-aware, i.e. aware of the surrounding paragraphs, the paragraph representations are fed into a 2-layered Transformer encoder with varying specifications depending on the model type. Finally, max-pooling over the context-aware paragraph representations is deployed, which results in a document representation that is fed to a classification layer.

5 Experimental Setup

Some datasets were highly imbalanced, one of the best examples being BCD-U with a proportion of the minority class of about 2%. Therefore, we applied random oversampling on all tasks of the SLTC datasets, except for GLC, since all its subsets have too many labels, which would have led to a drastic increase in the data size and thus in the computational costs for finetuning. For each run, we used the same hyperparameters, as described in Section A.2.

As described in section 4.1, some tasks contain very long documents, which required the usage of hierarchical variants with sequence lengths that go beyond 512. Based on the distribution of the sequence length per example for each task (cf. section D), we decided on suitable sequence lengths for each task before finetuning. A list of suitable sequence lengths can be found in A.1. Tasks with a maximum sequence length of over 512 required the usage of hierarchical variants.

Evaluation Metrics We use the macro-F1 score for all datasets to ensure comparability across the entire benchmark, since it can be computed for both text classification and NER tasks. Mathew's Correlation Coefficient (MCC) is a suitable score for evaluating text classification tasks but its applicability to NER tasks is unclear. For brevity, we do not display additional scores, but more detailed (such as precision and recall, and scores per seed) and additional scores (such as MCC) can be found online on our Weights and Biases project[5].

Aggregate Score We acknowledge that the datasets included in LEXTREME are diverse and hard to compare due to variations in the number of samples and task complexity (Raji et al., 2021a). This is why we always report the scores for each dataset subset, enabling a fine-grained analysis. However, we believe that by taking the following three measures, an aggregate score can provide more benefits than drawbacks, encouraging the community to evaluate multilingual legal models on a curated benchmark facilitating comparisons.

We (a) evaluate all datasets with the same score (macro-F1) making aggregation more intuitive and easier to interpret, (b) aggregating the F1 scores again using the harmonic mean, since F1 scores are already rates and obtained using the harmonic mean over precision and recall, following Tatiana and Valentin (2021), and (c) basing our final aggregate score on two intermediate aggregate scores — the dataset aggregate and language aggregate score – thus weighing datasets and languages equally promoting model fairness and robustness.

The final LEXTREME score is computed using the harmonic mean of the dataset and the language aggregate score. We compute the dataset aggregate score by taking the successive harmonic mean of (1.) the languages inside the configurations (e.g., de,fr,it within SJP), (2.) the configurations inside the datasets (e.g., OTS-UL, OTS-CT within OTS), and (3.) the datasets inside LEXTREME (BCD, GAM, etc.). We compute the language aggregate score by taking the successive harmonic mean of (1.) the configurations inside the datasets, (2.) the datasets for the given language (e.g., MAP and MEU for lv), and (3.) the languages inside LEXTREME (bg,cs, etc.).

[5] https://wandb.ai/lextreme/paper_results

Model	BCD	GAM	GLC	SJP	OTS	C19	MEU	GLN	LNR	LNB	MAP	Agg.
MiniLM	53.0	**73.3**	42.1	67.7	44.1	2.6	62.0	40.5	46.8	86.0	55.5	52.2
DistilBERT	54.5	69.5	**62.8**	66.8	56.1	22.2	63.6	38.1	48.4	78.7	55.0	56.0
mDeBERTa v3	57.6	70.9	52.2	69.1	66.5	25.5	65.1	42.2	46.6	87.8	**60.2**	58.5
XLM-R base	**63.5**	72.0	56.8	**69.3**	67.8	26.4	65.6	47.0	47.7	86.0	56.1	59.9
XLM-R large	58.7	73.1	57.4	69.0	**75.0**	**29.0**	**68.1**	**48.0**	**49.5**	**88.2**	58.5	**61.3**

Table 5: Dataset aggregate scores for multilingual models. The best scores are in bold.

Model	bg	cs	da	de	el	en	es	et	fi	fr	ga	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv	Agg.
MiniLM	64.0	57.7	55.4	60.1	48.9	42.8	63.8	59.7	56.6	48.5	41.5	62.2	41.8	45.6	59.8	60.2	55.7	38.8	33.5	63.5	58.4	58.9	62.2	59.4	54.1
DistilBERT	65.3	60.2	57.4	64.1	53.1	54.0	66.9	57.4	55.7	55.8	45.5	63.1	39.9	54.9	58.0	57.7	57.3	42.0	43.6	64.7	57.4	59.0	63.3	59.2	56.5
mDeBERTa v3	61.9	60.6	59.3	66.6	54.0	58.9	66.9	60.3	**61.1**	57.0	50.2	65.0	44.2	59.7	63.7	61.4	**61.2**	**48.1**	50.2	67.9	**60.8**	**65.2**	65.2	**65.4**	59.8
XLM-R base	**68.3**	61.3	58.5	66.0	54.7	58.6	63.8	59.3	57.5	57.7	47.8	65.9	43.3	59.6	60.3	60.8	58.0	45.0	52.0	**68.2**	59.2	60.3	66.2	61.7	58.9
XLM-R large	64.5	**63.3**	**65.1**	**68.3**	**59.6**	**61.9**	**70.0**	**61.3**	60.9	**57.9**	**50.3**	**68.3**	**44.7**	**62.9**	**66.1**	**65.5**	60.1	43.9	**55.0**	68.1	60.2	62.8	**68.2**	62.5	**61.3**

Table 6: Language aggregate scores for multilingual models. The best scores are in bold.

6 Results

In this section, we discuss the main result of our evaluation of the baseline models. Scores on the validation datasets and standard deviations across seeds can be found in Appendix C.

We show the dataset and language aggregated results in Tables 5 and 6 respectively. For both the dataset aggregate and the language aggregate scores, we see a clear trend that larger models perform better. However, when looking at the individual datasets and languages, the scores are more erratic. We notice that on some datasets, such as C19, GLC or OTS, the models vary greatly, with differences as large as 29.2 between the worst performing MiniLM and the best performing XLM-R large. MiniLM seems to struggle greatly with these three datasets, while even achieving the best performance on GAM. On other datasets, such as SJP, MEU, LNR, and MAP the models are very close together (6 points or fewer between best and worst model). SJP, MEU and MAP are the largest datasets in LEXTREME, thus probably decreasing the influence of the pretraining on downstream performance and leveling the playing field. LNR, however, is the smallest NER task, opposing this hypothesis. In contrast to the inconsistent results on the datasets, we notice, that XLM-R performs best on most languages. Additionally, we note that the variability of the models within a language is similar to the variability within a dataset, however, we don't see extreme cases such as GLC or OTS.

7 Conclusions and Future Work

Conclusions

We survey the literature and select 11 datasets out of 108 papers with rigorous criteria to compile the first multilingual benchmark for legal NLP. By open-sourcing both the dataset and the code, we invite researchers and practitioners to evaluate any future multilingual models on our benchmark. We provide baselines for five popular multilingual encoder-based language models of different sizes. We hope that this benchmark will foster the creation of novel legal multilanguage models and therefore contribute to the progress of natural legal language processing. We imagine this work as a living benchmark and invite the community to extend it with new suitable datasets.

Future Work

In future work, we will extend this benchmark with other NLU tasks and also generation tasks such as summarization, simplification, or translation. Another avenue of future work can be the extension with datasets in more languages or from jurisdictions not yet covered in the current version. Finally, we leave the evaluation of other models such as mT5 (Xue et al., 2021) to future work.

Limitations

It is important to not exceed with the enthusiasm for language models and the ambitions of benchmarks: many recent works have addressed the limits of these tools and analyzed the consequences of their misuses. For example, Bender and Koller (2020) argue that language models do not really learn "meaning". Koch et al. (2021) evaluate the use of datasets inside scientific communities and highlight that many machine learning communities focus on very few datasets and that often these dataset are "borrowed" from other communities. Raji et al. (2021b) offer a detailed exploration of the limits of popular "general" benchmarks, such as

GLUE (Wang et al., 2019b) and ImageNET (Deng et al., 2009). Their analysis covers 3 aspects: limited task design, de-contextualized data and performance reporting, inappropriate community use.

The first problem concerns the fact that typically tasks are not chosen considering proper theories and selecting what would be needed to prove generality. Instead, they are limited to what is considered interesting by the community, what is available, or other similar criteria. These considerations hold also for our work. Therefore, we can not claim that our benchmark can be used to assess the “generality” of a model or proving that it “understands natural legal language”.

The second point address the fact that any task, data, or metric are limited to their context, therefore “data benchmarks are closed and inherently subjective, localized constructions”. In particular, the content of the data can be too different from real data and the format of the tasks can be too homogeneous compared to human activities. Moreover, any dataset inherently contains biases. We tackle this limitation by deciding to include only tasks and data that are based on real world scenarios, in an effort to minimize the difference between the performance of a model on our benchmark and its performance on a real world problem.

The last aspect regards the negative consequences that benchmarks can have. The competitive testing may encourage misbehavior and the aggregated performance evaluation does create a mirage of cross-domain comparability. The presence of popular benchmarks can influence a scientific community up to the point of steering towards techniques that perform well on that specific benchmark, in disfavor of those that do not. Finally, benchmarks can be misused in marketing to promote commercial products while hiding their flaws. Since these behaviour obviously can not be forecasted in advance, but we hope that this analysis of the shortcomings of our work will be sufficient to prevent misuses of our benchmark and will also inspire research directions for complementary future works. For what specifically concerns aggregated evaluations, they provide an intuitive but imprecise understanding of the performance of a model. While we do not deny their potential downsides, we believe that their responsible use is beneficial, especially when compared to the evaluation of a model on only an arbitrarily selected set of datasets. Therefore, we have decided to provide an aggregated performance evaluation and to weight languages and tasks equally to make it as robust and fair as possible.

It is important to remark that while Raji et al. and Koch et al. argument against the misrepresentations and the misuses of benchmarks and datasets, they do not argue against their usefulness. On the contrary, they consider the creation and adoption of novel benchmarks a sign of a healthy scientific community.

Ethics Statement

The scope of this work is to release a unified multilingual legal NLP benchmark to accelerate the development and evaluation of multilingual legal language models. A transparent multilingual and multinational benchmark for NLP in the legal domain might serve as an orientation for scholars and industry researchers by broadening the discussion and helping practitioners to build assisting technology for legal professionals and laypersons. We believe that this is an important application field, where research should be conducted (Tsarapatsanis and Aletras, 2021) to improve legal services and democratize law, while also highlight (inform the audience on) the various multi-aspect shortcomings seeking a responsible and ethical (fair) deployment of legal-oriented technologies.

Nonetheless, irresponsible use (deployment) of such technology is a plausible risk, as in any other application (e.g., online content moderation) and domain (e.g., medical). We believe that similar technologies should only be deployed to assist human experts (e.g., legal scholars in research, or legal professionals in forecasting or assessing legal case complexity) with notices on their limitations.

All datasets included in LEXTREME, are publicly available and have been previously published. We referenced the original work and encourage LEXTREME users to do so as well. In fact, we believe this work should only be referenced, in addition to citing the original work, when experimenting with multiple LEXTREME datasets and using the LEXTREME evaluation infrastructure. Otherwise, only the original work should be cited.

References

Muhammad Al-Qurishi, Sarah AlQaseemi, and Riad Souissi. 2022. Aralegal-bert: A pretrained language model for arabic legal text. In *NLLP*.

Nikolaos Aletras, Leslie Barrett, Catalina Chalkidis Ilias Goanta, and Daniel Preotiuc-Pietro, editors. 2022. *Proceedings of the Natural Legal Language Processing Workshop 2022*. Association for Computational Linguistics, Abu Dhabi, UAE.

Vít Baisa, Jan Michelfeit, Marek Medveď, and Miloš Jakubíček. 2016. European Union language resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2799–2803, Portorož, Slovenia. European Language Resources Association (ELRA).

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: on meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5185–5198. Association for Computational Linguistics.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the Opportunities and Risks of Foundation Models. ArXiv:2108.07258 [cs].

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021a. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021b. Multieurlex–a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2109.00904*. Dataset URL: https://huggingface.co/datasets/multi_eurlex.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021c. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, Online. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022a. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael J. Bommarito II, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2022b. Lexglue: A benchmark dataset for legal language understanding in english. In *ACL (1)*, pages 4310–4330. Association for Computational Linguistics.

Victor Hugo Ciurlino. 2021. Bertbr: a pretrained language model for law texts. Master's thesis, Universidade de Brasília.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ona de Gibert, A García-Pablos, Montse Cuadros, and Maite Melero. 2022. Spanish datasets for sensitive entity detection in the legal domain. In

Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22), Marseille, France, june. European Language Resource Association (ELRA). Dataset URL: https://tinyurl.com/mv65cp66.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. 2021. JuriBERT: A masked-language model adaptation for French legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 95–101, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kasper Drawzeski, Andrea Galassi, Agnieszka Jablonowska, Francesca Lagioia, Marco Lippi, Hans Wolfgang Micklitz, Giovanni Sartor, Giacomo Tagiuri, and Paolo Torroni. 2021. A corpus for multilingual analysis of online terms of service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 1–8. Dataset URL: https://claudette.eui.eu/corpus_multilingual_NLLP2021.zip.

Neel Guha, Daniel E. Ho, Julian Nyarko, and Christopher Ré. 2022. Legalbench: Prototyping a collaborative benchmark for legal reasoning. *CoRR*, abs/2209.06120.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. pages 1–17.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. ArXiv:2207.00220 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. 2022. A multi-task benchmark for korean legal language understanding and judgement prediction. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. 2023. Natural Language Processing in the Legal Domain.

Bernard Koch, Emily Denton, Alex Hanna, and Jacob G. Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. In *NeurIPS Datasets and Benchmarks*.

André Lage-Freitas, Héctor Allende-Cid, Orivaldo Santana, and Lívia Oliveira-Lage. 2022. Predicting brazilian court decisions. *PeerJ Computer Science*, 8:e904. Dataset URL: https://github.com/proflage/predicting-brazilian-court-decisions.

Daniele Licari and Giovanni Comandè. 2022. Italian-legal-bert: A pre-trained transformer language model for italian law. In *EKAW-C*, volume 3256. CEUR-WS.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (1).

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pages 313–323. Springer. Dataset URL: https://huggingface.co/datasets/lener_br.

Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. 2021. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-judgment-prediction: a multilingual legal judgment prediction benchmark. *arXiv preprint arXiv:2110.00806*. Dataset URL: https://huggingface.co/datasets/swiss_judgment_prediction.

Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? ArXiv:2211.17135 [cs].

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022a. An empirical study on cross-X transfer for legal judgment prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022b. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christos Papaloukas, Ilias Chalkidis, Konstantinos Athinaios, Despina-Athanasia Pantazi, and Manolis Koubarakis. 2021. Multi-granular legal topic classification on greek legislation. *arXiv preprint arXiv:2109.15298*. Dataset URL: https://huggingface.co/datasets/greek_legal_code.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In *BioNLP@ACL*, pages 58–65. Association for Computational Linguistics.

Inioluwa Deborah Raji, Emily M. Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021a. AI and the Everything in the Whole Wide World Benchmark. ArXiv:2111.15366 [cs].

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021b. AI and the everything in the whole wide world benchmark. In *NeurIPS Datasets and Benchmarks*.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John F. Canny, Pieter Abbeel, and Yun S. Song. 2019. Evaluating protein transfer learning with TAPE. In *NeurIPS*, pages 9686–9698.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. 2022. Multilexsum: Real-world summaries of civil rights lawsuits at multiple granularities. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Andrea Tagarelli and Andrea Simeri. 2022. Lamberta: Law article mining based on bert architecture for the italian civil code. In *ICRDL*.

Shavrina Tatiana and Malykh Valentin. 2021. How not to Lie with a Benchmark: Rearranging NLP Leaderboards. ArXiv:2112.01342 [cs].

Rena Torres Cacoullos. 2020. Code-switching strategies: Prosody and syntax. *Frontiers in Psychology*, 11.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-Wen Yang, Shuyan Dong, Andy T. Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2022. SUPERB-SG: enhanced speech processing universal performance benchmark for semantic and generative capabilities. In *ACL (1)*, pages 8479–8492. Association for Computational Linguistics.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Georgios Tziafas, Eugenie de Saint-Phalle, Wietse de Vries, Clara Egger, and Tommaso Caselli. 2021. A multilingual approach to identify and classify exceptional measures against covid-19. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 46–62. Dataset URL: https://tinyurl.com/ycysvtbm.

Stefanie Urchs., Jelena Mitrović., and Michael Granitzer. 2021. Design and implementation of german legal decision corpora. pages 515–521. SciTePress.

Serena Villata, Michał Araszkiewicz, Kevin D. Ashley, Trevor J. M. Bench-Capon, L. Karl Branting, Jack G. Conrad, and Adam Zachary Wyner. 2022. Thirty years of artificial intelligence and law: the third decade. *Artificial Intelligence and Law*, 30:561–591.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. In *NeurIPS*, pages 3261–3275.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR (Poster)*. OpenReview.net.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for chinese legal long documents. *AI Open*, 2:79–84.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *COLING*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*. ArXiv: 2010.11934.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko-tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. 2021. SUPERB: speech processing universal performance benchmark. In *Interspeech*, pages 1194–1198. ISCA.

Yin Ying and Ivan Habernal. 2022. Privacy-Preserving Models for Legal Natural Language Processing. In *NLLP*, page (to appear), Abu Dhabi, UAE.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. 2022. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *ACL*, pages 7888–7915, Dublin, Ireland. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, page 159–168, New York, NY, USA. Association for Computing Machinery.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. How does NLP benefit legal system: A summary of legal artificial intelligence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5218–5230, Online. Association for Computational Linguistics.

A Experiment Details

A.1 Maximum Sequence Lengths

Brazilian Court Decisions: 1024 (128 x 8) CoVID19: 256 German Argument Mining: 256 Greek Legal Code: 4096 (if speed is important: 2048) (128 x 32 / 16) Greek Legal NER: 512 (max for non-hierarchical) LegalNERo: 512 (max for non-hierarchical) LeNER: 512 (max for non-hierarchical) MAPA: 512 (max for non-hierarchical) MultiEURLEX: 4096 (or for maximum performance 8192) (128 x 32 / 64) Online Terms of Service: 256 Swiss Judgment Prediction: 2048 (or for maximum performance on fr: 4096) (128 x 16 / 32)

A.2 Hyperparameters

We used learning rate 1e-5 for all models and datasets without tuning. We ran all experiments with 3 random seeds (1-3) We always used batch size 64. In case the GPU memory was insufficient, we additionally used gradient accumulation. We trained using early stopping on the validation loss with patience of 5 epochs. Because MultiEURLEX is very large and the experiment very long, we just train for 1 epoch and evaluated after every 1000th step. We used AMP mixed precision training and evaluation to reduce costs. Mixed precision was not used in combination with microsoft/mdeberta-v3-base because it led to errors. The experiments were run the following NVIDIA GPUs: 24GB RTX3090, 32GB V100 and 80GB A100.

B Dataset Splits

Language	SJP	OTS-UL	OTS-CT	C19	MEU-1	MEU-2	MEU-3	MAP-C	MAP-F
bg					15986 / 5000 / 5000	15986 / 5000 / 5000	15986 / 5000 / 5000	1411 / 166 / 560	1411 / 166 / 560
cs					23187 / 5000 / 5000	23187 / 5000 / 5000	23187 / 5000 / 5000	1464 / 176 / 563	1464 / 176 / 563
da					55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1455 / 164 / 550	1455 / 164 / 550
de	35458 / 4705 / 9725	491 / 42 / 103	4480 / 404 / 1027		55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1457 / 166 / 558	1457 / 166 / 558
el					55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1529 / 174 / 584	1529 / 174 / 584
en		526 / 49 / 103	5378 / 415 / 1038	648 / 81 / 81	55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	893 / 98 / 408	893 / 98 / 408
es					52785 / 5000 / 5000	52785 / 5000 / 5000	52785 / 5000 / 5000	806 / 248 / 155	806 / 248 / 155
et					23126 / 5000 / 5000	23126 / 5000 / 5000	23126 / 5000 / 5000	1391 / 163 / 516	1391 / 163 / 516
fi					42497 / 5000 / 5000	42497 / 5000 / 5000	42497 / 5000 / 5000	1398 / 187 / 531	1398 / 187 / 531
fr	21179 / 3095 / 6820			1416 / 178 / 178	55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1297 / 97 / 490	1297 / 97 / 490
ga								1383 / 165 / 515	1383 / 165 / 515
hr					7944 / 2500 / 5000	7944 / 2500 / 5000	7944 / 2500 / 5000		
hu				75 / 10 / 10	22664 / 5000 / 5000	22664 / 5000 / 5000	22664 / 5000 / 5000	1390 / 171 / 525	1390 / 171 / 525
it	3072 / 408 / 812	517 / 50 / 102	4806 / 432 / 1057	742 / 93 / 93	55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1411 / 162 / 550	1411 / 162 / 550
lt					23188 / 5000 / 5000	23188 / 5000 / 5000	23188 / 5000 / 5000	1413 / 173 / 548	1413 / 173 / 548
lv					23208 / 5000 / 5000	23208 / 5000 / 5000	23208 / 5000 / 5000	1383 / 167 / 553	1383 / 167 / 553
mt					17521 / 5000 / 5000	17521 / 5000 / 5000	17521 / 5000 / 5000	937 / 93 / 442	937 / 93 / 442
nb				221 / 28 / 28					
nl				135 / 18 / 18	55000 / 5000 / 5000	55000 / 5000 / 5000	55000 / 5000 / 5000	1391 / 164 / 530	1391 / 164 / 530
pl		540 / 50 / 109	5278 / 439 / 1175	75 / 10 / 10	23197 / 5000 / 5000	23197 / 5000 / 5000	23197 / 5000 / 5000		
pt					52370 / 5000 / 5000	52370 / 5000 / 5000	52370 / 5000 / 5000	1086 / 105 / 390	1086 / 105 / 390
ro					15921 / 5000 / 5000	15921 / 5000 / 5000	15921 / 5000 / 5000	1480 / 175 / 557	1480 / 175 / 557
sk					22971 / 5000 / 5000	22971 / 5000 / 5000	22971 / 5000 / 5000	1395 / 165 / 526	1395 / 165 / 526
sl					23184 / 5000 / 5000	23184 / 5000 / 5000	23184 / 5000 / 5000		
sv					42490 / 5000 / 5000	42490 / 5000 / 5000	42490 / 5000 / 5000	1453 / 175 / 539	1453 / 175 / 539

Table 7: Overview of the number of examples for each language-specific subset of multilingual tasks. The order of the values is train / validation / test.

Language	SJP	OTS-UL	OTS-CT	C19	MEU-1	MEU-2	MEU-3	MAP-C	MAP-F
bg					21 / 21 / 21	127 / 126 / 127	481 / 454 / 465	11 / 11 / 8	24 / 16 / 13
cs					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 11 / 9	30 / 17 / 16
da					21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 10 / 11	26 / 14 / 14
de	2 / 2 / 2	3 / 3 / 3	9 / 7 / 9		21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 9 / 10	28 / 14 / 14
el					21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 11 / 11	31 / 17 / 20
en		3 / 3 / 3	9 / 8 / 9	6 / 6 / 5	21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 9 / 9	28 / 17 / 18
es					21 / 21 / 21	127 / 126 / 127	497 / 454 / 465	11 / 8 / 11	26 / 13 / 18
et					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 11 / 11	25 / 14 / 17
fi					21 / 21 / 21	127 / 126 / 127	493 / 454 / 465	11 / 11 / 10	24 / 19 / 16
fr	2 / 2 / 2			8 / 8 / 7	21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 11 / 11	32 / 19 / 26
ga								13 / 11 / 11	33 / 17 / 18
hr					21 / 21 / 21	127 / 126 / 127	469 / 437 / 465		
hu				4 / 1 / 1	21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 10 / 10	20 / 15 / 14
it	2 / 2 / 2	3 / 3 / 3	9 / 8 / 9	7 / 7 / 6	21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	11 / 10 / 11	25 / 15 / 16
lt					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 11 / 10	28 / 19 / 21
lv					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465	11 / 11 / 11	31 / 15 / 21
mt					21 / 21 / 21	127 / 126 / 127	485 / 454 / 465	11 / 11 / 11	27 / 15 / 15
nb				7 / 5 / 6					
nl				2 / 2 / 2	21 / 21 / 21	127 / 126 / 127	500 / 454 / 465	10 / 9 / 10	25 / 12 / 14
pl		3 / 3 / 3	9 / 8 / 9	7 / 5 / 3	21 / 21 / 21	127 / 126 / 127	486 / 454 / 465		
pt					21 / 21 / 21	127 / 126 / 127	497 / 454 / 465	11 / 10 / 11	29 / 14 / 18
ro					21 / 21 / 21	127 / 126 / 127	481 / 454 / 465	11 / 11 / 11	25 / 16 / 18
sk					21 / 21 / 21	127 / 126 / 127	485 / 454 / 465	11 / 11 / 11	25 / 16 / 18
sl					21 / 21 / 21	127 / 126 / 127	486 / 454 / 465		
sv					21 / 21 / 21	127 / 126 / 127	493 / 454 / 465	11 / 11 / 10	23 / 15 / 15

Table 8: Overview of the number of labels for each language-specific subset of multilingual tasks. The order of the values is train / validation / test.

C Detailed Multilingual Results

Model	**Mean**	**BCD-J**	**BCD-U**	**GAM**	**GLC-V**	**GLC-C**	**GLC-S**	**SJP**	**OTS-UL**	**OTS-CT**	**C19**	**MEU-1**	**GLN**	**LNR**	**LNB**	**MAP-C**	**MAP-F**
MiniLM	55.3	52.8 (±6.7)	55.1 (±6.6)	72.1 (±0.9)	82.0 (±1.0)	39.4 (±1.0)	**5.1 (±1.6)**	68.9 (±0.7)	71.0 (±5.0)	15.3 (±3.4)	5.9 (±1.5)	64.8 (±0.3)	41.5 (±3.1)	**63.5 (±5.3)**	86.0 (±0.4)	80.1 (±0.2)	62.8 (±2.3)
DistilBERT	61.7	52.1 (±4.5)	60.0 (±9.8)	70.6 (±1.7)	84.9 (±0.5)	68.0 (±0.6)	33.9 (±2.0)	68.7 (±0.7)	66.9 (±3.4)	49.6 (±9.1)	41.4 (±5.6)	68.2 (±0.1)	38.9 (±2.9)	**63.5 (±3.7)**	70.3 (±1.6)	78.7 (±0.2)	58.8 (±1.8)
mDeBERTa v3	64.7	**68.2 (±3.9)**	69.9 (±5.6)	69.5 (±2.0)	85.0 (±0.8)	58.2 (±7.5)	12.3 (±2.5)	**71.2 (±0.7)**	**85.2 (±2.9)**	52.1 (±4.6)	43.4 (±4.3)	68.4 (±0.6)	44.6 (±1.8)	62.3 (±3.1)	88.5 (±2.2)	**81.1 (±0.9)**	**67.6 (±0.9)**
XLM-R base	64.2	67.5 (±2.2)	63.4 (±12.3)	72.5 (±1.9)	**85.4 (±0.2)**	68.1 (±1.6)	15.7 (±12.7)	69.6 (±0.9)	72.6 (±4.2)	52.4 (±6.0)	44.1 (±7.9)	69.2 (±0.1)	45.9 (±1.8)	63.1 (±2.8)	85.3 (±1.5)	80.1 (±1.0)	63.0 (±0.7)
XLM-R large	**66.4**	58.1 (±9.3)	**70.4 (±3.7)**	**73.0 (±1.4)**	58.2 (±50.2)	**73.0 (±0.9)**	38.9 (±33.7)	70.0 (±1.8)	84.9 (±2.7)	**62.9 (±6.1)**	**53.8 (±10.5)**	**71.2 (±1.4)**	**47.5 (±3.7)**	54.9 (±3.7)	**88.7 (±1.1)**	**81.1 (±0.9)**	65.9 (±1.7)

Table 9: Macro-F1 and standard deviation for multilingual models from the validation set. The best scores are in bold.

Model	**Mean**	**BCD-J**	**BCD-U**	**GAM**	**GLC-V**	**GLC-C**	**GLC-S**	**SJP**	**OTS-UL**	**OTS-CT**	**C19**	**MEU-1**	**GLN**	**LNR**	**LNB**	**MAP-C**	**MAP-F**
MiniLM	51.7	49.4 (±7.4)	56.7 (±7.9)	**73.3 (±0.9)**	81.7 (±0.5)	39.4 (±1.4)	**5.2 (±1.6)**	67.6 (±1.2)	74.6 (±1.1)	14.1 (±3.1)	5.6 (±2.2)	62.0 (±0.4)	40.5 (±4.0)	46.8 (±1.9)	86.0 (±0.2)	63.0 (±2.1)	40.4 (±2.1)
DistilBERT	58.0	50.3 (±2.9)	58.8 (±8.7)	69.5 (±0.9)	85.2 (±0.8)	70.0 (±0.3)	33.2 (±1.9)	66.7 (±1.1)	67.2 (±4.1)	46.2 (±8.9)	39.5 (±6.3)	63.6 (±0.1)	38.1 (±2.0)	48.4 (±5.2)	78.7 (±1.1)	61.3 (±2.8)	40.6 (±0.7)
mDeBERTa v3	59.4	**65.8 (±3.5)**	49.3 (±0.1)	70.9 (±0.9)	85.6 (±1.0)	58.6 (±7.8)	12.4 (±2.8)	**69.0 (±0.8)**	79.7 (±3.8)	53.8 (±3.0)	40.7 (±5.0)	65.0 (±0.4)	42.2 (±1.6)	46.6 (±1.1)	87.8 (±0.7)	**65.3 (±3.1)**	**46.3 (±0.6)**
XLM-R base	61.2	65.4 (±3.6)	61.6 (±11.2)	72.0 (±2.4)	**85.9 (±0.1)**	69.3 (±1.6)	15.4 (±12.3)	68.3 (±1.0)	80.8 (±1.9)	55.9 (±2.6)	45.9 (±11.0)	65.6 (±0.1)	47.0 (±2.2)	47.7 (±2.9)	86.0 (±1.9)	61.4 (±2.8)	42.2 (±0.4)
XLM-R large	**63.2**	55.1 (±7.6)	**62.3 (±3.6)**	73.1 (±1.5)	58.3 (±50.3)	**74.7 (±0.9)**	39.1 (±33.9)	68.3 (±1.8)	**83.6 (±4.8)**	**66.9 (±0.5)**	**54.2 (±7.2)**	**68.1 (±1.2)**	**48.0 (±4.2)**	**49.5 (±11.3)**	**88.2 (±0.7)**	65.0 (±5.7)	46.2 (±2.1)

Table 10: Macro-F1 and standard deviation for multilingual models from the test set. The best scores are in bold.

D Histograms

In the following, we provide the histograms for the distribution of the sequence length of the input (sentence or entire document) from each dataset. The length is measured by counting the tokens using the tokenizers of the multilingual models, i.e., DistilBERT, MiniLM, mDeBERTa v3, XLM-R base, XLM-R large. We only display the distribution within the 99th percentile; the rest is grouped together at the end.

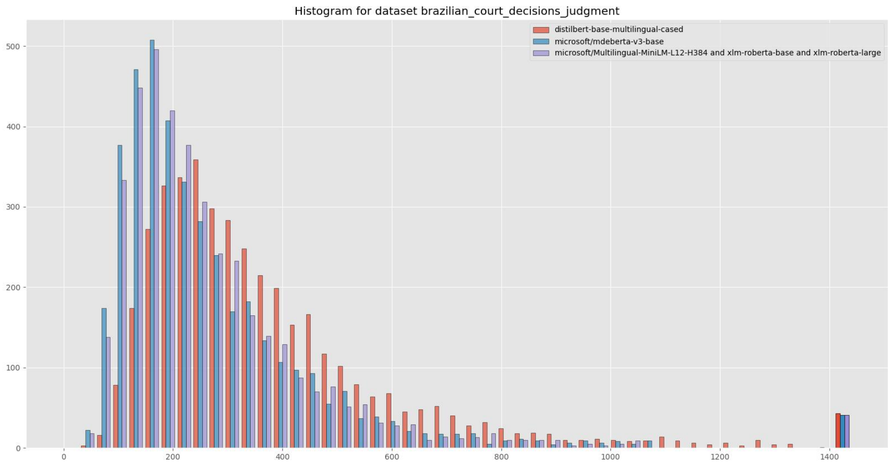


Figure 2: Histogram for dataset BCD-J

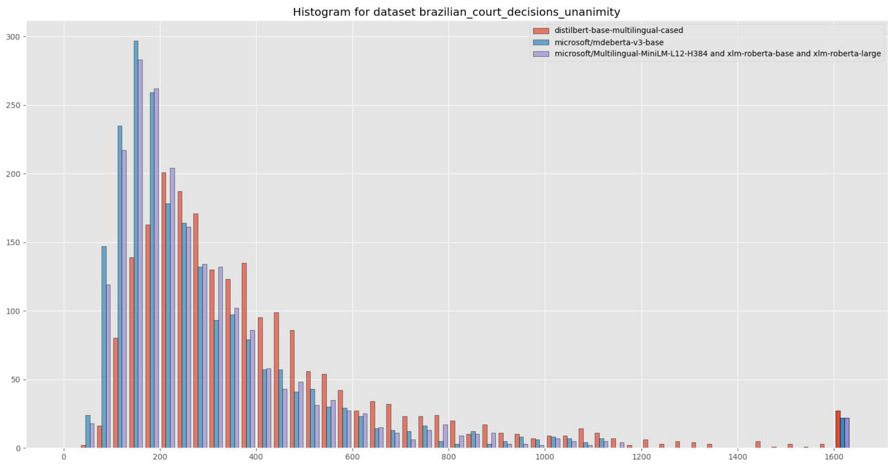


Figure 3: Histogram for dataset BCD-U

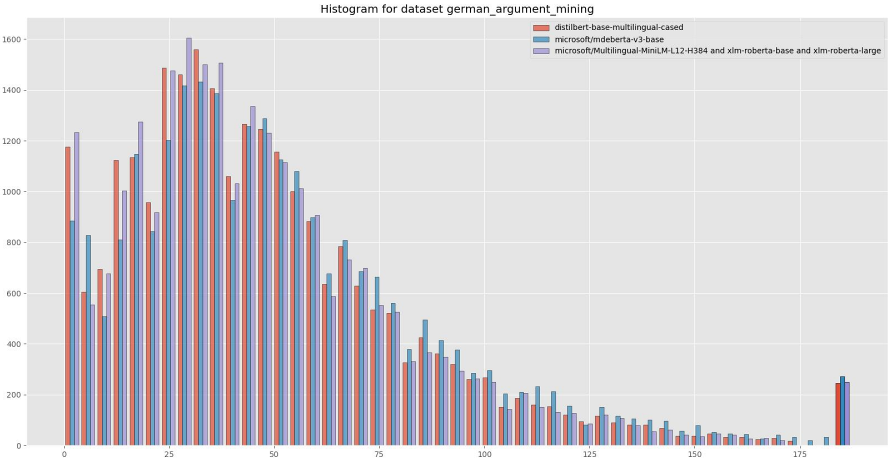


Figure 4: Histogram for dataset GAM

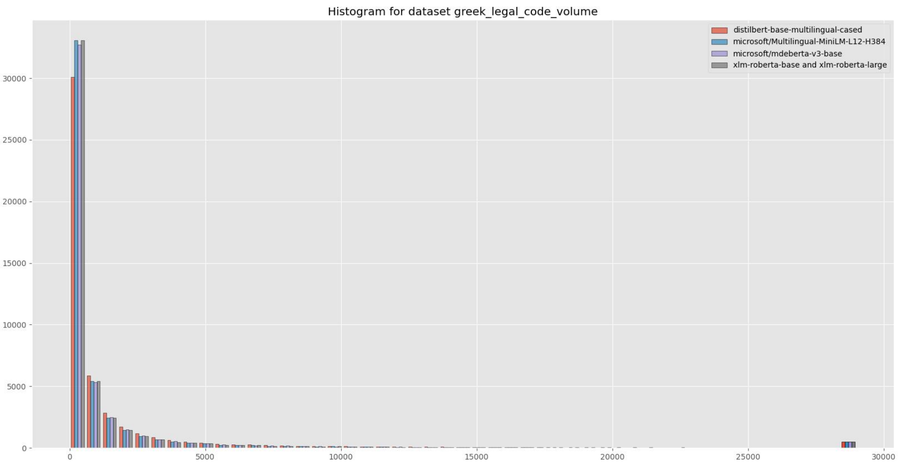


Figure 5: Histogram for dataset GLC-V

Figure 6: Histogram for dataset GLC-C

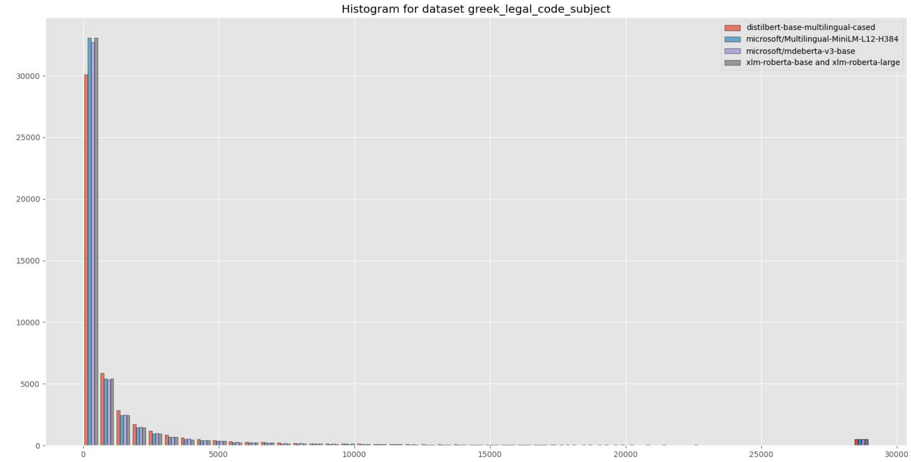


Figure 7: Histogram for dataset GLC-S

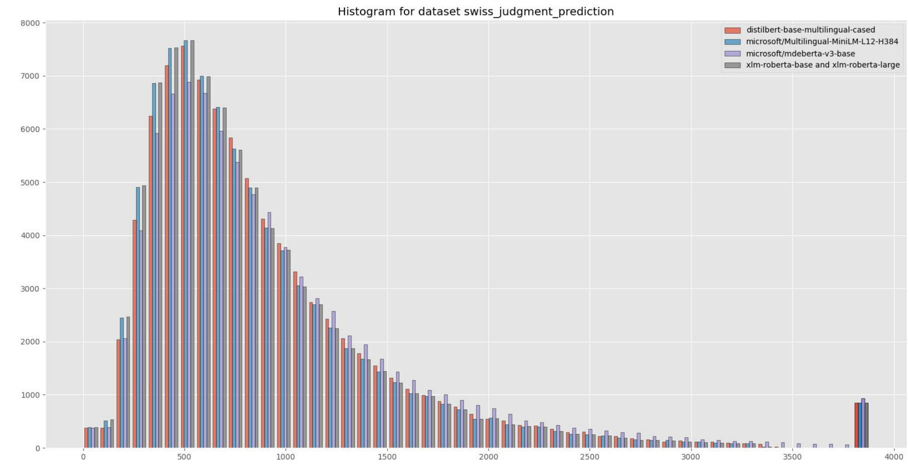


Figure 8: Histogram for dataset SJP

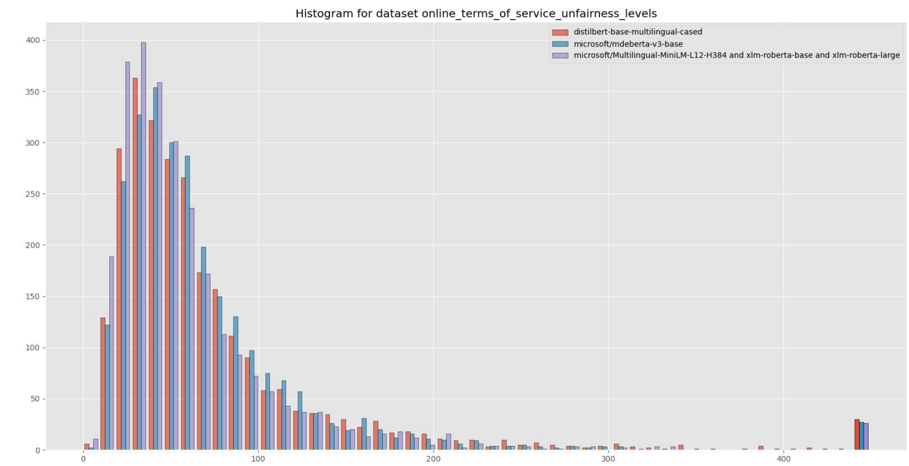


Figure 9: Histogram for dataset OTS-UL

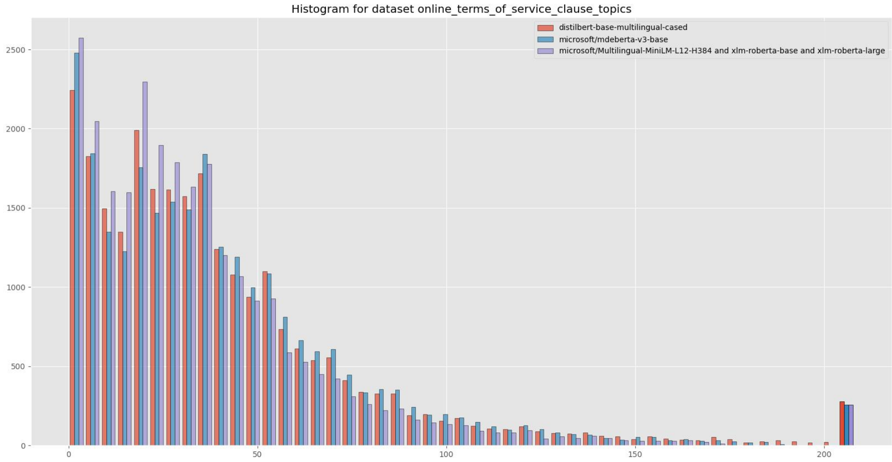


Figure 10: Histogram for dataset OTS-CT

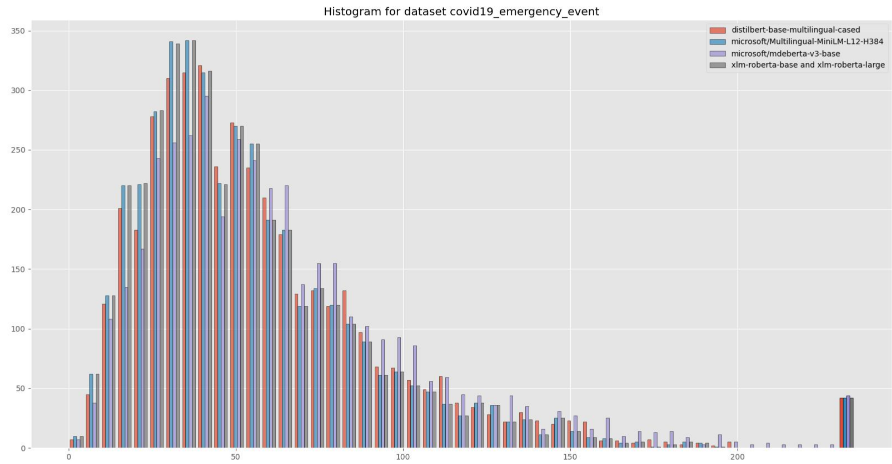


Figure 11: Histogram for dataset C19

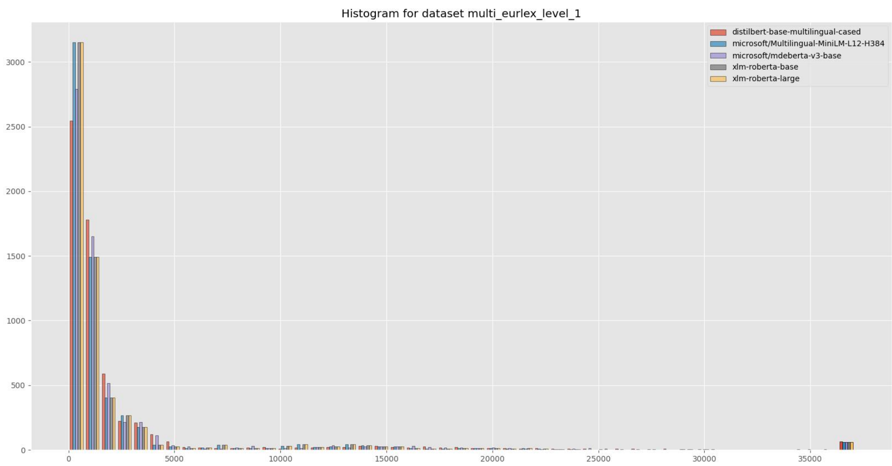


Figure 12: Histogram for dataset MEU-1

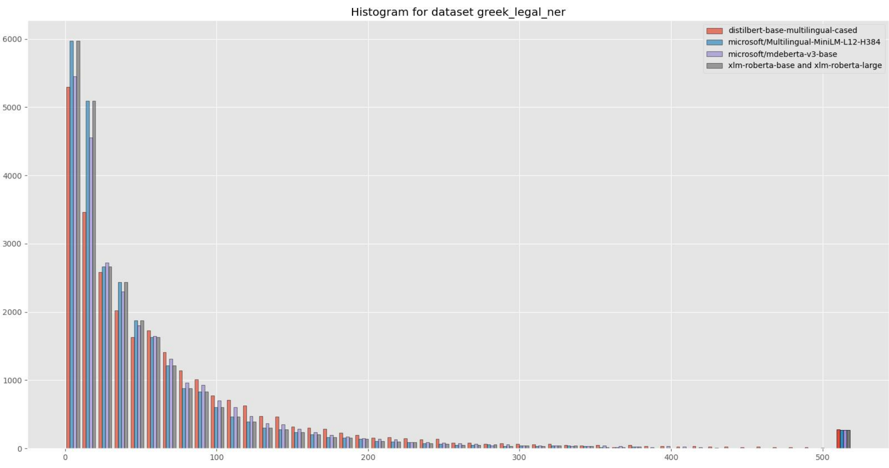


Figure 13: Histogram for dataset GLN

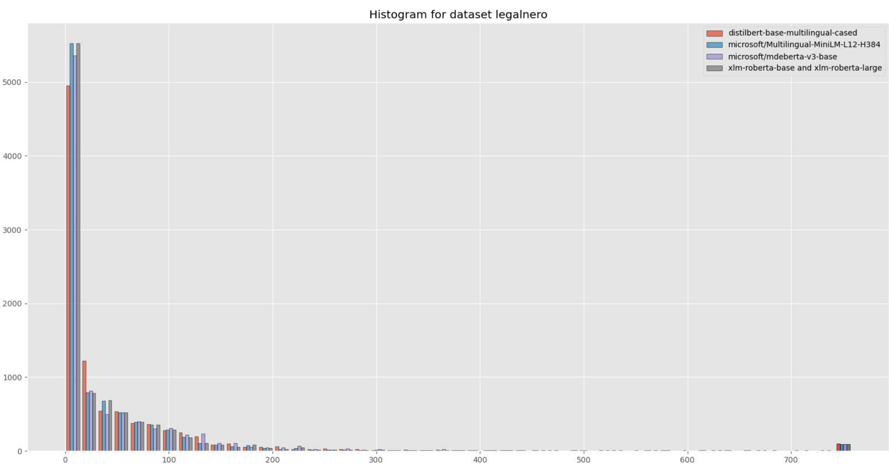


Figure 14: Histogram for dataset LNR

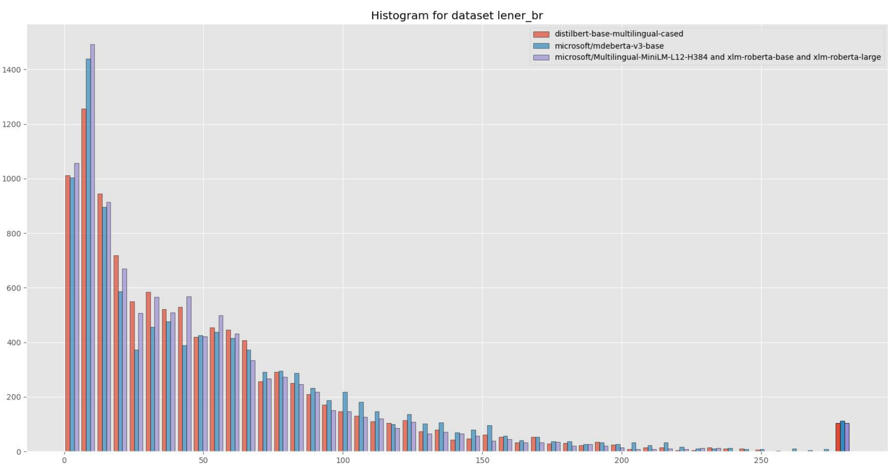


Figure 15: Histogram for dataset LNB

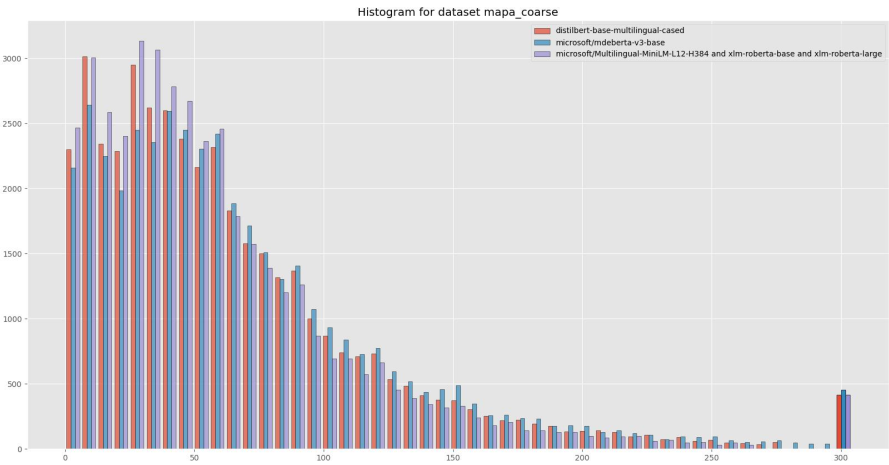


Figure 16: Histogram for dataset MAP-C

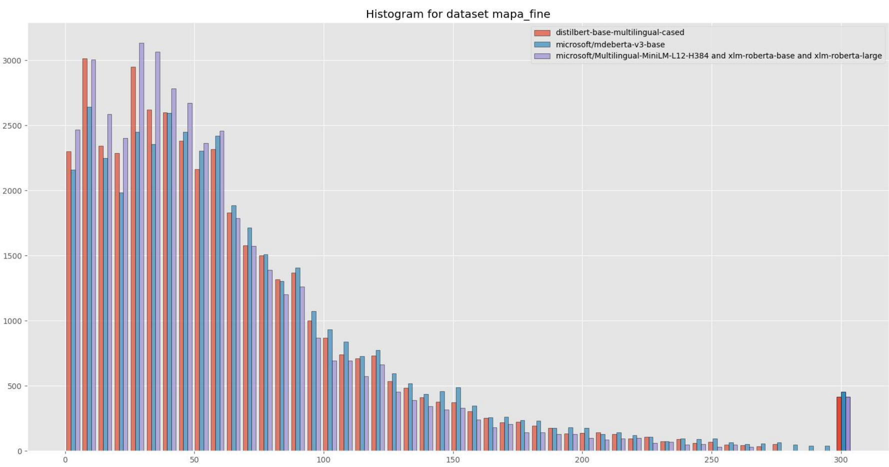


Figure 17: Histogram for dataset MAP-F

Automatic Anonymization of Swiss Federal Supreme Court Rulings

Joel Niklaus[1,2,3] * **Robin Mamié**[4] *

Matthias Stürmer[1,2] **Daniel Brunner**[4] **Marcel Gygli**[2]

[1]University of Bern [2]Bern University of Applied Sciences
[3]Stanford University [4]Swiss Federal Supreme Court

Abstract

Releasing court decisions to the public relies on proper anonymization to protect all involved parties, where necessary. The Swiss Federal Supreme Court relies on an existing system that combines different traditional computational methods with human experts. In this work, we enhance the existing anonymization software using a large dataset annotated with entities to be anonymized. We compared BERT-based models with models pre-trained on in-domain data. Our results show that using in-domain data to pre-train the models further improves the F1-score by more than 5% compared to existing models. Our work demonstrates that combining existing anonymization methods, such as regular expressions, with machine learning can further reduce manual labor and enhance automatic suggestions.

1 Introduction

The Swiss Federal Supreme Court (SFSC) is the highest judicial authority in Switzerland. It is the final arbiter in legal disputes and ensures the uniform application of federal law throughout the country. It consists of several divisions specialized in different areas of law, including civil, criminal, administrative, and social security matters (Glaser et al., 2021). In a year the SFSC roughly handles 7K cases and publishes its rulings. In this process, personal information must be anonymized from the rulings in order to protect involved parties. In the traditional setting, court rulings are anonymized by skilled experts. This task is highly complex as the removal/anonymization of a word is dependent on the context it is written in. For example *Zuerich* needs to be removed if it is part of the name of the legal entity "Zurich Insurance Group", but not if it is a reference to the city. At the SFSC, experts are already supported in their work through an application called *Anom2* (see Figure 1). *Anom2* provides access to various methods and algorithms for finding and replacing text entities (e.g., with regular expressions. The aim of this work is to enhance the capabilities of *Anom2* with Machine Learning capabilities, that provide the user with more suggestions that need to be anonymized. Our results show that this approach allows users to find more elements that require anonymization.

2 Related Work

For identifying elements that might require anonymization, a process called Named Entity Recognition (NER) is employed. Traditionally, NER recognizes and categorizes text parts according to a set of semantic categories like *Location (LOC)*, *Organization (ORG)*, or *Person (PER)* (Benikova et al., 2014). As these classes are not enough for the anonymization of court cases (Leitner et al., 2020) suggested enlarging this list to seven coarse, and 19 fine-grained classes, including entities such as *Judge (RR)*, or *Lawyer (AN)*. Using this dataset (Darji et al., 2023) fine-tuned GermanBERT (Chan et al., 2020), clearly outperforming a BiLSTM-CRF+ model. Similar approaches have been applied and tested in other languages, such as Romanian (Pais et al., 2021), Greek (Angelidis et al., 2018), Portuguese (Luz de Araujo et al., 2018), and multilingually (de Gibert et al., 2022; Niklaus et al., 2023a).

Domain specific pretraining has flourished in the legal domain recently. Chalkidis et al. (2020) pretrained LegalBERT on EU and UK legislation, ECHR and US cases and US contracts. Zheng et al. (2021) pretrained CaseHoldBERT on US case law, while Henderson et al. (2022) trained PoL-BERT on the 256 GB Pile of Law corpus. Niklaus and Giofré (2022) pretrained Longformer (Beltagy et al., 2020) models using the Replaced

* Equal contribution.

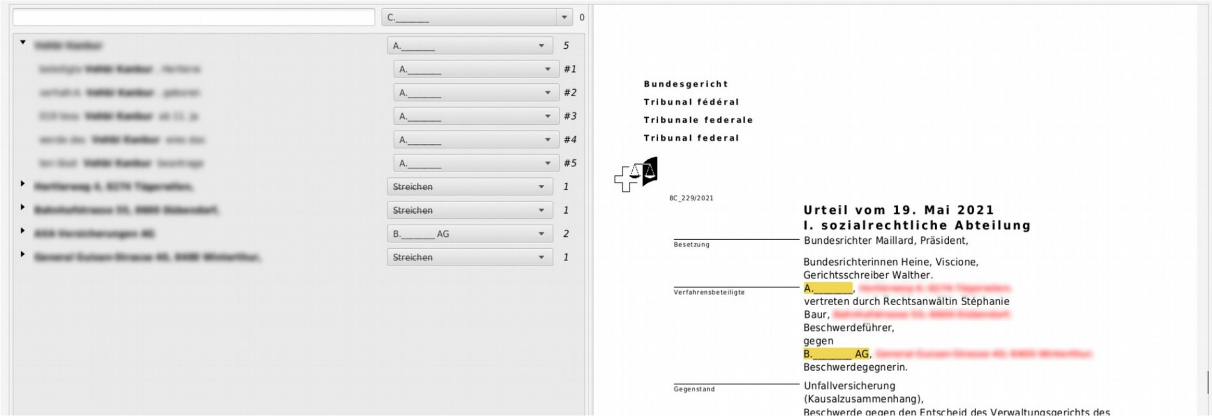


Figure 1: Main window of *Anom2*. Anonymizations are configured on the left, and the anonymized court ruling appears on the right. The system highlights completed anonymizations in gold and the current setting in yellow.

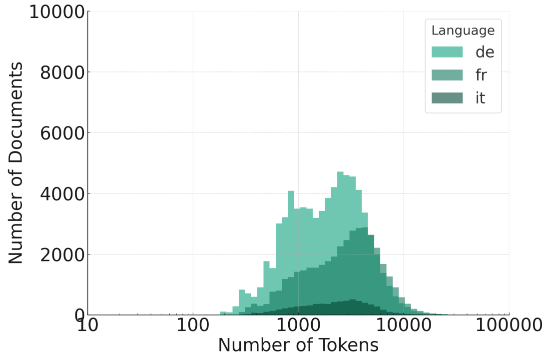
Token Detection (RTD) task on the Pile of Law. Hua et al. (2022) used RTD to pretrain Reformer (Kitaev et al., 2020) models on 6 GB of US case law. Finally, Niklaus et al. (2023b) released a large multilingual legal corpus and trained various legal models. We continue pretraining the German, French, and Italian models for 800K and 300K steps more for base and large models, respectively. Rasiah et al. (2023) pretrain models on Swiss legal data, termed Legal-Swiss-RoBERTa.

Document anonymization has a long tradition in the medical domain, where personal data needs to be removed from documents. Initially, this task was handled using methods like semantic lexicons (Ruch et al., 2000) or regular expressions to replace text occurrences. Recently, this has been expanded to include BERT-style models as well (Mao and Liu, 2019). In the legal domain, Glaser et al. (2021) worked on 1400 anonymized German rulings. Using already anonymized rulings, they trained different Recurrent Neural Networks (RNN) using BERT embeddings. Using this approach, they achieved a maximum of 68.9% precision and 79.1% recall rates. Garat and Wonsever (2022) performed similar work on 80K documents from Uruguayan courts. Our work specifically tackles court decisions by the SFSC. We compare the generic cased mBERT model (Devlin et al., 2019a) with models pre-trained on in-domain data (such as Legal-Swiss-RoBERTa-base (Rasiah et al., 2023)). We also investigate monolingual model performance in the three languages of the SFSC rulings: German, French, and Italian.

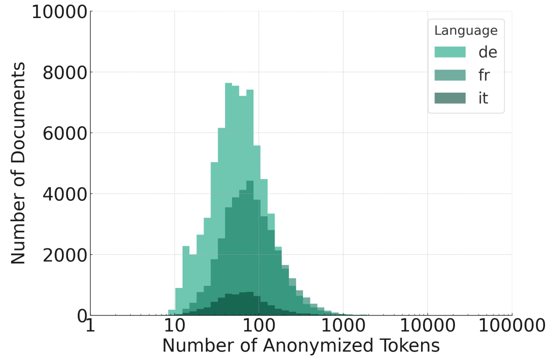
Much prior work used SFSC cases as data for their research due to wide availability in three languages, giving a good coverage of the most important Swiss case law. Niklaus et al. (2021, 2022) introduced and studied judgment prediction on SFSC rulings. Brugger et al. (2023) investigated and improved multilingual sentence boundary detection in the legal domain using SFSC decisions. Christen et al. (2023) studied negation scope resolution and Nyffenegger et al. (2023) investigated how easily LLMs can re-identify persons occurring in anonymized SFSC decisions. Rasiah et al. (2023) created a large benchmark of ten text classification tasks, two text generation tasks, an information retrieval and a citation extraction task.

3 Dataset

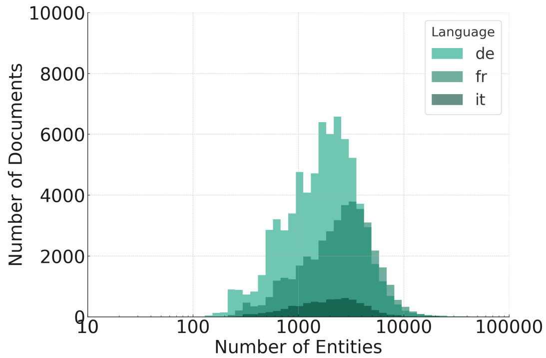
We used all supreme court decisions from the SFSC and split them into sentences using Spacy (Honnibal et al., 2020). We prepared court decisions for NER based on the manual labels from the paralegals who performed manual anonymizations. In total, we used 119156 rulings (77262 German, 40099 French, 6795 Italian). The histograms in Figure 2 illustrate the distribution of four key measures, namely Number of Tokens, Number of Anonymized Tokens, Number of Entities, and Number of Anonymized Entities, in three languages. German (de), French (fr), and Italian (it). Different color schemes for each language enhance the visual interpretability of the plots. The measures concerning tokens and entities exhibit a long-tailed distribution, signifying a concentration of instances at the lower end of the value spectrum. Specifically, the distribution of Number of Tokens and Number of Entities is examined within a range of 10 to 100,000, capturing the broad spread of these measures. In contrast, the measures concerning anonymized tokens and entities are evaluated



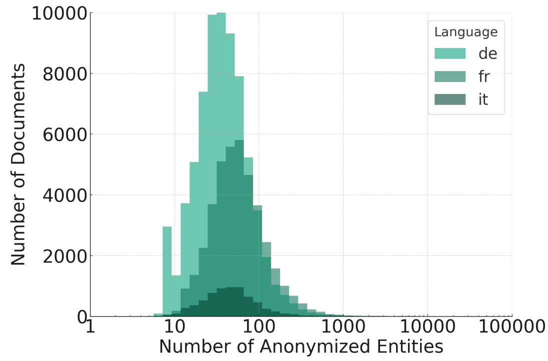
(a) Token Distribution



(b) Anonymized Token Distribution



(c) Entity Distribution



(d) Anonymized Entity Distribution

Figure 2: Histograms illustrating the distribution of (anonymized) tokens and entities across the three languages.

within a narrower range of 1 to 10,000, reflecting their more constrained distribution.

4 Legal Pretraining

To improve the SFSC anonymization system, we pretrained legal-specific models on diverse legal text in German, French, and Italian.

(a) We warm-start (initialize) our models from the original XLM-R checkpoints (base or large) of Conneau and Lample (2019). Model recycling is a standard process followed by many (Wei et al., 2021; Ouyang et al., 2022) to benefit from starting from an available "well-trained" PLM, rather from scratch (random). XLM-R was trained on 2.5 TB of cleaned CommonCrawl data in 100 languages.

(b) We train a new tokenizer of 32K BPEs on the training subsets to better cover legal language. However, we reuse the original XLM-R embeddings for all lexically overlapping tokens (Pfeiffer et al., 2021), i.e., we warm-start word embeddings for tokens that already exist in the original XLM-R vocabulary, and use random ones for the rest.

(c) We continue pretraining our monolingual models on our pretraining corpus with batches of 512 samples for an additional 1M/500K steps for the base/large model. We do initial warm-up steps for the first 5% of the total training steps with a linearly increasing learning rate up to $1e{-4}$, and then follow a cosine decay scheduling, following recent trends. For half of the warm-up phase (2.5%), the Transformer encoder is frozen, and only the embeddings, shared between input and output (MLM), are updated. We also use an increased 20/30% masking rate for base/large models respectively, where also 100% of the predictions are based on masked tokens, compared to Devlin et al. (2019b)[1], based on the findings of Wettig et al. (2023).

(d) We consider mixed cased models, i.e., both upper- and lowercase letters covered, similar to recently developed large PLMs (Conneau and Lample, 2019; Raffel et al., 2020; Brown et al., 2020).

(e) This leaves us with two models for each language (base and large). Additionally, we consider the multilingual legal models pretrained by Niklaus et al. (2023b) and the Swiss legal models pretrained by Rasiah et al. (2023).

[1] Devlin et al. (2019b) – and much follow-up work – used a 15% masking ratio, and a recipe of 80/10/10% of predictions made across masked/randomly-replaced/original tokens.

Table 1: Evaluation Results. Best results per setup are in **bold**.

Model	Normal			Uniformizing		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Multilingual Models						
bert-base-multilingual-cased	90.72	83.76	87.10	85.85	**94.95**	**90.17**
Legal-XLM-RoBERTa-base	**94.84**	81.98	87.94	**89.93**	86.85	88.36
Legal-Swiss-RoBERTa-base	92.26	**92.57**	**92.42**	83.13	94.85	88.60
Monolingual Models						
bert-base-german-cased	95.14	80.00	86.92	**91.49**	**85.86**	**88.58**
Legal-German-RoBERTa-base	**95.40**	**80.09**	**87.07**	89.20	84.97	87.03
dbmdz/bert-base-french-europeana-cased	**95.86**	81.84	88.30	**88.92**	89.14	**89.03**
Legal-French-RoBERTa-base	95.45	**83.48**	**89.06**	88.77	**89.17**	88.97
dbmdz/bert-base-italian-cased	93.49	80.21	86.35	76.71	83.85	80.12
Legal-Italian-RoBERTa-base	**94.16**	**80.59**	**86.85**	**84.03**	**84.06**	**84.05**

5 Anonymization System

The SFSC employs an anonymization system, *Anom2*, to assist paralegals in anonymizing rulings for public access. The main UI is shown in Figure 1. Upon loading a ruling, the application auto-identifies terms requiring anonymization and lists them on the left, along with replacement text. The search function allows direct term marking for anonymization. *Anom2* uses different algorithms for the search for text that needs to be anonymized:
Conventional is based on a statistical analysis of the loaded ruling. Using *polyglot*[2] an initial set of named entities is detected. Using the specific knowledge of the format, the rubrum is dynamically detected, allowing for the labelling of important names and addresses.
BERT performs the recognition of entities to be anonymized using a BERT (Devlin et al., 2019a) model fine-tuned for NER. Entity recognition is performed on the sentence level, as the rulings are often too long for the model. This approach could lead to inconsistencies in recognition, as a term identified in one sentence might not be identified in another. This is solved in post-processing, where any identified term is automatically anonymized in the whole document.
Legal-Swiss-RoBERTa-base works analogously to the BERT method, but uses a fine-tuned *Legal-Swiss-RoBERTa-base* (Rasiah et al., 2023) model.

[2]See: https://polyglot.readthedocs.io

6 Experimental Setup

We used the following hyperparameters for all evaluated models: batch size of 64, learning rate of 5e-5, and weight decay of 0.01. We employed the seqeval metric for evaluation. We set the maximum sequence length to 192 tokens, which we determined to be the optimal trade-off between average sentence size and training time for computational efficiency. We used early stopping based on the F1-score of the validation set, which constitutes 10% of the entire dataset, following an 80-10-10 split for the training, validation, and test sets, respectively. Training ceases once the F1-score on the validation set starts to decline. Due to resource constraints (we only had two Tesla T4 GPUs) we could only run one random seed per model.

We define and configure two special parameters:
1) *TruncationStrideRatio*: We set this parameter to 0.5. When a sentence exceeds 192 tokens, we truncate it using a specific overlap strategy. The overlap consists of half of the previous snippet and half of the next snippet.
2) *NonAnonymizedSentencesRatioToAnonymizedSentences*: We set the ratio at 1.5, including only 150% of sentences without anonymization examples compared to those with examples. This minimizes data redundancy and maximizes utility.

7 Results

Table 1 presents a comprehensive evaluation of various BERT and RoBERTa-based models on two different conditions: Normal and Uniformizing. For the Normal condition, in the multilingual setting, Legal-XLM-RoBERTa-base exhibits the highest Precision at 94.84%, while Legal-Swiss-RoBERTa-base demonstrates superior Recall and F1-Score values, achieving 92.57% and 92.42% respectively. With Uniformizing, we describe the process of forcing the model to replace all occurrences of a detected term across the whole document. This approach leads to better Recall, but reduces Precision. In the Uniformized case, again Legal-XLM-RoBERTa-base shows highest Precision, while mBERT achieves highest Recall and F1-Score. The improved Recall and F1-Score in the Normal condition show that pre-training on legal data can improve the performance of models. We observe similar behavior for the monolingual models. All models pre-trained on legal data achieve a higher F1-Score than generic monolingual models.

8 Discussion

We pretrained models on Swiss legal data and performed a detailed comparison of legal and generic models, both multilingually and monolingually in the ruling anonymization task. Our experiments indicate that pretraining on legal data improves the performance of models significantly compared to generic multi- or monolingual models.

To reduce errors in sentence splitting, we suggest future work to use legal specific sentence splitters (Brugger et al., 2023). Due to computational constraints we only experimented with base size encoder models. Future work may expand this by also testing larger models.

Acknowledgements

We greatly appreciate Google's generous support of TPUs v3-8 machines for pretraining the models. This work has been partially supported by the Swiss National Research Programme "Digital Transformation" (NRP-77) grant number 187477.

References

I. Angelidis, Ilias Chalkidis, and M. Koubarakis. 2018. Named Entity Recognition, Linking and Generation for Greek Legislation. In *JURIX*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tobias Brugger, Matthias Stürmer, and Joel Niklaus. 2023. MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset. ArXiv:2305.01211 [cs].

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. *arXiv:2010.10906 [cs]*. ArXiv: 2010.10906.

Ramona Christen, Anastassia Shaitarova, Matthias Stürmer, and Joel Niklaus. 2023. Resolving Legalese: A Multilingual Exploration of Negation Scope Resolution in Legal Documents.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2023. German BERT Model for Legal Named Entity Recognition:. In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence*, pages 723–728, Lisbon, Portugal. SCITEPRESS - Science and Technology Publications.

Ona de Gibert, A García-Pablos, Montse Cuadros, and Maite Melero. 2022. Spanish datasets for sensitive entity detection in the legal domain. In *Proceedings of the Thirteenth International Conference on Language Resources and Evaluation (LREC'22), Marseille, France, june. European Language Resource Association (ELRA)*. Dataset URL: https://tinyurl.com/mv65cp66.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference* *of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diego Garat and Dina Wonsever. 2022. Automatic Curation of Court Documents: Anonymizing Personal Data. *Information*, 13(1):27.

Ingo Glaser, Tom Schamberger, and Florian Matthes. 2021. Anonymization of german legal court rulings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, pages 205–209, São Paulo Brazil. ACM.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. 2022. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset. ArXiv:2207.00220 [cs].

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber. 2022. LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension. ArXiv:2212.08204 [cs].

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The Efficient Transformer. *arXiv:2001.04451 [cs, stat]*. ArXiv: 2001.04451.

Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. 2020. A Dataset of German Legal Documents for Named Entity Recognition. *arXiv:2003.13016 [cs]*. ArXiv: 2003.13016.

Pedro Henrique Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pages 313–323, Cham. Springer International Publishing.

Jihang Mao and Wanli Liu. 2019. Hadoken: a BERT-CRF Model for Medical Document Anonymization. In *Proceedings of the Iberian Languages Evaluation Forum co-located with 35th Conference of the Spanish Society for Natural Language Processing, IberLEF@SEPLN 2019, Bilbao, Spain, September 24th, 2019*, volume 2421 of *CEUR Workshop Proceedings*, pages 720–726. CEUR-WS.org.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus and Daniele Giofré. 2022. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch? ArXiv:2211.17135 [cs].

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023a. Lextreme: A multi-lingual and multi-task benchmark for the legal domain.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2023b. MultiLegalPile: A 689GB Multilingual Legal Corpus. ArXiv:2306.02069 [cs].

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only. Association for Computational Linguistics.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2023. Anonymity at Risk? Assessing Re-Identification Capabilities of Large Language Models. ArXiv:2308.11103 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Vasile Pais, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2021. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E. Ho, and Joel Niklaus. 2023. SCALE: Scaling up the Complexity for Advanced Language Model Evaluation. ArXiv:2306.09237 [cs].

P. Ruch, R. H. Baud, A. M. Rassinoux, P. Bouillon, and G. Robert. 2000. Medical document anonymization with a semantic lexicon. *Proceedings. AMIA Symposium*, pages 729–733.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. *arXiv:2104.08671 [cs]*. ArXiv: 2104.08671 version: 3.

MultiLegalPile: A 689GB Multilingual Legal Corpus

Joel Niklaus [1 2 3] Veton Matoshi [2] Matthias Stürmer [1 2] Ilias Chalkidis [4] Daniel E. Ho [3]

Abstract

Large, high-quality datasets are crucial for training Large Language Models (LLMs). However, so far, there are few datasets available for specialized critical domains such as law and the available ones are often only for the English language. We curate and release MULTILEGALPILE, a 689GB corpus in 24 languages from 17 jurisdictions. The MULTILEGALPILE corpus, which includes diverse legal data sources with varying licenses, allows for pretraining NLP models under fair use, with more permissive licenses for the Eurlex Resources and Legal mC4 subsets. We pretrain two RoBERTa models and one Longformer multilingually, and 24 monolingual models on each of the language-specific subsets and evaluate them on LEXTREME. Additionally, we evaluate the English and multilingual models on LexGLUE. Our multilingual models set a new SotA on LEXTREME and our English models on LexGLUE. We release the dataset, the trained models, and all of the code under the most open possible licenses.

1. Introduction

Recent years have seen LLMs achieving remarkable progress, as demonstrated by their performance on various benchmarks such as SuperGLUE (Wang et al., 2019), MMLU (Hendrycks et al., 2021), and several human Exams (OpenAI, 2023), including U.S. bar exams for admission to practicing the law (Katz et al., 2023). These models are typically trained on increasingly large corpora, such as the Pile (Gao et al., 2020a), C4 (Raffel et al., 2020b), and mC4 (Xue et al., 2021). However, it is important to note that public corpora available for training these models are predominantly in English, and often constitute web text with unclear licensing. This even led to lawsuits against LLM producers, highlighting this critical issue. Furthermore, there is a scarcity of large-scale, domain-specific pretraining corpora, which constitutes a significant gap in the current body of resources available for the training of LLMs. Similarly, LLMs are predominantly English, especially considering domain-specific models, e.g., ones specialized in biomedical, legal, or financial texts.

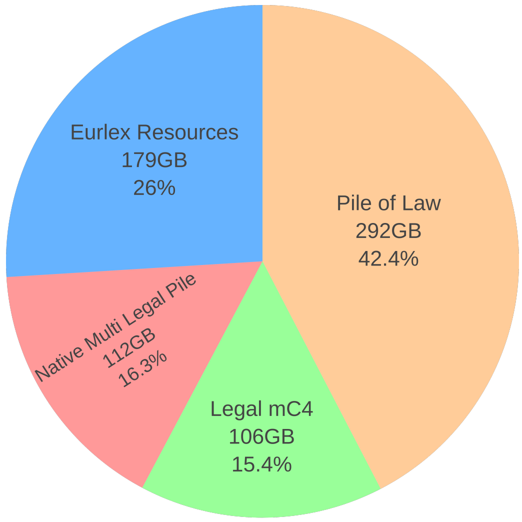


Figure 1. MULTILEGALPILE Source Distribution

Legal texts, often produced by public instruments (e.g., state governments, international organizations), are typically available under public licenses, offering a rich resource for domain-specific pretraining. Given this context, we curate a humongous, openly available, corpus of multilingual law text spanning across numerous jurisdictions (legal systems), predominantly under permissive licenses.

Further on, we continue pretraining XLM-R models (Conneau & Lample, 2019) on our corpus and evaluated these models on the recently introduced LEXTREME (Niklaus et al., 2023) and LexGLUE (Chalkidis et al., 2021e) benchmarks. Given the often extensive nature of legal text, we also pretrained a Longformer model (Beltagy et al., 2020) for comparison with hierarchical models (Chalkidis et al., 2019b; Niklaus et al., 2021; 2022).

Our multilingual models set new SotA on LEXTREME overall. Our legal Longformer outperforms all other models in four LEXTREME datasets and reaches the highest dataset aggregate score. Our monolingual models outperform their base model XLM-R in 21 out of 24 languages, even reaching

[1]Institute of Computer Science, University of Bern, Bern, Switzerland [2]Bern University of Applied Sciences [3]Stanford University [4]University of Copenhagen. Correspondence to: Joel Niklaus <joel.niklaus@unibe.ch>.

language specific SotA in five. On LexGLUE our English models reach SotA in five out of seven tasks with the large model achieving the highest aggregate score.

In the spirit of open science, we provide the dataset under a CC BY-NC-SA 4.0 license, with some subsets licensed more permissively. Dataset creation scripts, models, and pretraining code are public under Apache 2.0 licenses. This open-source approach encourages further research and advancements in the field of legal text analysis and understanding using large language models.

Contributions

The contributions of this paper are three-fold:

1. We curate and release a large scale multilingual legal text corpus, dubbed MULTILEGALPILE,[1] covering 24 languages and 17 legal systems (jurisdictions).
2. We release 2 multilingual and 24 monolingual new legal-oriented PLMs, dubbed LEGALXLMS, warm-started from the XLM-R (Conneau & Lample, 2019) models, and further pretrained on the MULTILEGALPILE. Additionally, we pretrain a Longformer (Beltagy et al., 2020) based on our multilingual base-size model on context lengths of up to 4096 tokens.
3. We benchmark the newly released models on the LEXTREME and LexGLUE benchmarks, achieving new SotA for base and large size models and increasing performance drastically in Greek legal code. Our Longformer model reaches SotA in four tasks and the highest dataset aggregate score. Our monolingual models set language specific SotA in five languages.

2. Related Work

2.1. General Pretraining Corpora

The use of pretrained Language Models (PLMs) has become increasingly popular in NLP tasks, particularly with the advent of models such as BERT (Devlin et al., 2019) that can be finetuned for specific applications. One key factor in the success of pretraining is the availability of large and diverse text corpora, which can help the model learn the nuances of natural language. In the following, we discuss large-scale general-purpose text corpora used for pretraining.

Wikipedia is a commonly used multilingual dataset for pretraining language models, and has been used to pretrain BERT (Devlin et al., 2019), MegatronBERT (Shoeybi et al., 2020), T5 (Raffel et al., 2020a), and GPT-3 (Brown et al., 2020b), among others.

Based on Wikipedia, Merity et al. (2016) created WikiText by selecting articles fitting the Good or Featured article criteria. The dataset contains 103M words and has two versions: WikiText2 and the larger WikiText103. It has been used to pretrain models like MegatronBERT (Shoeybi et al., 2020) and GPT-2 (Radford et al., 2019).

The BookCorpus (Zhu et al., 2015), also known as the Toronto Books Corpus, is an English dataset used for pretraining BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and T5 (Raffel et al., 2020a). It consists of almost 1B words from over 11K books collected from the web.

The Common Crawl corpus is a publicly available multilingual dataset of scraped web pages, regularly updated with new "snapshots". It has been used to pretrain GPT-3 (Brown et al., 2020b) as well as XLM-R (Conneau et al., 2020a). One significant drawback of Common Crawl is the presence of uncleaned data, which includes a considerable amount of "gibberish or boiler-plate text like menus, error messages, or duplicate text" (Raffel et al., 2020a). As a result, utilizing the Common Crawl dataset necessitates additional post-filtering and cleaning procedures. To address this issue, Raffel et al. (Raffel et al., 2020a) performed several cleaning steps on the April 2019 snapshot of Common Crawl, resulting in the creation of the Colossal Clean Crawled Corpus (C4), comprising 750 GB of English-language text. It was used for pretraining models such as T5 (Raffel et al., 2020a) and Switch Transformer (Fedus et al., 2022).

OpenWebText (Gokaslan & Cohen, 2019) openly replicates OpenAI's closed English WebText dataset (Radford et al., 2019), used to pretrain GPT-2 (Radford et al., 2019). WebText comprises over 8M documents with a combined text size of 40 GB. To ensure data uniqueness, any documents sourced from Wikipedia were excluded from WebText, as they are commonly utilized in other datasets. OpenWebText, on the other hand, consists of 38 GB of text data from 8M documents and was used for pretraining RoBERTa (Liu et al., 2019) and MegatronBERT (Shoeybi et al., 2020).

News articles are also a common source for pretraining corpora. The RealNews dataset (Zellers et al., 2019) is a large corpus extracted from Common Crawl, containing news articles from December 2016 to March 2019 (training) and April 2019 (evaluation), totaling 120 GB. It was used for pretraining MegatronBERT (Shoeybi et al., 2020). For pretraining RoBERTa, Liu et al. (2019) used an English subset of RealNews, comprising 63M English news articles crawled from September 2016 to February 2019.

The rise of LLMs brought about the creation of ever larger training datasets. The Pile (Gao et al., 2020b) combines 22 distinct, well-curated datasets, such as Wikipedia (English), OpenWebText2 (Gokaslan & Cohen, 2019), OpenSubtitles (Tiedemann, 2016) etc., encompassing 825 GB of data. Besides general-purpose textual datasets, it also contains domain-specific datasets, such as ArXiv (Science),

[1] https://huggingface.co/datasets/joelito/Multi_Legal_Pile

FreeLaw (Legal), PubMed Abstracts (Biomedicine), and GitHub data (to improve code-related task performance (Gao et al., 2020b)). GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020b) were evaluated on this dataset.

In their work, Touvron et al. (2023) compiled a substantial dataset from various publicly available sources, including CommonCrawl, C4, Github, Wikipedia, etc., totaling 1.4T tokens. They trained the 13B-parameter LLaMA model using this dataset, surpassing the performance of the 175B-parameter GPT-3 on most benchmark tasks. However, the dataset itself is not publicly available. To address this, a collaborative effort resulted in the creation of the RedPajama-Data-1T dataset, replicating LLaMA's dataset with a similar size of 1.2T tokens.

Some of the afore-mentioned datasets, such as Common Crawl, are used to pretrain multilingual versions of BERT, DistilBERT, RoBERTa etc. These models were pretrained on datasets that cover approximately 100 languages, thereby neglecting low-resource languages. ImaniGooghari et al. (2023) addressed this by compiling Glot500, a 700 GB dataset covering 500 diverse languages, with a focus on low-resource ones. The Glot500-m model, pretrained on this dataset, outperformed the XLM-RoBERTa base model on six out of seven tasks.

2.2. Domain Specific Corpora

While pretraining on general-purpose text like Wikipedia and news articles shows promise, evidence suggests that pretraining on domain-specific text can enhance language model performance on related tasks (Beltagy et al., 2019; Gu et al., 2021; Chalkidis et al., 2020b; Niklaus & Giofré, 2022). Domain-specific text corpora include texts specific to fields like medicine, law, or science.

Several studies have examined pretraining on scientific text corpora. Beltagy et al. (2019) pretrained SciBERT, a BERT-based model, on a random subset of 1.14M papers sourced from Semantic Scholar. This collection comprises 18% of computer science papers and 82% of papers from the broader biomedical field. Similarly, PubMed and PubMedCentral are common sources for biomedical datasets. Gu et al. (2021) trained PubMedBERT using PubMed abstracts and PubMedCentral articles; BioBERT (Lee et al., 2020) was pretrained similarly. Johnson et al. (2016) compiled the Medical Information Mart for Intensive Care III (MIMIC-III) dataset, a large single-center database of critical care patients. Huang et al. (2019) used over 2 million de-identified clinical notes from this dataset to pretrain ClinicalBERT. These models outperformed general-purpose models on biomedical NLP tasks.

In the legal domain, similar strategies are observed. Chalkidis et al. (2020a) collected 12 GB of diverse English legal texts, including legislation, court cases, and contracts. They pretrained LegalBERT on this dataset, showing state-of-the-art performance, especially in tasks requiring domain knowledge. Another study by Zheng et al. (2021) used the entire English Harvard Law case corpus (1965-2021) comprising 37 GB of text to pretrain CaseLaw-BERT.

Recently, Chalkidis* et al. (2023) released LexFiles, an English legal corpus with 11 sub-corpora covering legislation and case law from six English-speaking legal systems (EU, Council of Europe, Canada, US, UK, India). The corpus contains approx. 6M documents or approx. 19B tokens. They trained two new legal English PLMs, showing improved performance in legal probing and classification tasks.

Efforts to pretrain legal language models also exist for Italian (Licari & Comandè, 2022), Romanian (Masala et al., 2021), and Spanish (Gutiérrez-Fandiño et al., 2021). However, English dominates, underscoring the importance of compiling multilingual legal corpora.

Model	Domain	Languages	Size in # Words
SciBERT (Beltagy et al., 2019)	scientific	English	2.38B (3.17B tokens)
Galactica (Taylor et al., 2022)	scientific	English	79.5B (106B tokens)
BioBERT (Lee et al., 2019)	biomedical	English	18B
LegalBERT (Chalkidis et al., 2020b)	legal	English	1.44B (11.5GB)
CaselawBERT (Zheng et al., 2021)	legal	English	4.63B (37GB)
LegalXLMs (ours)	legal	24 EU langs	87B (689GB)

Table 1. Previous domain specific pretraining corpora. For some corpora only GB or tokens were available. We converted 8 GB into 1B words and 1 token to 0.75 words.

Table 1 compares previous domain-specific corpora, all in English. In terms of size, none reach the MULTILEGALPILE proposed here.

3. MULTILEGALPILE

3.1. Construction

We transformed all datasets into xz compressed JSON Lines (JSONL) format. The combination of XZ compression and JSONL is ideal for streaming large datasets due to reduced file size and efficient decompression and reading.

Filtering mC4 We employed the vast multilingual web crawl corpus, mC4 (Xue et al., 2021), as our foundation. To effectively filter this corpus for legal content, we utilized regular expressions to identify documents with legal references. We found that detecting legal citations, such as references to laws and rulings, served as a reliable indicator of legal-specific documents in the corpus.

Iteration	German	English	Spanish	French	Italian
1st	100%	20%	100%	65%	80%
2nd	100%	85%	100%	100%	95%

Table 2. Precision of investigated languages in legal mC4 (n=20)

In order to ensure the accuracy of our filtering, we engaged legal experts to aid in identifying citations to laws and rulings across different jurisdictions and languages. We manually reviewed the precision of the retrieved documents for five languages, namely German, English, Spanish, French, and Italian, as shown in Table 2. The proficiency levels of the evaluators included native German, fluent English and Spanish, intermediate French, and basic Italian.

Subsequent to the initial review, we performed a second round of precision evaluation, during which we refined our regex expressions based on our findings from the first iteration. This iterative process not only enhanced the precision of the legal content detection, but also resulted in a reduction of the corpus size from 133GB to 106GB. Although the overall volume of data was reduced, this process significantly improved the quality and specificity of the corpus by focusing on legal content with a higher degree of precision.

A major reason for utilizing regexes instead of a Machine Learning (ML) based classifier was speed. Already when utilizing regexes, filtering through such a huge corpus like mC4 (27TB in total, of which 10.4TB are in English) took several days. An ML model based on Bag-of-Words, Word vectors or even contextualized embeddings would a) need an annotated dataset and b) likely be much slower.

Figure 2. MULTILEGALPILE Text Type Distribution

Compiling Native MULTILEGALPILE To compile the corpus, we scraped several sources containing legal language materials. Our search was conducted in a loose manner, meaning that when we found a suitable source with legal text data, we included it in our corpus. It is important to note that we do not claim completeness, as we were unable to perform quality analysis for all available languages. For a detailed overview of sources used for the Native MULTILEGALPILE corpus, please refer to Table 9.

The majority of sources provided a link to download the data directly. In cases where data was formatted differently, we converted it into a unified format, such as jsonl. The post-processing steps involved performing various tasks depending on the initial data format. For example, in the case of CASS, we extracted the textual data from XML tags.

Curating Eurlex Resources To curate the Eurlex resources, we utilized the eurlex R package to generate SPARQL queries and download the data. Subsequently, we converted the data into a format more amenable to handling large datasets using Python.

Integrating Pile of Law Henderson et al. (2022) released a large corpus of diverse legal text in English mainly originating from the US. We integrated the latest version with additional data (from January 8, 2023) into our corpus.

3.2. Description

MULTILEGALPILE consists of four large subsets: a) Native Multi Legal Pile (112 GB), b) Eurlex Resources[2] (179 GB), c) Legal MC4[3] (106 GB) and d) Pile of Law (Henderson et al., 2022) (292 GB).

Figure 3 details the distribution of languages. Note that due to the integration of the Pile of Law, English is by far the most dominant language, representing over half of the words. In Figure 2 we show the distribution across text types. Caselaw makes up over half of the corpus, due to the good public access to court rulings especially in common law countries. Note, that even in civil law countries — where legislation is much more important – caselaw is usually more plentiful than legislation (as can be seen in the Swiss case in Table 9). It is hard to find publicly available contracts, leading to the relatively low percentage of the total corpus ($< 10\%$), even though they could potentially make up most of the legal texts in existence (from the private sector). Note that most of the contracts in our corpus are from the US or international treaties with the EU. Table 9 in Appendix C provides additional of the MULTILEGALPILE, including sources and licenses.

3.3. Licenses and Usage of MULTILEGALPILE

The Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license applied for the released MULTILEGALPILE corpus depends on the upstream licenses of the data subsets described above.

First, our *Native Multi Legal Pile* consists of data sources with different licenses. They range from restrictive licenses such as CC BY-NC-SA 4.0 up to the most liberal Creative Commons Zero (CC0) license, which, in essence, releases the data into the public domain. Many sources, however,

[2] https://huggingface.co/datasets/joelito/eurlex_resources
[3] https://huggingface.co/datasets/joelito/legal-mc4

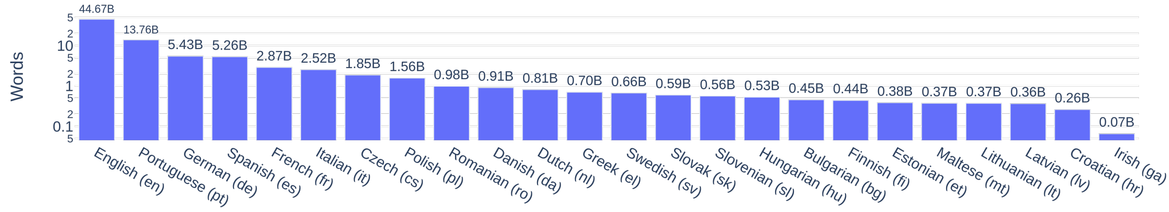


Figure 3. MULTILEGALPILE Language Distribution (Note the log-scaled y-axis)

do not explicitly state the license used for the available data. We assume that such data sources allow pretraining usage, since the creators are usually public agencies such as courts and administrations. Such legislation and caselaw is usually not protected by copyright law. Table 9 provides an overview of the license or copyright situation for each of the 29 sources in the Native Multi Legal Pile.

Second, the *Eurlex Resources* is licensed under CC BY 4.0 by the European Union.[4] Thus, including this corpus does not pose legal issues for pretraining.

Third, the *Legal mC4* corpus was created by filtering multilingual C4 (Xue et al., 2021) for legal content as described above. As mC4 is licensed under ODC-BY, we also release the filtered Legal mC4 corpus under the same license.

Finally, the *Pile of Law* (Henderson* et al., 2022) is published under CC BY-NC-SA 4.0 and the dataset is not altered, therefore the license remains the same.

Usage of the MULTILEGALPILE corpus is presumably possible for pretraining of NLP models. In general, we assume that the fair use doctrine allows employing the data for legal NLP models because the results are rather transformative (Henderson et al., 2023). Nevertheless, copyright issues in generative AI remain an unresolved problem for the moment. Several court cases are currently pending, such as Getty Images suing Stability AI for intellectual property infringement (Sag, 2023).

4. Pretraining Legal Models

As part of this study, we release 2 new multi-lingual legal-oriented PLMs, dubbed Legal-XLM-Rs, trained on the newly introduced MULTILEGALPILE corpus (Section 3). For the newly released Legal-XLM-Rs we followed a series of best-practices in language model development literature:

(a) We warm-start (initialize) our models from the original XLM-R checkpoints (base or large) of Conneau & Lample (2019). Model recycling is a standard process followed by many (Wei et al., 2021; Ouyang et al., 2022) to benefit from starting from an available "well-trained" PLM, rather from scratch (random). XLM-R was trained on 2.5TB of cleaned CommonCrawl data in 100 languages.

[4]EUR-Lex Legal notice

(b) We train a new tokenizer of 128K BPEs on the training subsets of MULTILEGALPILE to better cover legal language across all available legal systems and languages. However, we reuse the original XLM-R embeddings for all lexically overlapping tokens (Pfeiffer et al., 2021), i.e., we warm-start word embeddings for tokens that already exist in the original XLM-R vocabulary, and use random ones for the rest.

(c) We continue pretraining our models on the diverse MULTILEGALPILE corpus with batches of 512 samples for an additional 1M/500K steps for the base/large model. We do initial warm-up steps for the first 5% of the total training steps with a linearly increasing learning rate up to $1e-4$, and then follow a cosine decay scheduling, following recent trends. For half of the warm-up phase (2.5%), the Transformer encoder is frozen, and only the embeddings, shared between input and output (MLM), are updated. We also use an increased 20/30% masking rate for base/large models respectively, where also 100% of the predictions are based on masked tokens, compared to Devlin et al. (2019)[5], based on the findings of Wettig et al. (2023).

(d) For both training the tokenizer and our legal models, we use a sentence sampler with exponential smoothing of the sub-corpora sampling rate following Conneau & Lample (2019) and Raffel et al. (2020b), since there is a disparate proportion of tokens across sub-corpora and languages (Figures 1 and 3) and we aim to preserve per-corpus and language capacity, i.e., avoid overfitting to the majority (approx. 50% of the total number of tokens) US-origin English texts.

(e) We consider mixed cased models, i.e., both upper- and lowercase letters covered, similar to all recently developed large PLMs (Conneau & Lample, 2019; Raffel et al., 2020b; Brown et al., 2020a).

To better account for long contexts often found in legal documents, we continue training the base-size multilingual model on long contexts (4096 tokens) with windowed attention (128 tokens window size) (Beltagy et al., 2020) for 50K steps, dubbing it Legal-XLM-LF-base. We use the standard 15% masking probability and increase the learning rate to $3e-5$ before decaying but otherwise use the same settings as for training the small-context models.

[5]Devlin et al. – and many other follow-up work – used a 15% masking ratio, and a recipe of 80/10/10% of predictions made across masked/randomly-replaced/original tokens.

Model	Source	Params	Vocab	Specs	Corpus	# Langs
MiniLM	Wang et al. (2020)	118M	250K	1M steps / BS 256	2.5TB CC100	100
DistilBERT	Sanh et al. (2020)	135M	120K	BS up to 4000	Wikipedia	104
mDeBERTa-v3	He et al. (2021b;a)	278M	128K	500K steps / BS 8192	2.5TB CC100	100
XLM-R base	Conneau et al. (2020b)	278M	250K	1.5M steps / BS 8192	2.5TB CC100	100
XLM-R large	Conneau et al. (2020b)	560M	250K	1.5M steps / BS 8192	2.5TB CC100	100
Legal-XLM-R-base	ours	184M	128K	1M steps / BS 512	689GB MLP	24
Legal-XLM-R-large	ours	435M	128K	500K steps / BS 512	689GB MLP	24
Legal-XLM-LF-base	ours	208M	128K	50K steps / BS 512	689GB MLP	24
Legal-mono-R-base	ours	111M	32K	200K steps / BS 512	689GB MLP	1
Legal-mono-R-large	ours	337M	32K	500K steps / BS 512	689GB MLP	1

Table 3. Models: All models can process up to 512 tokens, except Legal-XLM-LF-base which can process up to 4096 tokens. BS is short for batch size. MLP is short for MULTILEGALPILE. Params is the total parameter count (including the embedding layer).

In addition to the multilingual models, we also train 24 monolingual models on each of the language-specific subsets of the corpus. Except for choosing a smaller vocab size of 32K tokens, we use the same settings as for the multilingual models. Due to resource constraints, we only train base-size models and stop training at 200K steps. Due to limited data available in some low-resource languages, these models sometimes do multiple passes over the data. Because of plenty of data and to achieve a better comparison on LexGLUE, we continued training the English model for 1M steps and also trained a large-size model for 500K steps. See Table 7 in appendix A for an overview.

We make all our models publicly available alongside all intermediate checkpoints (every 50K/10K training steps for RoBERTa/Longformer models) on the Hugging Face Hub.[6]

5. Evaluating on LEXTREME and LexGLUE

5.1. Benchmark Description

Below we briefly describe each dataset. We refer the interested reader to the original papers for more details.

LEXTREME (Niklaus et al., 2023) is a multilingual legal benchmark. It includes five single label text classification datasets, three multi label text classification datasets and four Named Entity Recognition (NER) datasets.

Brazilian Court Decisions (BCD) (Lage-Freitas et al., 2022) is from the State Supreme Court of Alagoas (Brazil) and involves predicting case outcomes and judges' unanimity on decisions. **German Argument Mining (GAM)** (Urchs et al., 2021) contains200 German court decisions for classifying sentences according to their argumentative function. **Greek Legal Code (GLC)** (Papaloukas et al., 2021) tackles topic classification of Greek legislation documents. Tasks involve predicting topic categories at volume, chapter, and subject levels. **Swiss Judgment Prediction (SJP)** (Niklaus et al., 2021) focuses on predicting the judgment outcome from 85K cases from the Swiss Federal Supreme Court. **Online Terms of Service (OTS)** (Drawzeski et al., 2021) contains 100 contracts for detecting unfair clauses with the tasks of classifying sentence unfairness levels and identifying clause topics. **COVID19 Emergency Event (C19)** (Tziafas et al., 2021): consists of legal documents from several European countries related to COVID-19 measures where models identify the type of measure described in a sentence. **MultiEURLEX (MEU)** (Chalkidis et al., 2021b) is a corpus of 65K EU laws annotated with EUROVOC taxonomy labels. Task involves identifying labels for each document. **Greek Legal NER (GLN)** (Angelidis et al., 2018) is a dataset for NER in Greek legal documents. **LegalNERo (LNR)** (Pais et al., 2021) tackles NER in Romanian legal documents. **LeNER BR (LNB)** (Luz de Araujo et al., 2018) addresses NER in Brazilian legal documents. **MAPA (MAP)** (Baisa et al., 2016) is a multilingual corpus based on EUR-Lex for NER annotated at a coarse-grained and fine-grained level.

LexGLUE (Chalkidis et al., 2021d) is a legal benchmark covering two single label text classification datasets, four multi label text classification datasets and a multiple choice question answering dataset.

ECtHR Tasks A & B (Chalkidis et al., 2019a; 2021c) contain approx. 11K cases from the European Court of Human Rights (ECtHR) public database. Based on case facts, Task A involves predicting violated articles of the European Convention of Human Rights (ECHR) and Task B involves predicting allegedly violated articles. **SCOTUS** (Spaeth et al.) combines information from US Supreme Court (SCOTUS) opinions with the Supreme Court DataBase (SCDB). The task is to classify court opinions into 14 issue areas. **EUR-LEX** (Chalkidis et al., 2021a) contains 65K EU laws from the EUR-Lex portal, annotated with EuroVoc concepts. The task is to predict EuroVoc labels for a given document. **LEDGAR** (Tuggener et al., 2020) contains approx. 850K contract provisions from the US Securities and Exchange Commission (SEC) filings. The task is to classify contract provisions into categories. **UNFAIR-ToS** (Lippi et al., 2019)

[6]https://huggingface.co/joelito

Model	BCD	GAM	GLC	SJP	OTS	C19	MEU	GLN	LNR	LNB	MAP	Agg.
MiniLM	53.0	73.3	42.1	67.7	44.1	5.0	29.7	74.0	84.5	93.6	57.8	56.8
DistilBERT	54.5	69.5	62.8	66.8	56.1	25.9	36.4	71.0	85.3	89.6	60.8	61.7
mDeBERTa-v3	60.2	71.3	52.2	69.1	66.5	29.7	37.4	73.3	85.1	94.8	67.2	64.3
XLM-R-base	63.5	72.0	57.4	69.3	67.8	26.4	33.3	**74.6**	**85.8**	94.1	62.0	64.2
XLM-R-large	58.7	73.1	57.4	69.0	**75.0**	29.0	**42.2**	74.1	85.0	**95.3**	68.0	66.1
Legal-XLM-R-base	62.5	72.4	68.9	70.2	70.8	30.7	38.6	73.6	84.1	94.1	**69.2**	66.8
Legal-XLM-R-large	63.3	73.9	59.3	70.1	74.9	**34.6**	39.7	73.1	83.9	94.6	67.3	66.8
Legal-XLM-LF-base	**72.4**	**74.6**	**70.2**	**72.9**	69.8	26.3	33.1	72.1	84.7	93.3	66.2	**66.9**

Table 4. Dataset aggregate scores for multilingual models on LEXTREME. We report macro-F1 and the best scores in bold.

Model	bg	cs	da	de	el	en	es	et	fi	fr	ga	hr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv	Agg.
MiniLM	52.7	48.6	42.8	54.6	50.3	34.3	40.1	46.3	42.2	39.0	42.8	29.7	29.6	40.5	44.2	40.8	40.8	29.5	22.7	61.6	59.6	44.3	30.0	43.4	40.5
DistilBERT	54.2	48.6	46.0	60.1	58.8	48.0	50.0	48.8	49.6	47.9	51.4	35.9	31.2	50.1	51.9	41.5	44.4	34.6	34.5	63.2	63.8	51.3	36.2	50.1	46.7
mDeBERTa-v3	54.1	51.3	51.7	63.6	57.7	50.7	53.3	50.8	54.6	49.2	54.9	37.4	37.5	55.1	53.9	47.0	52.5	42.1	41.0	65.7	65.3	55.4	37.5	56.1	50.5
XLM-R-base	56.4	48.3	48.3	60.6	57.6	50.1	47.2	46.7	48.6	49.4	50.1	33.6	32.8	53.4	50.0	44.1	43.8	35.2	41.3	66.1	63.7	45.3	33.7	50.0	47.1
XLM-R-large	**59.9**	56.0	**56.3**	65.4	60.8	56.2	**56.6**	56.5	**56.9**	51.4	55.4	42.5	38.1	**58.5**	58.1	49.9	53.9	39.5	**46.4**	**68.6**	**66.8**	**57.9**	42.4	**59.1**	**53.7**
Legal-XLM-R-base	55.6	**58.8**	50.4	63.6	**63.7**	66.8	56.3	**57.0**	52.6	50.1	56.6	38.7	**56.5**	56.1	57.2	49.1	56.0	41.6	43.9	68.2	66.1	55.6	38.6	54.9	53.5
Legal-XLM-R-large	57.8	55.6	50.4	**65.7**	60.7	**69.3**	55.7	54.5	56.6	**53.3**	**57.2**	39.7	39.1	58.1	**60.6**	48.4	57.2	39.4	45.5	67.3	65.5	49.3	39.7	56.4	53.6
Legal-XLM-LF-base	54.4	49.3	48.1	64.0	60.5	52.8	49.2	52.2	48.2	48.5	55.4	33.0	34.7	54.6	54.8	45.2	52.5	40.1	40.6	68.3	64.1	48.4	33.0	51.3	48.9
NativeLegalBERT	-	-	-	-	-	53.1	46.9	-	-	-	-	-	-	45.3	-	-	-	-	-	-	59.0	-	-	-	51.1
NativeBERT	54.8	57.3	51.2	63.0	62.3	52.0	42.6	47.2	52.4	49.4	50.1	-	37.4	47.1	-	-	-	37.0	40.5	66.5	63.1	44.8	-	55.1	50.2
Legal-mono-R-base	55.9	49.5	51.5	61.3	61.3	50.5	52.1	53.5	53.6	51.1	52.2	**44.1**	54.1	51.8	55.5	**50.0**	**59.1**	**54.3**	34.4	67.1	61.5	48.8	**53.4**	58	53.5

Table 5. Language aggregate scores on LEXTREME. We report macro-F1 and best scores in bold. For each language, we also list the best-performing monolingual legal model under *NativeLegalBERT*, the best-performing monolingual non-legal model under *NativeBERT* and our monolingual legal models under *Legal-mono-R-base*. Missing values indicate that no suitable models were found.

contains 50 Terms of Service (ToS) from online platforms, annotated with types of unfair contractual terms. The task is to predict unfair types for a given sentence. **CaseHOLD** (Zheng et al., 2021) contains approx. 53K multiple choice questions about holdings of US court cases. The task is to identify the correct holding statement from a selection of five choices.

5.2. Experimental Setup

To ensure comparability, we followed the experimental setups described in the original papers (Niklaus et al., 2023; Chalkidis et al., 2021d) using hierarchical transformers for datasets where the sequence length of most documents exceeds the maximum sequence length of the model (Aletras et al., 2016; Niklaus et al., 2022). The hyperparameters used for running experiments on each dataset are provided in Table 8 in the appendix. To obtain Table 6, we followed Chalkidis et al. (2021d), running five repetitions with different random seeds (1-5) and reporting the test scores based on the seed that yielded the best scores on the development data. For values in Tables 4 and 5, we followed the procedure in Niklaus et al. (2023), taking the mean of the results of 3 random seeds (1-3). We show an overview of the evaluated models in Table 3.

5.3. Evaluation on LEXTREME

We evaluate our models on LEXTREME (Niklaus et al., 2023) and show results across datasets in Table 4 and across languages in Table 5.

We notice that our Legal-XLM-R-base model is on par with XLM-R large even though it only contains 33% of the parameters (184M vs 560M). All our models outperform XLM-R large on the dataset aggregate score. Our base model sets a new SotA on MAPA (MAP), the large model on CoViD 19 emergency event (C19) and the Longformer on Brazilian court decisions (BCD), German argument mining (GAM), Greek legal code (GLC) and Swiss judgment prediction (SJP). Surprisingly, the legal models slightly underperform in three NER tasks (GLN, LNR, and LNB). Sensitivity to hyperparameter choice could be a reason for this underperformance (we used the same hyperparameters for all models without tuning due to limited compute resources). We see the largest improvements over prior art in Brazilian court decisions (72.4 vs. 63.5) and in Greek legal code (70.2 vs 62.8). Maybe these tasks are particularly hard and therefore legal in-domain pretraining helps more. For BCD especially, the large amount of Brazilian caselaw in the pretraining corpus may offer an additional explanation.

The monolingual models underperform their base model XLM-R base only in Italian, Polish, and Romanian. In some languages the monolingual model even outperforms XLM-R base clearly (Croatian, Hungarian, Latvian, Maltese, Dutch, Slovakian, and Swedish), and in five of them even set the new SotA for the language, sometimes clearly outperforming all other models (the Dutch model even outperforms its closest competitor mDeBERTa-v2 by 11.2 macro F1 and

Model	ECtHR-A	ECtHR-B	SCOTUS	EUR-LEX	LEDGAR	UNFAIR-ToS	CaseHOLD	Agg.
TFIDF+SVM *	48.9	63.8	64.4	47.9	81.4	75.0	22.4	49.0
BERT *	63.6	73.4	58.3	57.2	81.8	81.3	70.8	68.2
DeBERTa *	60.8	71.0	62.7	57.4	83.1	80.3	72.6	68.5
RoBERTa-base *	59.0	68.9	62.0	57.9	82.3	79.2	71.4	67.5
RoBERTa-large *	67.6	71.6	66.3	58.1	**83.6**	81.6	74.4	70.9
Longformer *	64.7	71.7	64.0	57.7	83.0	80.9	71.9	69.5
BigBird *	62.9	70.9	62.0	56.8	82.6	81.3	70.8	68.4
Legal-BERT *	64.0	74.7	66.5	57.4	83.0	**83.0**	75.3	70.8
CaseLaw-BERT *	62.9	70.3	65.9	56.6	83.0	82.3	75.4	69.7
Legal-en-R-base (ours)	65.2	73.7	66.4	**59.2**	82.7	78.7	73.3	70.5
Legal-en-R-large (ours)	**70.3**	**77.0**	**67.7**	58.4	82.5	82.4	**77.0**	**72.7**
Legal-XLM-R-base (ours)	64.8	73.9	63.9	58.2	82.8	79.6	71.7	69.7
Legal-XLM-R-large (ours)	68.2	74.2	67.5	58.4	82.7	79.9	75.1	71.4
Legal-XLM-LF-base (ours)	67.9	76.2	61.6	59.1	82.1	78.9	72.0	70.2

Table 6. Results on LexGLUE. We report macro-F1 and best scores in bold. Results from models marked with * are from Chalkidis et al. (2021d). Similar to LEXTREME, we computed the aggregate score as the harmonic mean of individual dataset results.

its base model XLM-R by almost 20 macro F1). These languages are all in the lower end of the data availability in the MULTILEGALPILE with the richest language (Dutch) containing only 810M words (see Figure 3). Pretraining a monolingual model on in-domain data may therefore be worth it, especially in low-resource languages.

Even though our legal Longformer model performs best on the dataset level, it performs much worse on the language level, possibly due to its lower scores in the most multilingual tasks MEU, MAP and C19 (24, 24 and 6 languages, respectively). Our legal base and large models achieve SotA in some languages, and are on aggregate almost as robust across languages as XLM-R.

Computing the final LEXTREME scores (harmonic mean of dataset aggregate and language aggregate scores), we find that the Legal-XLM-R-large is the new SotA on LEXTREME with a score of 59.5 vs 59.4 for Legal-XLM-R-base and 59.3 for XLM-R large. The legal Longformer's LEXTREME scores is with 56.5 not competitive due to its low language aggregate score.

5.4. Evaluation on LexGLUE

We evaluate our English and multilingual models on LexGLUE (Chalkidis et al., 2021e) and compare against baselines (see Table 6). Our models excel on the ECtHR, SCOTUS, EUR-LEX, and CaseHOLD tasks, achieving new SotA. In the other two tasks our models match general-purpose models such as RoBERTa. A reason for slight underperformance of the legal models in the LEDGAR and especially the Unfair ToS tasks may be the relatively low availability of contracts in the MULTILEGALPILE.

6. Conclusions and Future Work

Limitations We did not perform deduplication, thus data from the legal mC4 part might be present in other parts. However, recent work (Muennighoff et al., 2023) suggests that data duplication does not degrade performance during pretraining for up to four epochs. Overlap between the other parts is highly unlikely, since they are from completely different jurisdictions.

Conclusions Due to a general lack of multilingual pretraining data especially in specialized domains such as law, we curate a large-scale high-quality corpus in 24 languages from 17 jurisdictions. We continue pretraining XLM-R checkpoints on our data, achieving a new SotA for base and large models on the LEXTREME benchmark and vastly outperforming previous methods in greek legal code. We turn our XLM-R base model into a Longformer and continue pretraining on long documents. It reaches a new SotA in four LEXTREME datasets and reaches the overall highest dataset aggregate score. Monolingual models achieve huge gains over their base model XLM-R in some languages and even set language specific SotA in five languages outperforming other models by as much as 11 macro F1. On LexGLUE our English models reach SotA in five out of seven tasks with the large model achieving the highest aggregate score.

Future Work We leave the pretraining of a large generative multilingual legal language model for future work. Here we limited the corpus to the EU languages due to resource constraints, but in the future, we would like to expand the corpus in terms of languages and jurisdictions covered. Especially in China there exist many accessible sources suitable to extend the corpus. Finally, it would be very interesting to study in more detail the specific contents of the MULTILEGALPILE.

References

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., and Lampos, V. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93, October 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.93. URL https://peerj.com/articles/cs-93. Publisher: PeerJ Inc.

Angelidis, I., Chalkidis, I., and Koubarakis, M. Named entity recognition, linking and generation for greek legislation. 2018.

Baisa, V., Michelfeit, J., Medveď, M., and Jakubíček, M. European Union language resources in Sketch Engine. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2799–2803, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1445.

Beltagy, I., Lo, K., and Cohan, A. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*, 2019.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*, December 2020. URL http://arxiv.org/abs/2004.05150. arXiv: 2004.05150.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T. J., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020b.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4317–4323, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1424. URL https://aclanthology.org/P19-1424.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4317–4323, Florence, Italy, July 2019b. Association for Computational Linguistics. doi: 10.18653/v1/P19-1424. URL https://www.aclweb.org/anthology/P19-1424.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2898–2904, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.261. URL https://aclanthology.org/2020.findings-emnlp.261.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. LEGAL-BERT: The Muppets straight out of Law School. *arXiv:2010.02559 [cs]*, October 2020b. URL http://arxiv.org/abs/2010.02559. arXiv: 2010.02559.

Chalkidis, I., Fergadiotis, M., and Androutsopoulos, I. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *EMNLP*, 2021a.

Chalkidis, I., Fergadiotis, M., and Androutsopoulos, I. MultiEURLEX – A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv:2109.00904 [cs]*, September 2021b. URL http://arxiv.org/abs/2109.00904. arXiv: 2109.00904.

Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., and Malakasiotis, P. Paragraph-level rationale extraction through regularization: A case study on European court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 226–241, Online, June 2021c. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.22. URL https://aclanthology.org/2021.naacl-main.22.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., and Aletras, N. Lexglue:

A benchmark dataset for legal language understanding in english, 2021d. URL https://arxiv.org/abs/2110.00976.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M. J., Androutsopoulos, I., Katz, D. M., and Aletras, N. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. SSRN Scholarly Paper ID 3936759, Social Science Research Network, Rochester, NY, October 2021e. URL https://papers.ssrn.com/abstract=3936759.

Chalkidis*, I., Garneau*, N., Goanta, C., Katz, D. M., and Søgaard, A. Lexfiles and legallama: Facilitating english multinational legal language model development, 2023.

Conneau, A. and Lample, G. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*, April 2020b. URL http://arxiv.org/abs/1911.02116. arXiv: 1911.02116.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Drawzeski, K., Galassi, A., Jablonowska, A., Lagioia, F., Lippi, M., Micklitz, H. W., Sartor, G., Tagiuri, G., and Torroni, P. A Corpus for Multilingual Analysis of Online Terms of Service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 1–8, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.nllp-1.1.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23:1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv:2101.00027 [cs]*, December 2020a. URL http://arxiv.org/abs/2101.00027. arXiv: 2101.00027.

Gao, L., Biderman, S. R., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling. *ArXiv*, abs/2101.00027, 2020b.

Gokaslan, A. and Cohen, V. Openwebtext corpus, 2019. URL http://Skylion007.github.io/OpenWebTextCorpus.

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare*, 3 (1), oct 2021. ISSN 2691-1957. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Gonzalez-Agirre, A., and Villegas, M. Spanish Legalese Language Model and Corpora. oct 2021. URL http://arxiv.org/abs/2110.12201.

He, P., Gao, J., and Chen, W. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543 [cs]*, December 2021a. URL http://arxiv.org/abs/2111.09543. arXiv: 2111.09543.

He, P., Liu, X., Gao, J., and Chen, W. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*, October 2021b. URL http://arxiv.org/abs/2006.03654. arXiv: 2006.03654.

Henderson, P., Krass, M. S., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D., and Ho, D. E. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset, July 2022. URL http://arxiv.org/abs/2207.00220. arXiv:2207.00220 [cs].

Henderson*, P., Krass*, M. S., Zheng, L., Guha, N., Manning, C. D., Jurafsky, D., and Ho, D. E. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset, 2022. URL https://arxiv.org/abs/2207.00220.

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A., and Liang, P. Foundation models and fair use, 2023.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding, January 2021. URL http://arxiv.org/abs/2009.03300. arXiv:2009.03300 [cs].

Huang, K., Altosaar, J., and Ranganath, R. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. 2019. URL http://arxiv.org/abs/1904.05342.

ImaniGooghari, A., Lin, P., Kargaran, A. H., Severini, S., Sabet, M. J., Kassner, N., Ma, C., Schmid, H., Martins, A. F. T., Yvon, F., and Schütze, H. Glot500: Scaling multilingual corpora and language models to 500 languages, 2023.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL https://doi.org/10.1038/sdata.2016.35.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. GPT-4 Passes the Bar Exam, March 2023. URL https://papers.ssrn.com/abstract=4389233.

Lage-Freitas, A., Allende-Cid, H., Santana, O., and Oliveira-Lage, L. Predicting Brazilian Court Decisions. *PeerJ Computer Science*, 8:e904, March 2022. ISSN 2376-5992. doi: 10.7717/peerj-cs.904. URL https://peerj.com/articles/cs-904. Publisher: PeerJ Inc.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, pp. btz682, September 2019. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btz682. URL http://arxiv.org/abs/1901.08746. arXiv: 1901.08746.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. ISSN 14602059. doi: 10.1093/bioinformatics/btz682.

Licari, D. and Comandè, G. ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law. Technical report, 2022. URL http://ceur-ws.org.

Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., and Torroni, P. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2):117–139, 2019. ISSN 1572-8382. doi: 10.1007/s10506-019-09243-2. URL https://doi.org/10.1007/s10506-019-09243-2.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. 2019. URL http://arxiv.org/abs/1907.11692.

Luz de Araujo, P. H., Campos, T. E. d., de Oliveira, R. R., Stauffer, M., Couto, S., and Bermejo, P. Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on Computational Processing of the Portuguese Language*, pp. 313–323. Springer, 2018. Dataset URL: https://huggingface.co/datasets/lener_br.

Masala, M., Iacob, R. C. A., Uban, A. S., Cidota, M., Velicu, H., Rebedea, T., and Popescu, M. jurBERT: A Romanian BERT model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 86–94, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nllp-1.8. URL https://aclanthology.org/2021.nllp-1.8.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models, 2016.

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., and Raffel, C. Scaling Data-Constrained Language Models, May 2023. URL http://arxiv.org/abs/2305.16264. arXiv:2305.16264 [cs].

Niklaus, J. and Giofré, D. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch?, November 2022. URL http://arxiv.org/abs/2211.17135. arXiv:2211.17135 [cs].

Niklaus, J., Chalkidis, I., and Stürmer, M. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 19–35, Punta

Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.nllp-1.3.

Niklaus, J., Stürmer, M., and Chalkidis, I. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 32–46, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.3.

Niklaus, J., Matoshi, V., Rani, P., Galassi, A., Stürmer, M., and Chalkidis, I. Lextreme: A multi-lingual and multi-task benchmark for the legal domain, 2023. URL https://arxiv.org/abs/2301.13126.

OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Pais, V., Mitrofan, M., Gasan, C. L., Coneschi, V., and Ianov, A. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 9–18, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nllp-1.2. URL https://aclanthology.org/2021.nllp-1.2.

Papaloukas, C., Chalkidis, I., Athinaios, K., Pantazi, D.-A., and Koubarakis, M. Multi-granular legal topic classification on greek legislation. *arXiv preprint arXiv:2109.15298*, 2021. Dataset URL: https://huggingface.co/datasets/greek_legal_code.

Pfeiffer, J., Vulić, I., Gurevych, I., and Ruder, S. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL https://aclanthology.org/2021.emnlp-main.800.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. 2019.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67, 2020a. URL http://jmlr.org/papers/v21/20-074.html.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020b. ISSN 1533-7928. URL http://jmlr.org/papers/v21/20-074.html.

Sag, M. Copyright safety for generative ai. *Forthcoming in the Houston Law Review*, 2023. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, February 2020. URL http://arxiv.org/abs/1910.01108. arXiv:1910.01108.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020.

Spaeth, H. J., Epstein, L., Martin, A. D., Segal, J. A., Ruger, T. J., and Benesh, S. C. Supreme Court Database, Version 2020 Release 01.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A Large Language Model for Science, November 2022. URL http://arxiv.org/abs/2211.09085. arXiv:2211.09085 [cs, stat].

Tiedemann, J. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 3518–3522, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1559.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. LLaMA: Open and Efficient Foundation Language Models. *ArXiv*, abs/2302.1, 2023.

Tuggener, D., von Däniken, P., Peetz, T., and Cieliebak, M. LEDGAR: A large-scale multi-label corpus for text classification of legal provisions in contracts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1235–1241, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.155.

Tziafas, G., de Saint-Phalle, E., de Vries, W., Egger, C., and Caselli, T. A multilingual approach to identify and classify exceptional measures against covid-19. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 46–62, 2021. Dataset URL: https://tinyurl.com/ycysvtbm.

Urchs, S., Mitrović, J., and Granitzer, M. Design and Implementation of German Legal Decision Corpora:. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, pp. 515–521, Online Streaming, — Select a Country —, 2021. SCITEPRESS - Science and Technology Publications. ISBN 978-989-758-484-8. doi: 10.5220/0010187305150521. URL https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010187305150521.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. pp. 30, 2019.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. URL https://arxiv.org/abs/2109.01652.

Wettig, A., Gao, T., Zhong, Z., and Chen, D. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2985–3000, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.217.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*, March 2021. URL http://arxiv.org/abs/2010.11934. arXiv: 2010.11934.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. *Defending against Neural Fake News*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and Ho, D. E. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21, pp. 159–168, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385268. doi: 10.1145/3462757.3466088. URL https://doi.org/10.1145/3462757.3466088.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books, 2015.

Model Name	# Steps	Vocab Size
Legal-bg-R-base	200K	32K
Legal-hr-R-base	200K	32K
Legal-cs-R-base	200K	32K
Legal-da-R-base	200K	32K
Legal-nl-R-base	200K	32K
Legal-en-R-base	200K	32K
Legal-en-R-large	500K	32K
Legal-et-R-base	200K	32K
Legal-fi-R-base	200K	32K
Legal-fr-R-base	200K	32K
Legal-de-R-base	200K	32K
Legal-el-R-base	200K	32K
Legal-hu-R-base	200K	32K
Legal-ga-R-base	200K	32K
Legal-it-R-base	200K	32K
Legal-lv-R-base	200K	32K
Legal-lt-R-base	200K	32K
Legal-mt-R-base	200K	32K
Legal-pl-R-base	200K	32K
Legal-pt-R-base	200K	32K
Legal-ro-R-base	200K	32K
Legal-sk-R-base	200K	32K
Legal-sl-R-base	200K	32K
Legal-es-R-base	200K	32K
Legal-sv-R-base	200K	32K
Legal-XLM-R-base	1M	128K
Legal-XLM-R-large	500K	128K
Legal-XLM-LF-base	50K	128K

Table 7. Model Details

A. Training Details

B. Hyperparameter Details

source	Dataset	Task	Task type	Hierarchical	Seeds	lower case	Batch size	Metric for best model	Evaluation strategy	Epochs	Early stopping patience	Learning rate
(Niklaus et al., 2023)	GLN	GLN	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	LNR	LNR	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	LNB	LNB	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	MAP	MAP-F	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	MAP	MAP-C	NER	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	BCD	BCD-J	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	BCD	BCD-U	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	GAM	GAM	SLTC	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	GLC	GLC-C	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	GLC	GLC-S	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	GLC	GLC-V	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	SJP	SJP	SLTC	True	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	OTS	OTS-UL	SLTC	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	OTS	OTS-CT	MLTC	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	C19	C19	MLTC	False	1,2,3	True	64	evaluation loss	epoch	50	5	1e-5
(Niklaus et al., 2023)	MEU	MEU-1	MLTC	True	1,2,3	True	64	evaluation loss			5	1e-5
(Niklaus et al., 2023)	MEU	MEU-2	MLTC	True	1,2,3	True	64	evaluation loss			5	1e-5
(Niklaus et al., 2023)	MEU	MEU-3	MLTC	True	1,2,3	True	64	evaluation loss			5	1e-5
(Chalkidis et al., 2021d)	ECtHR	ECtHR-A	MLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	ECtHR	ECtHR-B	MLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	EUR-LEX	EUR-LEX	MLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	SCOTUS	SCOTUS	SLTC	True	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	LEDGAR	LEDGAR	SLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	UnfairToS	UnfairToS	MLTC	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5
(Chalkidis et al., 2021d)	CaseHOLD	CaseHOLD	MCQA	False	1,2,3,4,5	True	8	micro-f1	epoch	20	3	3e-5

Table 8. Hyperparameters for each dataset and task. However, there were a few exceptions. For the multilingual MEU tasks, given the dataset's size, we trained them for only 1 epoch with 1000 steps as the evaluation strategy when using multilingual models. When using monolingual models, we trained for 50 epochs with epoch-based evaluation strategy, as we utilized only the language-specific subset of the dataset. Regarding LexGlue, we followed the guidelines of Chalkidis et al. (2021d) for RoBERTa-based large language models, which required a maximum learning rate of 1e-5, a warm-up ratio of 0.1, and a weight decay rate of 0.06.

C. Dataset Details

Language	Text Type	Words	Documents	Words per Document	Jurisdiction	Source	License/Copyright
Native Multi Legal Pile							
bg	legislation	309M	262k	1178	Bulgaria	MARCELL	CC0-1.0
cs	caselaw	571M	342k	1667	Czechia Czechia Czechia	CzCDC Constitutional Court CzCDC Supreme Administrative Court CzCDC Supreme Court	CC BY-NC 4.0 CC BY-NC 4.0 CC BY-NC 4.0
da	caselaw	211M	92k	2275	Denmark	DDSC	CC BY 4.0 and other, depending on the dataset
da	legislation	653M	296k	2201	Denmark	DDSC	CC BY 4.0 and other, depending on the dataset
de	caselaw	1786M	614k	2905	Germany Switzerland	openlegaldata entscheidsuche	ODbL-1.0 similar to CC BY
de	legislation	513M	302k	1698	Germany Switzerland	openlegaldata lexfind	ODbL-1.0 not protected by copyright law
en	legislation	2539M	713k	3557	Switzerland UK	lexfind uk-lex	not protected by copyright law CC BY 4.0
fr	caselaw	1172M	495k	2363	Belgium France Luxembourg Switzerland	jurportal CASS judoc entscheidsuche	not protected by copyright law Open Licence 2.0 not protected by copyright law similar to CC BY
fr	legislation	600M	253k	2365	Switzerland Belgium	lexfind ejustice	not protected by copyright law not protected by copyright law
hu	legislation	265M	259k	1019	Hungary	MARCELL	CC0-1.0
it	caselaw	407M	159k	2554	Switzerland	entscheidsuche	similar to CC BY
it	legislation	543M	238k	2278	Switzerland	lexfind	not protected by copyright law
nl	legislation	551M	243k	2263	Belgium	ejustice	not protected by copyright law
pl	legislation	299M	260k	1148	Poland	MARCELL	CC0-1.0
pt	caselaw	12613M	17M	728	Brazil Brazil Brazil	RulingBR CRETA CJPG	not protected by copyright law CC BY-NC-SA 4.0 not protected by copyright law
ro	legislation	559M	396k	1410	Romania	MARCELL	CC0-1.0
sk	legislation	280M	246k	1137	Slovakia	MARCELL	CC0-1.0
sl	legislation	366M	257k	1418	Slovenia	MARCELL	CC-BY-4.0
total		**24236M**	**23M**	**1065**		**Native Multi Legal Pile**	
Overall statistics for the remaining subsets							
total		**12107M**	**8M**	**1457**	EU	**Eurlex Resources**	**CC BY 4.0**
total		**43376M**	**18M**	**2454**	US (99%), Canada, and EU	**Pile of Law**	CC BY-NC-SA 4.0; **See Henderson* et al. (2022) for details**
total		**28599M**	**10M**	**2454**		**Legal mC4**	**ODC-BY**

Table 9. Information about size and number of words and documents for *Native* Multi Legal Pile are provided according to language and text type. For the remaining subsets of Multi Legal Pile we provide general statistics.

⚖️ *SCALE*: *S*caling up the *C*omplexity for *A*dvanced *L*anguage Model *E*valuation

Vishvaksenan Rasiah[1*] **Ronja Stern**[1*] **Veton Matoshi**[2]

Matthias Stürmer[1,2] **Ilias Chalkidis**[3] **Daniel E. Ho**[4]

Joel Niklaus[1,2,4*]

[1]University of Bern [2]Bern University of Applied Sciences
[3]University of Copenhagen [4]Stanford University

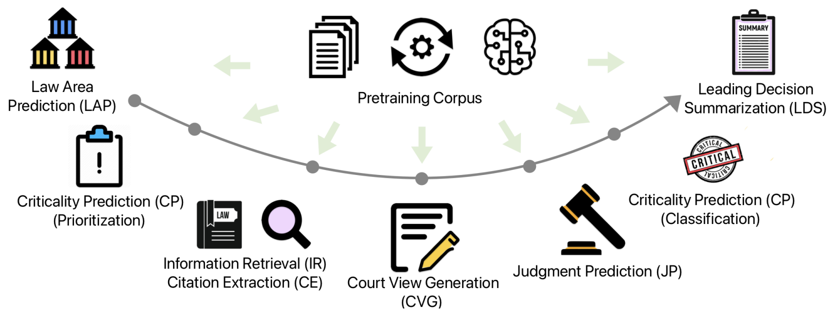


Figure 1: Sequence of tasks for support in the judicial system.

Abstract

Recent strides in Large Language Models (LLMs) have saturated many NLP benchmarks (even professional domain-specific ones), emphasizing the need for more challenging ones to properly assess LLM capabilities. In this paper, we introduce a novel NLP benchmark that poses challenges to current LLMs across four key dimensions: processing *long documents* (up to 50K tokens), utilizing *domain specific knowledge* (embodied in legal texts), *multilingual* understanding (covering five languages), and *multitasking* (comprising legal document to document Information Retrieval, Court View Generation, Leading Decision Summarization, Citation Extraction, and eight challenging Text Classification tasks). Our benchmark comprises diverse legal NLP datasets from the Swiss legal system, allowing for a comprehensive study of the underlying Non-English, inherently multilingual, federal legal system. Despite recent advances, efficiently processing long documents for intense review/analysis tasks remains an open challenge for LLMs. Also, comprehensive, domain-specific benchmarks requiring high expertise to develop are rare, as are multilingual benchmarks. This scarcity underscores our contribution's value, considering most public models are trained predominantly on English corpora, while other languages remain understudied, particularly for practical domain-specific NLP tasks. Our benchmark allows for testing and advancing the state-of-the-art LLMs. As part of our study, we evaluate several pre-trained multilingual language models on our benchmark to establish strong baselines as a point of reference. Despite the large size of our datasets (tens to hundreds of thousands of examples),

* Equal contribution.

Preprint. Under review.

existing publicly available models struggle with most tasks, even after extensive in-domain pretraining. We publish all resources (benchmark suite, pre-trained models, code) under a fully permissive open CC BY-SA license.

1 Introduction

The history of legal Natural Language Processing (NLP) is extensive Ashley [2017], with remarkable progress recently Katz et al. [2023a]. Notably, the introduction of datasets containing legal data from various jurisdictions worldwide Paul et al. [2021], Chalkidis et al. [2019b], as well as the development of more domain-specific tasks and benchmarks Hendrycks et al. [2021b], Li and Zhang [2021a], Semo et al. [2022], Brugger et al. [2023], Hwang et al. [2022], Niklaus et al. [2023a], Thakur et al. [2021], Chen et al. [2022], Guha et al. [2022] have significantly contributed to the progress in the field. General benchmarks such as SuperGLUE Wang et al. [2019] are saturated and ineffective at differentiating Large Language Models (LLMs). Hence, larger, challenging benchmarks are urgently needed, especially in the domain-specific context. In the context of Switzerland, the availability of only one dataset for evaluating LLMs hampers the assessment of their performance and effectiveness within the country's diverse linguistic and legal landscape Niklaus et al. [2021, 2022]. In this paper, we introduce seven related datasets covering a range of tasks and spanning across five languages within the same overarching jurisdiction. These datasets are derived from 26 cantons and the Swiss Federal Supreme Court (FSCS), each with distinct legal frameworks, in the uniquely multilingual and multi-jurisdictional context of Switzerland. The country's multiple official languages and a wealth of data for its size, position Switzerland as an exemplary testbed for assessing LLMs in a multilingual and multi-jurisdictional environment. Our assessment concentrates on three classification tasks – Criticality Prediction (CP), Judgment Prediction (JP), and Law Area Prediction (LAP) – an Information Retrieval (IR) task and two generative tasks – Court View Generation (CVG) and Leading Decision Summarization (LDS). To facilitate a comprehensive analysis and provide baselines for future research, we evaluate an array of models on our datasets similar to Hwang et al. [2022] or Niklaus et al. [2023a]. Furthermore, we have pretrained our own Swiss legal models, Legal Swiss RoBERTa$_{\text{Base/Large}}$ and Legal Swiss Longformer$_{\text{Base}}$. Our tasks challenge current models significantly, with the best performing model only achieving an aggregated Macro F1 score of 48.4. ChatGPT was not able to solve the text classification tasks well, considerably lagging behind fine-tuned models. The results for CP, IR and CVG are particularly underwhelming, seeming rather arbitrary. We invite the research community to develop new methods to tackle these hard tasks. All data employed in this study is in the public domain (see https://entscheidsuche.ch/dataUsage and https://www.fedlex.admin.ch/en/legal-information and is available on the HuggingFace Hub under a CC BY-SA license (https://huggingface.co/rcds).

This paper makes three contributions. First, we present seven public multilingual datasets containing Swiss legal documents. Second, we release two large, in-domain pretraining datasets, and pretrain three new models - Legal-Swiss-RoBERTa$_{\text{Base/Large}}$ and Legal-Swiss-LongFormer$_{\text{base}}$. Third, we evaluate multilingual baselines on our datasets and compare them to our models. Although in-domain pretraining improves performance, significant room for improvement remains in most tasks.

2 Related Work

We briefly discuss prior work on benchmarks for long documents, domain specificity, multilinguality, and multitasking. Additional task-specific related work is presented in Appendix F.

Long Documents SCROLLS consists of summarization, Question Answering (QA), and Natural Language Inference (NLI) tasks with example inputs typically in the thousands of English words [Shaham et al., 2022]. MuLD is a set of six tasks (twice QA, style change detection, classification, summarization, and translation) where each input is at least 10K tokens, with some up to almost 500K tokens [Hudson and Moubayed, 2022].

Domain Specificity The BLUE benchmark Peng et al. [2019a] contains five tasks over ten datasets for biomedical and clinical texts. CBLUE Zhang et al. [2022] is a Chinese biomedical benchmark with eight NLU tasks including Named Entity Recognition (NER), information extraction, diagnosis normalization, Text Classification (TC), QA, intent classification, and semantic similarity. LexGLUE covers six predictive tasks over five datasets made of English documents from the US, EU, and Council of Europe Chalkidis et al. [2022]. LEXTREME is a multi-lingual and multi-task benchmark for the legal domain Niklaus et al. [2023a]. LegalBench Guha et al. [2022] covers zero-shot and few-shot Language Model (LM) evaluation for diverse realistic legal tasks in English. LBOX OPEN Hwang et al. [2022] consists of five legal tasks from South Korea.

Multilinguality XTREME Hu et al. [2020], designed to evaluate cross-lingual generalization, includes six tasks across ten datasets, covering 40 languages. Some datasets were cross-lingual, others were extended via professional and automatic translations. XTREME-UP expands XTREME, emphasizing the evaluation of multilingual models in a few-shot setting for user-centric tasks [Ruder et al., 2023]. It covers 88 under-represented languages such as Swahili, Burmese, or Telugu where only few datasets exist.

Multitasking GLUE Wang et al. [2018], an early benchmark of sentence NLU tasks evaluating general-purpose neural LMs, quickly became obsolete due to advanced models like BERT Devlin et al. [2019a]. Its updated version, SUPERGLUE [Wang et al., 2019], introduced new tasks challenging for machines yet solvable by humans. MMLU features only zero-shot and few-shot learning tasks Hendrycks et al. [2021a], containing about 16K multiple-choice questions divided into 57 subtasks, spanning subjects in the humanities, social and hard sciences, etc. CLUE Xu et al. [2020] is the first Chinese language multitask benchmark that includes single sentence classification, sentence pair classification, and machine reading comprehension. BIG-Bench Srivastava et al. [2022] consists of 204 language tasks created by 450 authors from 132 institutions. The tasks cover topics such as linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development. HELM Liang et al. [2022] is a multi-metric benchmark covering seven metrics and seven targeted evaluations and involves 42 test scenarios with a large-scale evaluation of 30 LMs.

3 Background on the Swiss Legal System

Switzerland comprises 26 cantons, each with unique jurisdiction and court organization. The Swiss Federal Supreme Court (FSCS) is Switzerland's highest legal authority and final arbiter for federal criminal, administrative, patent, and cantonal courts. Its decisions make gaps in legislation explicit, and shape the development of the law and its adaptation to changing circumstances. The FSCS has seven divisions, specializing in public, penal, and civil law [Bundesgericht, 2019]. All cases before the Supreme Court are Federal Supreme Court Decisons (FSCD), but only a few are designated as Leading Decisions (BGE for Bundesgerichtsentscheid in German) and are separately published, influencing future jurisprudence significantly. Cantonal court proceedings begin at the lowest instance and may be appealed higher, with appeal stages varying by canton and legal area. Swiss court decisions typically consist of four major sections: 1) the *rubrum* (introduction) contains the date and chamber, mentions the involved judge(s) and parties and finally states the topic of the decision. 2) the *facts* describe what happened in the case and form the basis for the considerations of the court. The higher the level of appeal, the more general and summarized the facts. 3) the *considerations* reflect the formal legal reasoning, citing laws and other influential rulings, and forming the basis for the final ruling. 4) the *rulings*, the final section, are an enumeration of the binding decisions made by the court. This section is normally rather short and summarizes the considerations.

4 SCALE: The Datasets

Table 1 introduces our eleven datasets. Data were collected from 26 cantons (in addition to federal decisions), 184 courts, 456 chambers, four main law areas, and five languages as seen in Table 18. There are significant differences in the availability of documents across cantons and courts.[2] Most courts are monolingual, but there are cantons where multiple languages are used in documents. In addition to the decision-based datasets, we also provide a collection of approx. 35K laws from cantonal and federal jurisdictions in Switzerland.

[2]The FSCS is the only court where we have complete data, since all decisions since 2007 have been published.

While most FSCD are written in German, French is more common for cantonal cases. We partitioned downstream datasets into training (until 2015), validation (2016-2017), and test (2018-2022) sets. We opted for a relatively large test split, because LLMs seem to need relatively little training data Brown et al. [2020]. A large test set allows longitudinal studies, including COVID-19 pandemic years. This large temporal gap between the newest training (2015) and test (2022) samples might contribute to model difficulties (see section 6). Source data undergo rigorous curation by established Swiss institutions such as courts and administrative bodies, including manual anonymization at an approximate cost of 45 minutes per case Niklaus et al. [2021].

Table 1: Overview over all datasets and their multilingualism: Abbreviations: **Cant**onal, **Fed**eral, **Fac**ts, **Cons**iderations Column Fac and Cons report the mean number of tokens. Sections Facts and Considerations are not available for Ruling Summarization, Legislation and Rulings due to different format, thus mean number of tokens for the full text is reported and marked with *.

Name	Level	Total	DE	FR	IT	RM	EN	Fac	Cons
Rulings	Cant + Fed	638K	320K	247K	71K	-	180	-	*7K
Leading Decisions	Fed	21K	14K	6K	1K	-	-	689	3K
Legislation	Cant + Fed	36K	18K	11K	6K	534	207	-	*7K
Doc2Doc IR	Fed	141K	87K	46K	8K	-	-	847	3K
Citation Extraction	Fed	131K	85K	38K	8K	-	-	-	204
Criticality	Fed	139K	85K	45K	8K	-	-	828	3K
Law Area	Cant + Fed	329K	127K	156K	46K	-	-	2K	4K
Judgment Prediction	Cant + Fed	329K	160K	128K	41K	-	-	2K	4K
Court View	Cant + Fed	404K	197K	163K	44K	-	-	2K	5K
Court View Origin	Fed	270	49	221	-	-	-	1K	6K
Leading Decision Summarization	Fed	18K	12K	5K	835	-	-	-	*3K

4.1 Database Creation Pipeline

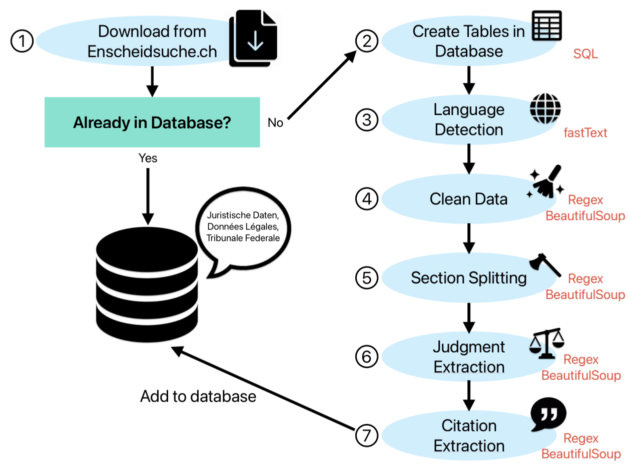


Figure 2: Database Creation Pipeline

Every day, new cases are published on Entscheidsuche.ch, allowing daily document retrieval (see Figure 2). **(1)** We scrape all files from Entscheidsuche.ch, including each court's folder metadata. Only new case documents are sent through the pipeline. **(2)** We used BeautifulSoup / tika-python library to extract text from HTML / PDF. **(3)** Language is identified using fastText Grave et al. [2018] for subsequent tasks. **(4)** A cleaner removes irregular patterns or redundant text to avoid extraction errors. **(5)** Cases are segmented into header, facts, considerations, rulings, and footer via regex patterns. **(6)** To extract the judgement outcome, a word set is defined for each outcome. As these indicators are not context-exclusive, considering only the ruling section is crucial to avoid false positives. Therefore, accurate judgment outcome extraction relies on precise section splitting. **(7)** Leading Decision (BGE) and law citations are obtained through Regex (cantonal) or BeautifulSoup (federal). The FSCS labels citations with HTML tags, ensuring a high quality of citations for federal cases.

Providing an objective metric for quality is hard and expensive to obtain. Multiple people repeated quality checks over multiple months during this process to ensure the highest quality. The parsers and regexes were double-checked by senior people before integration. We wrote a series of tests to make sure that the pipeline is robust to changes (test_utils.py). Finally, we wrote code to easily inspect samples at various stages of the pipeline to ensure quality (debug_utils.py).

4.2 Pretraining

Legislation The Swiss Legislation dataset comprises 35.7K legislative texts (182M tokens) distributed across five languages: German, French, Italian, Romansh, and English (see Table 7). Table 9 details its coverage of federal, cantonal, and inter-cantonal legislation on a broad array of legal topics including public health, education, civil rights, societal matters, energy, environment, infrastructure, and visa regulations. It also includes instances of the same legislation texts across different languages, useful for enhancing the multilingual capabilities of legal LMs.

Rulings The Swiss Rulings dataset is a comprehensive collection of Swiss court rulings designed for pretraining purposes. It consists of 638K cases (3.3B tokens) distributed across three languages: German (319K), French (247K) and Italian (71K). Spanning several decades and covering multiple areas of law, this dataset provides an extensive representation of Swiss law practice.

4.3 Text Classification

We work with eight different configurations, built from the LAP, CP and JP datasets. While the Law Area and Judgment Prediction datasets include both federal and cantonal cases, the Criticality dataset considers only FSCD. All tasks presented in this section involve Single Label Text Classification (SLTC), which required either extracting or defining labels. For each of the tasks we both consider the facts and the considerations as input. The facts represent the most similar available proxy to the complaints, useful for predictive tasks. The considerations as input make the tasks considerably easier, since they include the legal reasoning. These tasks can be used as post-hoc analyses for verification (e.g., in judgment prediction whether the made judgment is congruent with the given reasoning).

Table 2: Task Configurations. Label names are *Critical* (C), *Non-critical* (NC), *Critical-1* (C1) to *Critical-4* (C4), *Approval* (A), *Dismissal* (D). For Law-Sub-Area we reported only the two most common labels *Substantive Criminal* (SC), *Criminal Procedure* (CP), and the two least common *Intellectual Property* (IP), *Other Fiscal* (OF). Abbreviations: **Val**idation, **Con**siderations, **Fac**ts

Task Name	Train	Labels Train				Val	Labels Val				Test	Labels Test			
		C	**NC**				**C**	**NC**				**C**	**NC**		
BGE-Fac	**75K**	3K	72K	-	-	**12K**	580	13K	-	-	**26K**	950	25K	-	-
BGE-Con	**91K**	3K	85K	-	-	**15K**	580	13K	-	-	**32K**	948	29K	-	-
		C-1	**C-2**	**C-3**	**C-4**		**C-1**	**C-2**	**C-3**	**C-4**		**C-1**	**C-2**	**C-3**	**C-4**
Citation-Fac	**2.5K**	782	626	585	513	**563**	186	152	131	94	**725**	137	177	224	187
Citation-Con	**2.5K**	779	624	586	520	**563**	186	154	131	92	**723**	137	177	224	185
		D	**A**				**D**	**A**				**D**	**A**		
Judgment-Fac	**197K**	135K	62K	-	-	**37K**	27K	11K	-	-	**94K**	67K	27K	-	-
Judgment-Con	**188K**	130K	59K	-	-	**37K**	26K	11K	-	-	**92K**	66K	26K	-	-
		SC	**CP**	**IP**	**OF**		**SC**	**CP**	**IP**	**OF**		**SC**	**CP**	**IP**	**OF**
Law-Sub-Area-Fac	**10K**	3K	3K	6	2	**9K**	2K	1K	11	1	**3K**	1K	509	5	1
Law-Sub-Area-Con	**8K**	2K	1K	6	2	**7K**	2K	750	11	1	**3K**	885	401	3	1

Law Area Prediction The **Law Area** label was established by associating a law-area to each chamber where a case was adjudicated. Using metadata from Entscheidsuche.ch, a lawyer helped define the law areas for each chamber, resulting in chambers being classified into one of four main law-area categories (civil, public, criminal and social law) and 12 sub-law-areas. Due to many chambers operating in various law areas, it was not always feasible to assign a single law area label to each chamber. Particularly for the more detailed sub-law areas, where several chambers could not be uniquely linked, this resulted in a small subset of cases with the subset label. Initial results on the full dataset including the four main law areas showed that current models achieve near perfect accuracy, which is why we only consider the smaller filtered dataset of sub areas for this benchmark.

Judgment Prediction We created the **Judgment** label by extracting the judgment outcome with regex patterns and assign a binary label with two classes: approval and dismissal, similar to Niklaus et al. [2021]. For partially approved or dismissed judgments, we labeled them as approval or dismissal, respectively.

Criticality Prediction We quantified **Criticality** in two ways: First, the **BGE-Label** is binary: *critical* and *non-critical*. FSCD are labeled as *critical* if additionally published as Leading Decision (see section 3). To achieve this, we extracted the FSCD file names from the headers of BGE cases using regex patterns. Cases not found in the header of a BGE were labeled as *non-critical* (since we have all BGE but not all FSCD, there are missing *critical* cases). Second, to create a more precise adaptation of the BGE-Label, we developed the **Citation-Label**, which involved counting all citations of BGE in all FSCD cases. The BGE frequency was weighted based on recency, with older citations receiving a smaller weight: $score = count * \frac{year-2002+1}{2023-2002+1}$. This resulted in a ranking of BGEs, which were then divided into four categories of criticality *critical-1* to *critical-4*. We used the 25, 50 and 75% quartiles as separation for our four classes.

4.4 Text Generation

Court View Generation Clerks and judges dedicate a significant portion of their time to preparing considerations for court cases - approximately 50% in penal law and as much as an estimated 85% in other law areas [Niklaus et al., 2021]. Crafting considerations is arguably the central task of a judge's role, requiring intricate legal knowledge of applicable legislation, caselaw and legal analysis, and advanced reasoning skills to connect this myriad of information. The complexity of this task, especially in the Supreme Court, is reflected in the average appointment age of judges being 50 years[3], highlighting the length and difficulty of their professional journey. Given these demands for time and expertise, the necessity of the CVG task emerges, aiming to create case considerations from the facts. Generating court views is challenging for several reasons: Both the facts (input) and the considerations (output) can be long and complex. Current models, constrained by their limitations in handling long context, often fail to fully process this extensive input. This shortcoming, when coupled with the input's inherent complexity, underscores the deficiencies of current models. To overcome these limitations, we present a novel CVG dataset containing over 400K cases, covering a wide range of legal scenarios. With an average length of 1522 tokens for the facts and 4673 tokens for the considerations, this dataset provides a challenging benchmark for models to generate coherent and accurate case considerations from legal facts. Furthermore, we provide a Court View origin dataset featuring federal rulings, enriched with data from the lower courts, including their facts and considerations, as well as those of the federal court. This provides a multilevel judicial perspective, contributing to a more comprehensive understanding of case progression and further augmenting the challenge of CVG.

Leading Decision Summarization BGE are crucial in the Swiss legal system, often cited to clarify legislative gaps. Access to their summaries simplifies searching and understanding key concepts, the most important citations, and main themes. In the LDS dataset, we include 18K BGE with their summaries, penned by FSCS clerks and judges.

4.5 Information Retrieval

In our IR task, data is organized into queries, qrels, and corpus (see Figure 3). The corpus includes all Swiss legislation and leading decisions; queries come from FSCS cases in German, French, and Italian. The mean token count for our queries significantly exceeds that in other IR benchmarks, owing to the use of entire documents as queries (see Table 1). The goal is to find laws and decisions cited in a given case. We use the facts as a proxy for an appeal drafted by a lawyer. Ground truth is based on citations from the considerations. We find relevant laws and decisions by extracting cited law articles and decisions from the Swiss legislation and leading decisions datasets, respectively. Document lengths mirror tasks like EU2UK Chalkidis et al. [2021a]. Laws written in all three official languages result in cross-lingual query-corpus pairs, logged as qrels. Long documents and cross-lingual factors may challenge retrieval models. In total, we have 10K documents, 101K queries, and 2K qrels, yielding an average of 19 relevant documents per query.

4.6 Citation Extraction

The FSCS annotates citations with special HTML tags, which we used to create a token classification dataset for Citation Extraction (CE). Solving the task with regexes is complicated due to extensive citation rules, but the transformer-based model MiniLM Wang et al. [2020b] achieves over 95 macro F1. For brevity, we omit experiments on this dataset, but release the dataset and the trained model as a resource to the community (both available under https://huggingface.co/organizations/rcds).

4.7 The Big Picture

The pretraining corpus and our seven datasets JP, LAP, CP, IR, CVG, LDS, and CE form a unified framework that resembles an artificial judicial system (see Figure 1). Pre-training serves as the foundation, equipping models with the ability to specialize in the respective tasks and thereby enhancing their performance. The remaining tasks, all interconnected, focus on the output of the judicial system. Superior performance in one task can bolster the effectiveness of the others. LAP

[3] *Mit 28 Jahren Mitglied des Bundesgerichts*, SonntagsZeitung, December 19, 2019, p. 24

facilitates routing decisions to the correct chambers inside a court. CP enables courts allocating resources and setting priorities. IR identifies the relevant documents for a case, facilitating the JP and CVG tasks, which predicts the case's outcome and synthesizes a coherent text to explain the decision's rationale. CE automatically extracts citations to enrich the final decision before publication. LDS condenses the reasoning into a short summary. Together, these tasks model (albeit still primitively) the flow of the judicial system end-to-end, the first of this kind, to the best of our knowledge.

5 Experiments

In this section, we present the pretraining of our legal models and describe the experimental setup for each of the tasks. Besides BLOOM Scao et al. [2022] and mT5 Xue et al. [2021], there is a scarcity of open multilingual LLMs (> 500M parameters), with most recent work pretraining on English only. Only 4.5% of LLaMA2's Touvron et al. [2023] pretraining data is non-English text, whereas XLM-R's Conneau et al. [2020b] data contained 87% non-English text. Table 3 shows an overview of models we evaluated across tasks.

Table 3: Models: InLen is the maximum input length the model has seen during pretraining. # Paramaters is the total parameter count (including embedding). Our models were built upon the pre-trained RoBERTa/Longformer. SwissBERT was further trained from X-MOD. Utilizing three language adapters with X-MOD and SwissBERT led to fewer parameters and languages. (%de/%fr/%it) shows the percentages of the Swiss languages in the corpus. A question mark in brackets (?) indicates that the we could not find reliable sources.

Model	Source	InLen	# Parameters	Vocab	# Steps	BS	Corpus (%de/%fr/%it)	# Langs
MiniLM	Wang et al. [2020b]	512	118M	250K	1M	256	2.5TB CC100 (2.9/2.5/1.3)	100
DistilmBERT	Sanh et al. [2020]	512	135M	120K	n/a	< 4000	Wikipedia (na/na/na)	104
mDeBERTa-v3	He et al. [2021b,a]	512	278M	128K	500K	8192	2.5TB CC100 (2.9/2.5/1.3)	100
XLM-$R_{Base/Large}$	Conneau et al. [2020b]	512	278M/560M	250K	1.5M	8192	2.5TB CC100 (2.9/2.5/1.3)	100
X-MOD_{Base}	Pfeiffer et al. [2022]	512	299M	250K	1M	2048	2.5TB CC100 (2.9/2.5/1.3)	3 (81)
SwissBERT (XLM vocab)	Vamvas et al. [2023]	512	299M	250K	364K	768	Swissdox (80/18/1)	3 (4)
$mT5_{Small/Base/Large}$	Xue et al. [2021]	1K	300M/580M/1.2B	250K	1M	1024	mC4 (CC) (3.1/2.9/2.4)	101
$BLOOM_{560M}$	Scao et al. [2022]	2K	560M	250K	1.3M	256	ROOTS (0/5/0)	59
Legal-Swiss-R_{Base}	ours	512	184M	128K	1M	512	CH Legal (50/27/23)	3
Legal-Swiss-R_{Large}	ours	512	435M	128K	500K	512	CH Legal (50/27/23)	3
Legal-Swiss-LF_{Base}	ours	4096	208M	128K	50K	512	CH Legal (50/27/23)	3
Claude-2	Anthropic	100K	137B (?)	na	na	na	na	na
Claude-Instant	Anthropic	100K	52B (?)	na	na	na	na	na
GPT-3.5	Brown et al. [2020]	16K	175B	na	na	na	na	na
GPT-4	OpenAI [2023]	32K	1.8T (?)	na	na	na	na	na
PaLM-2	Anil et al. [2023]	8K	340B (?)	na	na	na	na	na
LLaMA-2	Touvron et al. [2023]	4K	7B/13B/70B	32K	na	na	LLaMA-2 (0.2/0.2/0.1)	27

5.1 Pretraining Legal Models

We release two multi-lingual legal-oriented PLMs, dubbed Legal-Swiss-RoBERTa and a Longformer, dubbed Legal-Swiss-LF_{Base}trained on Swiss rulings and legislation additional to EUR-LEX data Niklaus et al. [2023b] (https://huggingface.co/datasets/joelito/eurlex_resources). For the newly released Legal-Swiss-RoBERTa models we followed a series of best-practices in LM development literature described in more detail in Appendix G. We make all our models publicly available alongside all intermediate checkpoints (every 50K/10K training steps for RoBERTa/Longformer models) on the HuggingFace Hub (https://huggingface.co/joelito). Limited resources prevented us from pretraining a large generative model, so we leave this to future work.

5.2 Text Classification

For our TC tasks, namely LAP, JP, and CP, we adopted the LEXTREME benchmark setup [Niklaus et al., 2023a], namely hierarchical aggregation of macro-averaged F1 scores using harmonic mean for fairness (the harmonic mean is biased more towards lower scores than the geometric or arithmetic mean). We averaged in order over random seeds, languages (de, fr, it), configurations (e.g., JP-F and JP-C), and datasets (LAP, JP, and CP). This setup punishes models with outlier low scores in certain languages, or configurations, thus promoting fairer models. We fine-tuned all models below 2B parameters per task on our training datasets with early stopping on the validation dataset. We evaluated closed models zero-shot as per the setup of Chalkidis [2023], providing one instruction and example as input. We randomly selected samples from the validation set instead of the test set to avoid leaking the test set for future evaluations. For each sample, we checked whether it exceeded the model's maximum token limit of 4096 and truncated it if necessary. To manage costs, we limited the validation set to 1000 samples. Our experiments focused solely on zero-shot classification due to the long input lengths. We show an overview of the prompts used in Appendix K. We used the ChatCompletion API for GPT-3.5 (gpt3.5-turbo as of June 7, 2023), the Anthropic Claude API for Claude-2, the Vertex AI API for PaLM-2 (text-bison@001), and ran LLaMA-2 locally with 4-bit quantization.

5.3 Text Generation

We evaluated using **BERT**Score [Zhang et al., 2020b], **BLEU** [Papineni et al., 2002], **MET**EOR [Banerjee and Lavie, 2005], and **R**OUGE [Lin, 2004]. Each individual metric has inherent weaknesses [Zhang et al., 2020b], so it is necessary to employ multiple metrics for a more comprehensive assessment. We suggest future work to evaluate predictions with trained lawyers.

Court View Generation Due to lengthy input (avg. 1522 tokens) and output (avg. 4673 tokens) for CVG, we truncated input facts to 2048 tokens and output considerations to 512 tokens. In 90% of cases, complete facts were retained. This truncation was driven by resource constraints and the task's complexity, which remained challenging with only 512 output tokens. Owing to test data volume and compute limits, the evaluation was limited to a subset of 1K instances. For the origin dataset, input was evenly divided between origin facts and considerations.

Leading Decision Summarization In our LDS experiments, we faced large input text (avg. 3081 tokens) but shorter output text (avg. 168 tokens). To manage this, we truncated input to 4096 tokens and output to 256 tokens, preserving full output in over 80% of cases.

5.4 Information Retrieval

Finding relevant legal references for FSCD is challenging due to (a) legal language complexity, (b) multilinguality, and (c) long documents. We explore multilingual Doc2Doc IR in the legal domain using our new dataset, which includes FSCD with unique identifiers for law citations and BGE. Performance is expected to decline as document count increases. We conducted an ablation study to assess minor dataset adjustments on performance (see Appendix H). We employ BM25 for its scalability to long documents but note its limitations in contextual processing and multilingual handling Robertson and Zaragoza [2009]. Neural methods like Sentence-Bert (SBERT) show promise but degrade on long texts due to truncation-induced context loss. We also investigate training with hard negatives, using the distiluse-base-multilingual-cased-v1 SBERT model Zhan et al. [2021]. This 135M parameter model employs DistilmBERT as the student and Multilingual Universal Sentence Encoder (mUSE) as the teacher, trained on 15 languages. It has a 128 token maximum sequence length and outputs 512-dimensional embeddings via mean pooling, suitable for cosine similarity scoring. We excluded cross-encoder models Wang et al. [2020a], due to their high computational cost.[4] In addition to Normalized Discounted Cumulative Gain (NDCG) Wang et al. [2013] we use Capped Recall@k Thakur et al. [2021].[5]

6 Results

6.1 Text Classification

We present results in Table 4, with detailed information including standard deviations in Table 15. Language-specific scores are in Table 16. Scores on the validation dataset are in Table 17. As expected, larger models generally perform better, with XLM-R_{Large} emerging on top. Our pre-trained model Legal-ch R_{Base} outperformed XLM-R_{Base}, indicating that domain-specific pre-training enhances performance. Overall, our pre-trained models showed better aggregated results compared to other models. However, unexpectedly, Legal Swiss $RoBERTa_{Large}$ underperformed compared to its base model XLM_{Large}. Due to the high weight to outliers allotted by the harmonic mean, Legal-ch-R_{Large} is severely penalized by its relatively low performance in CPC-C compared to XLM-R_{Large}.

[4] BEIR datasets Thakur et al. [2021] feature query lengths of 3-192 words and document lengths of 11-635 words. Our dataset surpasses these by 4-282x for queries (847 words on average) and approx. 8-500x for documents (approx. 4K/7K words on average for rulings/legislation documents), further increasing computational costs.

[5] The Capped Recall@k is computed as the proportion of relevant documents for a specific query, retrieved from the top k scored list of documents generated by the model. This is a good representation of model success in our specific task, as each query has multiple relevant documents without the need for intra-document ranking.

Despite extra training on longer texts up to 4096 tokens, Legal-ch-LF did not surpass the hierarchical Legal-ch-R-$_{Base}$ model. Large models such as GPT-3.5, Claude-2, and LLaMA-2 underperform fine-tuned models, underlining the need for specialized models for these tasks. The difference is largest in the JP and Sub Law Area Prediction (SLAP) tasks where the fine-tuned models are best.

Table 4: Results on the Text Classification datasets. Macro F1 score is reported. The highest values are in bold. The 'F' or 'C' following the dash represents input based on 'Facts' or 'Considerations' respectively. 'CPB' and 'CPC' refer to the CP task using BGE and Citation labels, respectively, while 'SLAP' denotes Sub Law Area Prediction. Note: Seeds that yielded very high evaluation losses were considered failed and, therefore, excluded from the analysis. The models marked with an asterisk (*) are LLMs that generated zero-shot predictions (based on prompts) on a maximum of samples from the validation dataset, as described in section 5.2.

Model	CPB-F	CPB-C	CPC-F	CPC-C	SLAP-F	SLAP-C	JP-F	JP-C	Agg.
MiniLM	54.7	65.8	9.8	20.8	59.7	61.1	58.1	78.5	32.4
DistilmBERT	56.2	65.4	19.6	22.1	63.7	65.9	59.9	75.5	42.1
mDeBERTa-v3	55.1	69.8	21.0	17.5	63.8	59.3	60.6	77.9	40.2
XLM-R_{Base}	57.2	65.9	21.3	23.7	67.2	73.4	60.9	79.7	44.6
XLM-R_{Large}	56.4	67.9	24.4	**29.1**	65.1	78.9	60.8	80.9	**48.6**
X-MOD_{Base}	56.6	67.8	20.0	20.6	63.9	64.4	60.5	79.1	41.9
$SwissBERT_{(xlm\text{-}vocab)}$	56.9	67.3	25.7	23.0	61.5	73.2	61.4	79.4	46.1
$mT5_{Small}$	52.2	62.1	13.2	17.9	53.1	60.9	58.9	74.2	34.4
$mT5_{Base}$	52.1	61.5	14.0	19.7	58.4	61.8	54.5	72.0	35.9
$BLOOM_{560M}$	53.0	61.7	10.7	8.0	52.6	53.2	60.5	73.4	24.9
Legal-ch-R_{Base}	57.7	70.5	16.2	20.1	77.0	79.7	64.0	86.4	40.9
Legal-ch-R_{Large}	55.9	68.9	**25.8**	16.3	76.9	**84.9**	62.8	**87.1**	43.3
Legal-ch-LF_{Base}	**58.1**	**70.8**	21.4	17.4	**80.1**	77.1	**65.4**	86.4	42.5
GPT-3.5*	46.6	44.8	25.7	16.7	67.9	69.5	51.3	61.9	38.6
Claude-2*	38.4	40.5	16.6	19.7	60.2	60.5	48.9	48.1	33.9
LLaMA-2*	45.2	26.6	7.0	8.5	58.7	55.6	40.3	37.8	19.7
PaLM-2*	40.6	38.5	16.6	14.3	57.6	67.8	52.6	65.3	32.3

6.2 Text Generation

Court View Generation We present CVG results in Table 5. Fine-tuning models generally leads to higher scores, with even small models like $mT5_{Small}$ outperforming 1-shot GPT-4. 1-shot prompting only leads to marginal gains over 0-shot for both GPT-4 and Claude-2 (LLaMA-2 1-shot was not possible due to limited context width). Generally, the generated text exhibited stylistic authenticity, resembling typical legal language. However, it often lacked logical coherence in terms of its content, which underscores the current models' limited capacity to fully grasp the complexities of generating coherent court views. In multiple court cases, target considerations contained similar paragraphs, which were generally well-predicted (see examples in Table 30). While fine-tuned models proficiently predict specific textual patterns, LLaMA-2-13B-Chat in the zero-shot setup struggles, often reverting from German to English and introducing linguistic errors, probably due to a highly English dominant training corpus. Despite their challenges, zero-shot models focus more on the main content, while fine-tuned models mirror target formalities. We provide a more detailed error analysis in Appendix J.1. Larger mT5 models consistently outperformed smaller ones, but performance increase with longer input was minimal, sometimes counterproductive (see Table 11). The results from the origin dataset were less conclusive (see Table 10), likely due to the smaller dataset size.

Table 5: Results on the Court View Generation task. The input is truncated to 2048 tokens. **Bold**: best within setup; underlined: best overall. (*) These models were fine-tuned on only 1'000 samples for 3 epochs. All models, except for the mT5 models, were evaluated on the validation set.

Model	Setup	BERT ↑	BLEU ↑	MET ↑	R1 / R2 / RL ↑
$mT5_{Large}$	Fine-tuned	**75.74**	**66.92**	**34.44**	**34.91 / 15.58 / 33.53**
$mT5_{Base}$	Fine-tuned	75.01	65.48	32.89	33.23 / 13.57 / 31.89
$mT5_{Small}$	Fine-tuned	74.13	63.97	30.96	31.29 / 11.01 / 29.90
GPT-3.5-Turbo	Fine-tuned*	72.31	62.23	28.08	26.06 / 7.19 / 24.54
LLaMA-2-13B Chat	Fine-tuned*	**74.22**	**63.51**	**33.33**	**34.36 / 16.68 / 33.20**
GPT-4	1-shot	**70.39**	59.69	24.63	23.87 / 4.64 / 22.32
Claude-2	1-shot	69.45	**61.26**	**24.85**	**24.98 / 5.39 / 23.61**
GPT-4	0-shot	69.41	58.16	23.25	22.61 / 3.95 / 21.10
Claude-2	0-shot	**69.46**	**61.36**	**24.90**	**24.66 / 5.30 / 23.24**
LLaMA-2-13B Chat	0-shot	67.23	55.01	20.18	19.76 / 3.26 / 18.57

Leading Decision Summarization In contrast to the very specific CVG task, requiring long-form output, the closed large models perform very well on LDS, at least in BLEU and METEOR (see Table 6). According to ROUGE, fine-tuned mT5 models are still better, while BERT-Score does not discriminate clearly. We assume that summarization is a much larger portion of internal instruction tuning datasets used for optimizing these models. The quality of the generated text demonstrated a good stylistic imitation of legal language and more consistent logical coherence compared to the CVG task (see examples in Table 31 and Table 29). GPT-3.5-Turbo (zero-shot) offered a narrative-style summary, while others adhered to the traditional 'Regeste' format. Notably, GPT-3.5-Turbo made a factual error by negating a crucial element, and Claude-2 referenced an outdated legal provision. See Appendix J.2 for more detailed error analysis. Table 13 shows two trends for fine-tuned mT5 models. First, longer input generally improved scores across models. Second, larger models outperformed smaller ones, although the differences between base and large models were subtle.

Table 6: Results on the Leading Decision Summarization task. The input is truncated to 4096 tokens. **Bold**: best within setup; underlined: best overall. All models, except for the mT5 models, were evaluated on the validation set.

Model	Setup	BERT ↑	BLEU ↑	MET ↑	R1 / R2 / RL ↑
mT5_{Base}	Fine-tuned	**73.33**	**30.81**	**23.50**	<u>**32.43**</u> / <u>**12.78**</u> / <u>**30.87**</u>
$\text{mT5}_{\text{Small}}$	Fine-tuned	72.04	28.68	21.29	29.61 / 10.31 / 28.12
GPT-4	1-shot	<u>**73.55**</u>	**47.75**	<u>**34.72**</u>	**30.82 / 9.68 / 28.89**
GPT-3.5-Turbo-16K	1-shot	72.89	45.21	32.76	29.69 / 9.25 / 27.94
Claude 2	1-shot	72.91	47.55	33.57	30.28 / 9.12 / 28.58
Claude Instant	1-shot	72.44	44.80	30.29	27.89 / 8.56 / 26.18
GPT-4	0-shot	**71.56**	48.35	**32.97**	26.52 / **8.93** / 24.51
GPT-3.5-Turbo-16K	0-shot	70.28	46.08	30.60	25.18 / 7.58 / 23.59
Claude 2	0-shot	71.13	<u>**49.20**</u>	32.54	**27.70** / 8.39 / **25.90**
Claude Instant	0-shot	71.33	45.65	29.22	26.13 / 8.16 / 24.15

6.3 Information Retrieval

Table 7 shows that most models failed to retrieve relevant documents, even with k=100. Lexical models outperformed others even without hyperparameter optimization for BM25 Chalkidis et al. [2021b]. Surprisingly, despite German prevalence in our dataset, a French language analyzer (used for stemming and stopword removal) demonstrated superior performance. For SBERT truncation led to context loss, negatively affecting scores, a problem absent in lexical models. Training SBERT models using Multiple Negative Ranking Loss Henderson et al. [2017] significantly improved performance, with hard negative examples beneficial. SBERT evaluation on single languages, denoted as DE, FR, and IT, revealed its inability to perform consistently across all languages, which could be caused by the training set consisting of more German than French or Italian documents. More experiments are in Tables 8 and 9. Overall, our study exposes limitations of models in handling multilingualism, long documents, and legal texts, areas relatively underexplored in previous research. These findings offer a foundation for the IR community to address these challenges.

Table 7: Results on Information Retrieval with best scores per section in **bold**. Abbreviations: **distil**use_{Base}-multilingual-cased-v1, joelito/**swiss**-legal-$\text{roberta}_{\text{Base}}$

Model	RCap@ 1 / 10 / 100 ↑	NDCG@ 1 / 10 / 100 ↑
BM25 (fr lang analyzer)	**11.37 / 7.74 / 16.54**	**11.37 / 8.34 / 11.51**
SBERT distil	0.90 / 0.75 / 2.64	2.06 / 1.70 / 3.31
SBERT distil + pos	**4.40** / 3.92 / 12.64	**10.11** / 8.76 / 16.16
SBERT distil + pos + h-neg	3.97 / **4.46** / **13.36**	9.12 / **9.21** / **16.87**
SBERT swiss + pos	3.97 / 3.47 / 12.28	9.12 / 7.76 / 15.16
SBERT distil eval on de queries	**4.22 / 4.49 / 15.21**	**8.21 / 8.15 / 15.86**
SBERT distil eval on fr queries	1.88 / 2.20 / 9.19	5.77 / 6.22 / 13.94
SBERT distil eval on it queries	0.22 / 0.24 / 0.79	5.43 / 5.74 / 11.44

7 Conclusions

We present SCALE, an end-to-end benchmark of seven datasets for the Swiss legal system, a worldwide unique possibility to study crosslinguality within the same jurisdiction. Our tasks require legal reasoning abilities and challenge models on four key aspects: long documents, domain-specificity, multilinguality, and multitasking. We evaluate 14 open and five closed multilingual models, including three in-domain pretrained, as a reference point. These models, including ChatGPT, Claude-2, LLaMA-2 and PaLM-2, show low performance, particularly in challenging tasks like CVG and IR. Our results highlight opportunities for improving models and set the stage for next-generation LLM evaluations in domain-specific, multilingual contexts.

References

Ehsan Aghaei, Xi Niu, Waseem Shadid, and Ehab Al-Shaer. SecureBERT: A Domain-Specific Language Model for Cybersecurity. In Fengjun Li, Kaitai Liang, Zhiqiang Lin, and Sokratis K. Katsikas, editors, *Security and Privacy in Communication Networks*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 39–56. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-25538-0_3.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective. *PeerJ Computer Science*, 2:e93, October 2016. ISSN 2376-5992. doi: 10.7717/peerj-cs.93. URL `https://peerj.com/articles/cs-93`. Publisher: PeerJ Inc.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. PaLM 2 Technical Report, May 2023. URL `http://arxiv.org/abs/2305.10403`. arXiv:2305.10403 [cs].

Kevin D. Ashley. *Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age*. Cambridge University Press, 2017. doi: 10.1017/9781316761380.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. Ms marco: A human generated machine reading comprehension dataset, 2018.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL `https://aclanthology.org/W05-0909`.

Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A Pretrained Language Model for Scientific Text. *arXiv:1903.10676 [cs]*, September 2019. URL `http://arxiv.org/abs/1903.10676`. arXiv: 1903.10676.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*, December 2020. URL `http://arxiv.org/abs/2004.05150`. arXiv: 2004.05150.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://aclanthology.org/2020.acl-main.463`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf`.

Tobias Brugger, Matthias Stürmer, and Joel Niklaus. MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset, May 2023. URL `http://arxiv.org/abs/2305.01211`. arXiv:2305.01211 [cs].

Schweizerisches Bundesgericht. The paths to the swiss federal supreme court. `https://www.bger.ch/files/live/sites/bger/files/pdf/en/BG_Brosch%C3%BCreA5_E_Onl.pdf`, 2019. Accessed: 2023-04-27.

Ilias Chalkidis. ChatGPT may Pass the Bar Exam soon, but has a Long Way to Go for the LexGLUE benchmark. *ArXiv*, abs/2304.1, 2023.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy, July 2019a. Association for Computational Linguistics. doi: 10.18653/v1/P19-1424. URL `https://www.aclweb.org/anthology/P19-1424`.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Extreme Multi-Label Legal Text Classification: A Case Study in EU Legislation. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 78–87, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/W19-2209. URL `https://www.aclweb.org/anthology/W19-2209`.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, 2020.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. Regulatory compliance through doc2doc information retrieval: A case study in eu/uk legislation where text similarity has limitations, 2021a.

Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. Regulatory Compliance through Doc2Doc Information Retrieval: A case study in EU/UK legislation where text similarity has limitations, 2021b. _eprint: 2101.10726.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 226–241, 2021c.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, 2022.

Huajie Chen, Deng Cai, Wei Dai, Zehui Dai, and Yadong Ding. Charge-Based Prison Term Prediction with Deep Gating Network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6362–6367, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1667. URL `https://aclanthology.org/D19-1667`.

Yiran Chen, Zhenqiao Song, Xianze Wu, Danqing Wang, Jingjing Xu, Jiaze Chen, Hao Zhou, and Lei Li. MTG: A benchmark suite for multilingual text generation. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2508–2527, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.192. URL https://aclanthology.org/2022.findings-naacl.192.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv:2003.10555 [cs]*, March 2020. URL http://arxiv.org/abs/2003.10555. arXiv: 2003.10555.

Alexis Conneau and Guillaume Lample. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/hash/c04c19c2c2474dbf5f7ac4372c5b9af1-Abstract.html.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020a.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*, April 2020b. URL http://arxiv.org/abs/1911.02116. arXiv: 1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Yi Feng, Chuanyi Li, and Vincent Ng. Legal judgment prediction: A survey of the state of the art. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5461–5469. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/765. URL https://doi.org/10.24963/ijcai.2022/765. Survey Track.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Claire Grover, Ben Hachey, and Ian Hughson. The HOLJ corpus. supporting summarisation of legal texts. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*, pages 47–54, Geneva, Switzerland, aug 29 2004. COLING. URL https://aclanthology.org/W04-1907.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1), oct 2021. ISSN 2691-1957. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Neel Guha, Daniel E. Ho, Julian Nyarko, and Christopher Ré. LegalBench: Prototyping a Collaborative Benchmark for Legal Reasoning, September 2022. URL http://arxiv.org/abs/2209.06120. arXiv:2209.06120 [cs].

Ben Hachey and Claire Grover. Extractive summarisation of legal texts. *Artificial Intelligence and Law*, 14(4):305–345, 2006. ISSN 1572-8382. doi: 10.1007/s10506-007-9039-z. URL `https://link.springer.com/article/10.1007/s10506-007-9039-z`.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *arXiv:2111.09543 [cs]*, December 2021a. URL `http://arxiv.org/abs/2111.09543`. arXiv: 2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with Disentangled Attention. *arXiv:2006.03654 [cs]*, October 2021b. URL `http://arxiv.org/abs/2006.03654`. arXiv: 2006.03654.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652, 2017.

Peter Henderson, Mark S. Krass, Lucia Zheng, Neel Guha, Christopher D. Manning, Dan Jurafsky, and Daniel E. Ho. Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset, July 2022. URL `http://arxiv.org/abs/2207.00220`. arXiv:2207.00220 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding, January 2021a. URL `http://arxiv.org/abs/2009.03300`. arXiv:2009.03300 [cs].

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review, 2021b.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization, September 2020. URL `http://arxiv.org/abs/2003.11080`. arXiv:2003.11080 [cs].

Yibo Hu, MohammadSaleh Hosseini, Erick Skorupa Parolin, Javier Osorio, Latifur Khan, Patrick Brandt, and Vito D'Orazio. ConfliBERT: A Pre-trained Language Model for Political Conflict and Violence. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5469–5482. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.400. URL `https://aclanthology.org/2022.naacl-main.400`.

Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber. LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension, December 2022. URL `http://arxiv.org/abs/2212.08204`. arXiv:2212.08204 [cs].

Allen H. Huang, Amy Y. Zang, and Rong Zheng. Evidence on the information content of text in analyst reports. *ERN: Econometric Modeling in Financial Economics (Topic)*, 2014.

G Thomas Hudson and Noura Al Moubayed. Muld: The multitask long document benchmark. *arXiv preprint arXiv:2202.07362*, 2022.

Wonseok Hwang, Dongjun Lee, Kyoungyeon Cho, Hanuhl Lee, and Minjoon Seo. A Multi-Task Benchmark for Korean Legal Language Understanding and Judgement Prediction, October 2022. URL `http://arxiv.org/abs/2206.05224`. arXiv:2206.05224 [cs].

Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. Summarization of legal documents: Where are we now and the way forward. *Computer Science Review*, 40:100388, 2021. ISSN 1574-0137. doi: https://doi.org/10.1016/j.cosrev.2021.100388. URL `https://www.sciencedirect.com/science/article/pii/S1574013721000289`.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35. URL `https://europepmc.org/articles/PMC4878278`.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II au2. Natural language processing in the legal domain, 2023a.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael James Bommarito. Natural Language Processing in the Legal Domain, January 2023b. URL https://papers.ssrn.com/abstract=4336224.

Kornraphop Kawintiranon and Lisa Singh. PoliBERTweet: A Pre-trained Language Model for Analyzing Political Content on Twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7360–7367. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.801.

Mi-Young Kim, Ying Xu, and Randy Goebel. Summarization of legal texts with high cohesion and automatic compression rate. In Yoichi Motomura, Alastair Butler, and Daisuke Bekki, editors, *New Frontiers in Artificial Intelligence*, pages 190–204, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-39931-2.

Anastassia Kornilova and Vladimir Eidelman. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, 2019.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *arXiv:1909.11942 [cs]*, February 2020. URL http://arxiv.org/abs/1909.11942. arXiv: 1909.11942.

Dawn Lawrie, Eugene Yang, Douglas W. Oard, and James Mayfield. Neural approaches to multilingual information retrieval, 2023.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682. URL https://doi.org/10.1093/bioinformatics/btz682.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. Give Me Convenience and Give Her Death: Who Should Decide What Uses of NLP are Appropriate, and on What Basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.261. URL https://aclanthology.org/2020.acl-main.261.

Johannes Leveling. On the effect of stopword removal for sms-based faq retrieval. In *International Conference on Applications of Natural Language to Data Bases*, 2012.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, April 2021. URL http://arxiv.org/abs/2005.11401. arXiv:2005.11401 [cs].

Quanzhi Li and Qiong Zhang. Court opinion generation from case fact description with legal basis. In *AAAI Conference on Artificial Intelligence*, 2021a.

Quanzhi Li and Qiong Zhang. Court opinion generation from case fact description with legal basis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14840–14848, May 2021b. doi: 10.1609/aaai.v35i17.17742. URL https://ojs.aaai.org/index.php/AAAI/article/view/17742.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models, November 2022. URL `http://arxiv.org/abs/2211.09110`. arXiv:2211.09110 [cs].

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013`.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4046–4062, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.313. URL `https://aclanthology.org/2021.acl-long.313`.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.*, 65(4):782–796, apr 2014. ISSN 2330-1635. doi: 10.1002/asi.23062. URL `https://doi.org/10.1002/asi.23062`.

Mercedes Martínez-González, Pablo de la Fuente, and Dámaso-Javier Vicente. Reference Extraction and Resolution for Legal Texts. In Sankar K. Pal, Sanghamitra Bandyopadhyay, and Sambhunath Biswas, editors, *Pattern Recognition and Machine Intelligence*, Lecture Notes in Computer Science, pages 218–221, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-32420-1. doi: 10.1007/11590316_29.

Maria Medvedeva, Michel Vols, and Martijn Wieling. Judicial decisions of the European Court of Human Rights: looking into the crystall ball. *Proceedings of the Conference on Empirical Legal Studies in Europe 2018*, 2018. URL `https://research.rug.nl/en/publications/judicial-decisions-of-the-european-court-of-human-rights-looking-`.

Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. Dynamic Context Extraction for Citation Classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 539–549, Online only, November 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.aacl-main.41`.

Usman Naseem, Adam G Dunn, Matloob Khushi, and Jinman Kim. Benchmarking for biomedical natural language processing tasks with a domain specific albert. *BMC bioinformatics*, 23(1):1–15, 2022.

Joel Niklaus and Daniele Giofré. BudgetLongformer: Can we Cheaply Pretrain a SotA Legal Language Model From Scratch?, November 2022. URL `http://arxiv.org/abs/2211.17135`. arXiv:2211.17135 [cs].

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.nllp-1.3`.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. An Empirical Study on Cross-X Transfer for Legal Judgment Prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 32–46, Online only, November 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.aacl-main.3`.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. Lextreme: A multi-lingual and multi-task benchmark for the legal domain, 2023a.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. MultiLegalPile: A 689GB Multilingual Legal Corpus, June 2023b. URL `http://arxiv.org/abs/2306.02069`. arXiv:2306.02069 [cs].

OpenAI. GPT-4 Technical Report, March 2023. URL `http://arxiv.org/abs/2303.08774`. arXiv:2303.08774 [cs].

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL `https://arxiv.org/abs/2203.02155`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL `https://doi.org/10.3115/1073083.1073135`.

Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. Lesicin: A heterogeneous graph-based approach for automatic legal statute identification from indian legal documents, 2021.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and elmo on ten benchmarking datasets. In *BioNLP@ACL*, pages 58–65. Association for Computational Linguistics, 2019a. doi: 10.18653/v1/W19-5006.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, 2019b. URL `http://arxiv.org/abs/1906.05474`.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. UNKs everywhere: Adapting multilingual language models to new scripts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10186–10203, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.800. URL `https://aclanthology.org/2021.emnlp-main.800`.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. Lifting the Curse of Multilinguality by Pre-training Modular Transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.255. URL `https://aclanthology.org/2022.naacl-main.255`.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. ISSN 1533-7928. URL `http://jmlr.org/papers/v21/20-074.html`.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009. ISSN 1554-0669. doi: 10.1561/1500000019. URL `http://dx.doi.org/10.1561/1500000019`.

Sebastian Ruder, Jonathan H. Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel A. Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana L. Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David I. Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. Xtreme-up: A user-centric scarce-data benchmark for under-represented languages, 2023.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*, February 2020. URL `http://arxiv.org/abs/1910.01108`. arXiv: 1910.01108.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model, November 2022. URL `http://arxiv.org/abs/2211.05100`. arXiv:2211.05100 [cs].

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language Models Can Teach Themselves to Use Tools, February 2023. URL `http://arxiv.org/abs/2302.04761`. arXiv:2302.04761 [cs].

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.nllp-1.3`.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. WHEN FLUE MEETS FLANG: Benchmarks and large pre-trained language model for financial domain, 2022. URL `http://arxiv.org/abs/2211.00083`.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*, 2022.

Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, Margo Schlanger, and Doug Downey. Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities, July 2022. URL `http://arxiv.org/abs/2206.10883`. arXiv:2206.10883 [cs].

Jerrold Soh, How Khang Lim, and Ian Ernst Chai. Legal Area Classification: A Comparative Study of Text Classifiers on Singapore Supreme Court Judgments. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 67–77, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-2208. URL `http://aclweb.org/anthology/W19-2208`.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning,

Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout

Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models, June 2022. URL `http://arxiv.org/abs/2206.04615`. arXiv:2206.04615 [cs, stat].

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A Large Language Model for Science, November 2022. URL `http://arxiv.org/abs/2211.09085`. arXiv:2211.09085 [cs, stat].

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models, October 2021. URL `http://arxiv.org/abs/2104.08663`. arXiv:2104.08663 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL `http://arxiv.org/abs/2307.09288`. arXiv:2307.09288 [cs].

Jannis Vamvas, Johannes Graën, and Rico Sennrich. Swissbert: The multilingual language model for switzerland. *arXiv e-prints*, pages arXiv–2303, 2023.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446`.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. page 30, 2019.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 5776–5788. Curran Associates, Inc., 2020b. URL `https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html`.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. A theoretical analysis of ndcg type ranking measures. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL `https://proceedings.mlr.press/v30/Wang13.html`.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652, 2021. URL `https://arxiv.org/abs/2109.01652`.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.217.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance, 2023. URL http://arxiv.org/abs/2303.17564.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction. *arXiv:1807.02478 [cs]*, July 2018. URL http://arxiv.org/abs/1807.02478. arXiv: 1807.02478.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, 2:79–84, January 2021. ISSN 2666-6510. doi: 10.1016/j.aiopen.2021.06.003. URL https://www.sciencedirect.com/science/article/pii/S2666651021000176.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. CLUE: A Chinese language understanding evaluation benchmark. In *COLING*, pages 4762–4772, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.419. URL https://aclanthology.org/2020.coling-main.419.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv:2010.11934 [cs]*, March 2021. URL http://arxiv.org/abs/2010.11934. arXiv: 2010.11934.

Yi Yang, Mark Christopher Siy UY, and Allen Huang. FinBERT: A pretrained language model for financial communications, 2020. URL http://arxiv.org/abs/2006.08097.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1503–1512, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462880. URL https://doi.org/10.1145/3404835.3462880.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. *arXiv:1912.08777 [cs]*, July 2020a. URL http://arxiv.org/abs/1912.08777. arXiv: 1912.08777.

Ningyu Zhang, Mosha Chen, Zhen Bi, Xiaozhuan Liang, Lei Li, Xin Shang, Kangping Yin, Chuanqi Tan, Jian Xu, Fei Huang, Luo Si, Yuan Ni, Guotong Xie, Zhifang Sui, Baobao Chang, Hui Zong, Zheng Yuan, Linfeng Li, Jun Yan, Hongying Zan, Kunli Zhang, Buzhou Tang, and Qingcai Chen. CBLUE: A Chinese biomedical language understanding evaluation benchmark. In *ACL*, pages 7888–7915, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.544. URL https://aclanthology.org/2022.acl-long.544.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*, February 2020b. URL http://arxiv.org/abs/1904.09675. arXiv: 1904.09675.

Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset. *arXiv:2104.08671 [cs]*, July 2021. URL http://arxiv.org/abs/2104.08671. arXiv: 2104.08671 version: 3.

Zhe Zheng, Xin-Zheng Lu, Ke-Yin Chen, Yu-Cheng Zhou, and Jia-Rui Lin. Pretrained domain-specific language model for natural language processing tasks in the aec domain. *Comput. Ind.*, 142(C), nov 2022. ISSN 0166-3615. doi: 10.1016/j.compind.2022.103733. URL https://doi.org/10.1016/j.compind.2022.103733.

Sicheng Zhou, Nan Wang, Liwei Wang, Hongfang Liu, and Rui Zhang. CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *Journal of the American Medical Informatics Association*, 29(7):1208–1216, 03 2022. ISSN 1527-974X. doi: 10.1093/jamia/ocac040. URL https://doi.org/10.1093/jamia/ocac040.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. Predicting the Law Area and Decisions of French Supreme Court Cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_092. URL https://doi.org/10.26615/978-954-452-049-6_092.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Appendix C
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Appendix E
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Sections 1, 4, and 5
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Appendix G.3
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We ran all the text classification experiments with three random seeds each and show standard deviations in the appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Appendix G.2
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] Sections 1, 2, 4, 5
   (b) Did you mention the license of the assets? [Yes] Section 1
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Sections 1 and 5
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] The data is already public and not copyrighted.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Section 4
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix Table of Contents

We include the following supplementary sections as an addition to the main paper:

B Access to the Provided Resources

In this section, we provide the URLs to the data, models, and code.

B.1 Data

- Judgment Prediction: *https://huggingface.co/datasets/rcds/swiss_judgment_prediction_xl*
- Law Area Prediction: *https://huggingface.co/datasets/rcds/swiss_law_area_prediction*
- Criticality Prediction: *https://huggingface.co/datasets/rcds/swiss_criticality_prediction*
- Court View Generation: *https://huggingface.co/datasets/rcds/swiss_court_view_generation*
- Leading Decision Summarization: *https://huggingface.co/datasets/rcds/swiss_leading_decision_summarization*
- Information Retrieval: *https://huggingface.co/datasets/rcds/swiss_doc2doc_ir*
- Citation Extraction: *https://huggingface.co/datasets/rcds/swiss_citation_extraction*
- Rulings: *https://huggingface.co/datasets/rcds/swiss_rulings*
- Legislation: *https://huggingface.co/datasets/rcds/swiss_legislation*
- Leading Decisions: *https://huggingface.co/datasets/rcds/swiss_leading_decisions*

B.2 Models

- Legal-CH-RoBERTa$_{\text{Base}}$: *https://huggingface.co/joelito/legal-swiss-roberta-base*
- Legal-CH-RoBERTa$_{\text{Large}}$: *https://huggingface.co/joelito/legal-swiss-roberta-large*
- Legal-CH-Longformer$_{\text{Base}}$: *https://huggingface.co/joelito/legal-swiss-longformer-base*
- Citation Extraction: *https://huggingface.co/rcds/MiniLM-swiss_citation_extraction-de-fr-it*

B.3 Code

- Data Preparation (SwissCourtRulingCorpus): *https://github.com/JoelNiklaus/SwissCourtRulingCorpus*
- Text Classification Experiments (LEXTREME): *https://github.com/JoelNiklaus/LEXTREME*
- Text Generation Experiments: *https://github.com/vr18ub/court_view_generation*
- Information Retrieval Experiments: (BEIR): *https://github.com/Stern5497/Doc2docBeirIR*
- Citation Extraction Experiments (LEXTREME): *https://github.com/JoelNiklaus/LEXTREME*
- Zero-Shot Text Classification with LLM: *https://github.com/kapllan/zeroshot_lexglue*

C Limitations

C.1 General

The research area for language models and benchmarks continues to evolve, and while there is palpable enthusiasm in the field, it is critical to maintain a balanced perspective. Studies, including Bender and Koller [2020], have shed light on the limitations of language models and benchmarks, stressing that language models do not truly "learn" meaning and that communities often focus on limited datasets, some of which are borrowed from other fields.

C.2 Models

Even though English models are plentiful, LLMs pre-trained multilingually are very rare. To the best of our knowledge, mT5 is the only multilingual model with variants over 1B parameters, covering German, French, and Italian (BLOOM does not contain German and Italian). Additionally, we were limited by very large sequence lengths. We did not have the resources available to run mT5_{XL} or mT5_{XXL} with sequence lengths greater than 1K.

C.3 Data

The process of selecting tasks for benchmarks is typically influenced by the interests of the community or the convenience of available resources, rather than being informed by all-encompassing theories. These constraints present difficulties when trying to explore a model's broader applicability or its capacity for understanding. The data employed in benchmarks is often tied to a specific context and is naturally susceptible to inherent biases. Furthermore, the content of such data may vary significantly from real-world data, is de-contextualized and the uniformity of the task formats may not adequately reflect the diversity of human activities. Regarding our specific context, it is crucial to acknowledge that we cannot generalize Swiss legal data to other countries or different legal systems.

Figures 6 and 8 show the language and cantonal distributions over rulings and Figures 7 and 9 over legislation. Note that the distribution is imbalanced for both rulings (50% German, 39% French, and 11% Italian) and legislation (49% German, 31% French, and 17% Italian). However, compared to (Swiss Speaker Distribution (63% German, 23% French, and 8% Italian), the legal text is actually more balanced. It looks similar in the cantonal distribution, with many sparsely inhabited cantons being represented above their weight, especially the ones in the non-German speaking regions such as Vaud, Geneva, and Ticino.

C.4 Labels

Annotating high-quality datasets is very expensive, especially when experts are needed, such as in the legal domain. Because of our limited budget and in order to arrive at a large amount of labels, we algorithmically generated labels based on metadata information present in the corpus. These metadata are of high quality, being provided by the courts themselves.

C.5 Text Classification

Judgment Prediction While the judgment prediction task is arguably very interesting and also very challenging, it is unlikely to be deployed in practice anytime soon. Ideally, we would want the complaints as input (similar to Semo et al. [2022]) instead of the facts description, since this is written by the court itself in part to justify its reasoning. Unfortunately, the complaints are not public in Switzerland, making us rely on the widely available facts description as a close proxy.

Law Area Prediction We used information about the chambers at the courts to determine the law areas. Predicting the main law area is not challenging for current models, leading to very high results and thus rendering this task unsuitable for a benchmark. Unfortunately, most chambers cover multiple sub areas, thus ruling them out for the sub area prediction task and considerably reducing dataset size. In conclusion, while this task is very useful in practice for routing requests to the different chambers inside a court, it relatively unsuitable for a challenging benchmark.

Criticality Prediction It is very difficult to estimate the importance of a case. By relying on proxies such as whether the case was converted to a leading decision (BGE-label) and how often this leading decision was cited (Citation-label), we were able to create labels semi-automatically. While we discussed this with lawyers at length and implemented the solution we agreed on finally, this task remains somewhat artificial.

C.6 Text Generation

Court View Generation Court View Generation is an extremely challenging task and thus very well suitable for a benchmark. Current multilingual transformer-based models do not allow processing text in the tens of thousands of tokens. As a consequence, we were forced to look at a simplified version of this task, only considering the facts as input and ignoring relevant case law and legislation. Additionally, we were only able to generate the first 512 tokens of the considerations. We thus invite the community to develop new methods potentially capable of tackling harder versions of this task.

Leading Decision Summarization Due to limited resources, we limited our evaluation to $\text{mT5}_{\text{Small/Base/Large}}$. Future work may investigate large multilingually pretrained generative models on this task. Additionally, one may want to conduct human evaluation on the generated summaries. Finally, we only considered the simple version of this task where we only generate a text-based output. Future work may treat the first and second parts of the summary as extreme multilabel classification problems of relevant citations and relevant keywords from the thesaurus Jurivoc respectively, possibly increasing performance.

C.7 Citation Extraction

Even though the citations are annotated by the FSCS, we encountered citations that were not marked. However, models achieved very high scores anyway, leading us to exclude it from the benchmark. Future work may investigate this in more detail.

C.8 Information Retrieval

The labels are constructed with the citations from the considerations. Due to most legal analyses being private, our corpus is restricted to case law and legislation. Constructing a ranking of relevant documents is challenging due to missing information, and thus probably requiring extensive human annotation. Additionally, S-BERT models are usually limited to 512 tokens, being a constraint for this task due to our long documents.

D Directions for Future Research

The political parties of the judges in the ruling determine in what direction the ruling will go. In future work we would like to enrich the dataset with this information to make models more accurate in the judgment prediction task.

For simplicity, we treat the summary (regeste) of the leading decisions as just one string. Actually, it is composed of important citations in the first part, keywords from a Thesaurus in the second part and a text-based summary in the third part. The first two tasks could be framed as classification or retrieval tasks, possibly improving model performance.

Due to limited context width, we only considered the facts as input to the court view generation task. However, judges and clerks do not only look at the facts when drafting a decision. They consider a myriad of information including possible lower court decisions and relevant case law, legislation and legal analyses. This information is available in our dataset. In the future, we would like to develop systems that are capable of integrating all this information to write the legal reasoning.

So far, to our knowledge, the largest model pretrained on legal data specifically is Legal-XLM-R_{Large} (435M parameters) Niklaus et al. [2023b]. Future work should look at pretraining larger generative models in the billions and tens of billions of parameters.

Future work may investigate the more difficult Citation Prediction task in addition to the Citation Extraction task. In Citation Prediction, the model only gets the context up to the citation as input and is tasked to predict the citation. This may help lawyers in drafting their texts.

We strongly suggest future work include relevant external information, like caselaw or legislation for solving these challenging tasks. Augmenting models with retrieval Lewis et al. [2021] and models using tools Schick et al. [2023] seem to be promising avenues.

Finally, to provide a better perspective on the results, we suggest future work to collect human performance as an additional reference point. This may be done at several levels, i.e., laypeople, law students, early career lawyers, expert lawyers in the respective field.

E Ethical Considerations

While our research has several positive applications, it is important to acknowledge potential negative societal impacts. Large Language Models (LLMs) and their applications in the legal domain could potentially automate certain tasks traditionally performed by legal professionals, such as legal IR and LDS. While our goal is to support lawyers, it could impact the job market for legal professionals.

Recent literature has identified potential ethical problems within legal NLP research. The study on legal judgement prediction in China regarding prison term duration demonstrated the criticalness of legal NLP datasets and analysis Chen et al. [2019]. As a response to this publication at EMNLP 2019, concerned researchers pointed out important questions to respond when working with ethically delicate data and NLP tasks Leins et al. [2020]. They suggest asking a series of ethical questions to assess the potential societal risks associated with a publication.

For example, they asked: "Does the dataset contain information that might be considered sensitive or confidential?" Leins et al. [2020]. In our case, we only used publicly available court decisions that are anonymized. Therefore, the person should not be identifiable. Another aspect is concerned with the possibility of future updates of the court decision due to new facts or an appeal (going to a higher court): "Will the dataset be updated?" We could update our data anytime. However, this maintenance would not happen automatically. We would need to be informed that a new decision was made regarding a certain case.

Another ethical concern concerns precision: Thus, LLMs can occasionally deliver results that are not entirely precise. This can have severe implications when it comes to the legal domain, where precision and factual accuracy are paramount. This could potentially lead to misinformation or misinterpretation of legal texts, impacting legal proceedings and decisions.

While the benchmark focuses on the Swiss legal system, it is important to recognize that law systems are highly culturally and contextually dependent. The understanding and interpretations of legal texts by these models, especially in a multilingual context, might not accurately reflect the nuanced cultural aspects of different regions. This could potentially lead to misrepresentations or misinterpretations, particularly when applied to other legal systems.

Finally, like any AI model, LLMs could be misused to create misinformation or misleading content at scale, especially in languages and domains where automated content generation is still a novel concept. It is crucial to develop and implement robust ethical guidelines and policies to mitigate these risks.

Therefore, while the new benchmark presents exciting opportunities to improve LLMs, it is essential to carefully consider the implications of its use and manage the associated risks effectively. The developers and users of such technology should adhere to ethical guidelines to ensure its responsible use.

F Additional Related Work

F.1 Domain-Specific Pretraining

General-purpose language models are trained on generic text corpora such as Wikipedia and evaluated on widely used benchmarks such as GLUE [Wang et al., 2018]. However, domain-specific models need focused datasets for training and specialized benchmarks for assessing the quality of the model. The following examples illustrate the increase in performance when using domain-specific datasets and benchmarks.

In the biomedical area of natural language processing (BioNLP), Lee et al. [2019] created for the first time a domain-specific LM based on BERT [Devlin et al., 2019b] by pre-training it on biomedical text corpora. They used PubMed abstracts (4.5B words) and PubMed Central (PMC) full-text articles (13.5B words). The resulting domain-specific LM BioBERT achieved higher F1 scores than BERT in the biomedical NLP tasks named entity recognition (0.62) and relation extraction (2.80), and a higher mean reciprocal rank (MRR) score (12.24) in the biomedical question-answering task. In 2022, those scores were outperformed. Naseem et al. [2022] conducted a domain-specific pre-training of ALBERT Lan et al. [2020] using only text from the biomedical field (PudMed) and from the "Medical Information Mart for Intensive Care" (MIMIC-III), a large, de-identified and publicly-available collection of medical records [Johnson et al., 2016]. One domain-specific benchmark applied to test BioALBERT originates from Gu et al. [2021] who created BLURB, the Biomedical Language Understanding and Reasoning Benchmark. Naseem et al. [2022] found that BioALBERT exceeded the state-of-the-art models by 11.09% in terms of micro averaged F1-score (BLURB score). Another biomedical NLP benchmark is BLUE, the "Biomedical Language Understanding Evaluation" [Peng et al., 2019b]. It covers five tasks (sentence similarity, named entity recognition, relation extraction, document classification, inference) with ten datasets from the biomedical and clinical area. BioALBERT also includes all datasets and tests from BLUE thus presenting the most comprehensive domain-specific model and benchmark in the biomedical area at the moment.

In the financial domain, FinBERT was pretrained 2020 by Yang et al. [2020] using financial data. The text corpora consisted of 203'112 corporate reports (annual and quarterly reports from the Securities Exchange Commission SEC), 136'578 earnings call transcripts (conference call transcripts from CEOs and CFOs), and 488'494 analyst reports (textual analysis of the company) resulting in 3.3B tokens. For testing FinBERT, Yang et al. [2020] used the Financial Phrase Bank dataset with 4'840 sentiment classifications [Malo et al., 2014], the AnalystTone Dataset with 10'000 sentences [Huang et al., 2014], and FiQA Dataset with 1'111 sentences from an open challenge dataset for financial sentiment analysis (Financial Opinion Mining and Question Answering). The results show that the domain-specific FinBERT outperforms the generic BERT models in all of these financial datasets. An improved financial domain LM was released 2022 by Shah et al. [2022] by introducing FLANG-BERT, the Financial LANGuage Model. They also created a domain-specific benchmark, Financial Language Understanding Evaluation (FLUE). Recently in May 2023, Bloomberg announced the BloombergGPT model, a Large Language Model (LLM) for the financial domain [Wu et al., 2023]. However, next to some experience on the training process no datasets, benchmarks, or weights have been released publicly.

Numerous other domain-specific LMs have been created since the rise of BERT. They all outperform general-purpose LMs. For instance, SciBERT is a pretrained LM based on scientific publications and evaluated on a suite of tasks in difference scientific domains [Beltagy et al., 2019]. ConfliBERT is built to improve monitoring of political violence and conflicts [Hu et al., 2022] and PoliBERTweet is used to analyze political content on Twitter [Kawintiranon and Singh, 2022]. Cybersecurity is another important area thus Aghaei et al. [2023] pretrained a M on a large corpus of cybersecurity text. To improve IR tasks in the architecture, engineering, and construction (AEC) industry, Zheng et al. [2022] pretrained BERT on a corpus of regulatory text. Also, the domain-specific model BlueBERT [Peng et al., 2019b] from the biomedical domain has been further pretrained and evaluated on more narrow, cancer-related vocabularies, resulting in CancerBERT [Zhou et al., 2022].

In the legal domain Chalkidis et al. [2020] pretrained LegalBERT on EU and UK legislation, ECHR and US cases and US contracts. Zheng et al. [2021] pretrained CaseHoldBERT on US caselaw. Henderson et al. [2022] trained PoL-BERT on their 256 GB diverse Pile of Law corpus. Niklaus and Giofré [2022] pretrained longformer models using the Replaced Token Detection (RTD) task Clark et al. [2020] on the Pile of Law. Hua et al. [2022] pretrained reformer models with RTD on 6 GB of US caselaw. Finally, Niklaus et al. [2023b] released a large multilingual legal corpus and trained various legal models on it.

F.2 Judgment Prediction

The domain of Legal Judgment Prediction (LJP) centers around the crucial task of predicting legal case outcomes given the provided facts. In the landscape of LJP research, there have been significant advances focusing on diverse languages, jurisdictions, and input types. Researchers have utilized a variety of datasets, each with their unique characteristics and annotations, to analyze and predict case outcomes [Feng et al., 2022, Aletras et al., 2016, Şulea et al., 2017, Medvedeva et al., 2018, Chalkidis et al., 2019a].

In the context of Chinese criminal cases, notable efforts have been made by Xiao et al. [2018, 2021], where they utilized the CAIL2018 dataset, which consists of over 2.6M cases and provides annotations for Law Article, Charge, and Prison Term, among others.

Focusing on the Indian and Swiss jurisdictions, Malik et al. [2021] and Niklaus et al. [2021, 2022] employed the ILDC and SJP datasets respectively, both using binary labels. The ILDC dataset, with over 34K Indian Supreme Court cases, offers sentence-level explanations along with Court Decision annotations, while the SJP corpus is trilingual, containing judgments from Switzerland in German, French, and Italian, and provides annotations like the publication year, legal area, and the canton of origin.

European jurisdictions have been explored using the ECHR2019 and ECHR2021 datasets [Chalkidis et al., 2019a, 2021c]. These corpora feature cases from the European Court of Human Rights, annotated for Violation, Law Article, and Alleged Law Article, among others, with the latter also providing paragraph-level rationales.

The FCCR dataset, containing over 126K cases from France, has been used to predict Court Decisions with different setups, offering additional annotations such as the date of the court ruling and the law area [Şulea et al., 2017].

Recently, Semo et al. [2022] introduced a new perspective on LJP, applying it to US class action cases. The proposed task involves predicting the judgment outcome based on the plaintiff's pleas, further expanding the scope of LJP research and making the task more realistic.

These efforts underscore the breadth of LJP research, demonstrating its applicability across multiple jurisdictions, languages, and legal systems, and its potential in assisting legal professionals and enhancing access to justice.

F.3 Criticality Prediction

Chalkidis et al. [2019a] introduced the Importance Prediction task, which predicts the importance of a ECtHR case on a scale from 1 (key case) to 4 (unimportant). Legal experts defined and assigned these labels for each case, representing a significant contrast to our approach where labels were algorithmically determined. This is to our knowledge the only comparable task to Criticality Prediction.

F.4 Law Area Prediction

Although not widely studied, several notable works have focused on LAP. Şulea et al. [2017] worked on the Law Judgment Prediction (LJP) task, using a dataset of over 126K cases from the French Supreme Court. The study used Linear Support Vector Machines (SVMs) to classify cases into one of eight law areas, using the entire case description as input. This approach yielded an F1 score of 90%. Soh et al. [2019] conducted a similar study using a dataset of 6K judgments in English from the Singapore Supreme Court. These judgments were mapped into 30 law areas. Several text classifiers were used in the study, achieving a macro F1 score of up to 63.2%.

F.5 Court View Generation

Over the past decade, text generation in the field of Legal NLP has been underexplored [Katz et al., 2023b], especially in comparison to tasks such as classification and information extraction. Li and

Zhang [2021b] utilize Chinese case facts, as well as charge (formal accusations of crimes) and law article information, to generate court opinions. A key difference from our task is the shortness of their opinions (avg. 31/34 tokens), while ours span approximately four thousand tokens on average. With the emergence of powerful generative models, we expect a surge in research activity in this area, necessitating challenging benchmarks to assess progress effectively.

F.6 Leading Decisions Summarization

In the field of legal text summarization, several noteworthy contributions have been made [Grover et al., 2004, Hachey and Grover, 2006, Kim et al., 2013, Jain et al., 2021], with the BillSum [Kornilova and Eidelman, 2019] and Multi-LexSum [Shen et al., 2022] datasets being particularly significant. The creators of the BillSum dataset focused on summarizing 22K bills from the US Congress and the state of California. They also applied transfer learning in summarization from federal to state laws. Models based on BERT and TF-IDF, as well as a combination of both, have been evaluated on this dataset. The BillSum dataset focuses on English language documents related to the US legislative environment. The Multi-LexSum dataset is another significant development in the area of legal text summarization. It targets long civil rights lawsuits, with an average length of over 75K words. This 9K-document dataset allows for in-depth study at different summary lengths: short (25 words), medium (130 words), and long (650 words), a unique feature of the Multi-LexSum dataset. Models based on BART Lewis et al. [2020] and PEGASUS Zhang et al. [2020a] were evaluated on this dataset. Like BillSum, the Multi-LexSum dataset is primarily for the English language and is relevant to the US legal setting.

F.7 Citation Extraction

Early work from Martínez-González et al. [2005] extract citations from legal text with patterns. Nambanoor Kunnath et al. [2022] studied the effect of differing context size for citation classification in scientific text. Taylor et al. [2022] considered the more difficult Citation Prediction task on scientific text and found that larger models are more true to the real citation distribution, whereas smaller models tend to output the most frequent citations most of the times.

F.8 Information Retrieval

Lawrie et al. [2023] revisited the challenges of multilingual IR and proposed neural approaches to address this issue. They demonstrated that combining neural document translation with neural ranking resulted in the best performance in their experiments conducted on the MS MARCO dataset Bajaj et al. [2018]. However, this approach is computationally expensive. To mitigate this issue, they showed that using a pre-trained XLM-R multilingual model to index documents in their native language resulted in only a two percent difference in effectiveness. XLM-R is a transformer-based masked language model that employs self-supervised training techniques for cross-lingual understanding Conneau et al. [2020a]. Lawrie et al. [2023] crucially utilized mixed-language batches from the neural translation of MS MARCO passages.

A widely used technique is BM-25, which is an improved retrieval method that considers the term frequencies and takes into account the saturation effect and document length Robertson and Zaragoza [2009]. The saturation effect refers to the point where the relevance of a term stops increasing, even if it appears many times in a document. This issue is mitigated through the use of an additional parameter, k. Additionally, longer documents are more likely to contain a higher number of occurrences of a term simply because they contain more words, not necessarily because the term is more relevant to the document, which is why parameter b is used. The BM-25 score is calculated using the Inverse Document Frequency (IDF), Term Frequency (TF), queries Q, documents d, and term t.

$$BM25(d, Q, b, k) = \sum_{t \in Q} IDF(t) \frac{(k+1)TF(t,d)}{(1-b)(b*A)+TF(t,d)}$$

Chalkidis et al. [2021b] proposed a new IR task called REG-IR, which deals with longer documents in the corpus and entire documents as queries. This task is an adaptation of Document-to-Document (Doc2Doc) IR, which aims to identify a relevant document for a given document. The authors observed that neural re-rankers underperformed due to contradicting supervision, where similar query-document pairs were labeled with opposite relevance. Additionally, they demonstrated for long documents that using BM25 as a document retriever in a two-stage approach often results in underperformance since the parameters k and b are often not optimal when using standard values. The problem of noise filtering of long documents was also addressed by using techniques like stopwords removal. However, as seen in Leveling [2012], this approach can have a negative effect on performance. The best pre-fetcher for long documents was found to be C-BERTs Chalkidis et al. [2021a], which are trained on classifying documents using predefined labels.

Thakur et al. [2021] proposed a novel evaluation benchmark for IR that encompasses a wide range of approaches, including BM25, dense, and re-ranking models. They found that while BM25 is computationally expensive, it provides a robust baseline, whereas other models failed to achieve comparable performance. Their findings suggest that there is still much room for improvement in this area of NLP. Efficient retrieval of relevant information is crucial for many NLP tasks, and these results highlight the need for continued research in this area.

G More Detailed Experimental Setup

G.1 Pretraining Legal Models

(a) We warm-start (initialize) our models from the original XLM-R checkpoints (base or large) of Conneau and Lample [2019]. Model recycling is a standard process followed by many Wei et al. [2021], Ouyang et al. [2022] to benefit from starting from an available "well-trained" PLM, rather from scratch (random). XLM-R was trained on 2.5TB of cleaned CommonCrawl data in 100 languages.

(b) We train a new tokenizer of 128K BPEs on the training subsets to better cover legal language across languages. However, we reuse the original XLM-R embeddings for all lexically overlapping tokens Pfeiffer et al. [2021], i.e., we warm-start word embeddings for tokens that already exist in the original XLM-R vocabulary, and use random ones for the rest.

(c) We continue pretraining our models on our pretraining corpus with batches of 512 samples for an additional 1M/500K steps for the base/large model. We do initial warm-up steps for the first 5% of the total training steps with a linearly increasing learning rate up to $1e-4$, and then follow a cosine decay scheduling, following recent trends. For half of the warm-up phase (2.5%), the Transformer encoder is frozen, and only the embeddings, shared between input and output (MLM), are updated. We also use an increased 20/30% masking rate for base/large models respectively, where also 100% of the predictions are based on masked tokens, compared to Devlin et al. [2019a][6], based on the findings of Wettig et al. [2023].

(d) For both training the tokenizer and our legal models, we use a sentence sampler with exponential smoothing of the sub-corpora sampling rate following Conneau and Lample [2019] and Raffel et al. [2020], since there is a disparate proportion of tokens across languages (Figure 7) and we aim to preserve per-language capacity, i.e., avoid overfitting to the majority (almost 50% of the total number of texts) German texts.

(e) We consider mixed cased models, i.e., both upper- and lowercase letters covered, similar to all recently developed large PLMs Conneau and Lample [2019], Raffel et al. [2020], Brown et al. [2020].

(f) To better account for long contexts often found in legal documents, we continue training the base-size multilingual model on long contexts (4096 tokens) with windowed attention (128 tokens window size) Beltagy et al. [2020] for 50K steps, dubbing it Legal-Swiss-LF-base. We use the standard 15% masking probability and increase the learning rate to $3e-5$ before decaying but otherwise use the same settings as for training the small-context models.

G.2 Resources Used

The experiments were performed on internal university clusters on NVIDIA GPUs with the following specifications: 24GB RTX3090, 32GB V100, 48GB A6000, and 80GB A100. We used an approximate total of 160, 20, and 2 GPU days for the text classification, text generation and information retrieval experiments.

Text Generation For inference and fine-tuning LLaMA-2 in the text generation task, we used the Together API.

G.3 Hyperparameters

Text Classification For all models and datasets, a learning rate of 1e-5 was used without any tuning. Each experiment was executed with three random seeds (1-3), and the batch size was tailored for each task and corresponding computational resource. If the GPU memory was inadequate, gradient accumulation was employed as a workaround to arrive at a final batch size of 64. The training was conducted with early stopping based on validation loss, maintaining a patience level of 5 epochs. Due to the considerable size of the judgment prediction dataset and the extended duration of the experiment, training was limited to a single epoch with evaluations after every 1000th step. To reduce costs, we utilized AMP mixed precision during the training and evaluation phases whenever it did not

[6] Devlin et al. [2019a] – and many other follow-up work – used a 15% masking ratio, and a recipe of 80/10/10% of predictions made across masked/randomly-replaced/original tokens.

lead to overflows (e.g., mDeBERTa-v3). We established the max-sequence-length (determined by the product of max-segment-length and max-segments in the hierarchical setup Aletras et al. [2016], Niklaus et al. [2021, 2022]) based on whether we used Facts: 2048 (128 X 16), or Considerations: 4096 (128 X 32).

Text Generation For the main CVG dataset, we trained our mT5 models for only one epoch (because of the large training set) with a final batch size of 16, using gradient accumulation as needed. We performed evaluations every 1000 steps. For the smaller origin dataset, we increased the number of epochs to 100 and evaluated every 100 steps. For the LDS task, we adjusted the training to 10 epochs.

Information Retrieval For the BM25 model, we used the same parameters as used in the BEIR paper Thakur et al. [2021], chosen were k = 0.9 and b = 0.4. For the SBERT model training, we employed the BEIR toolkit Thakur et al. [2021]. Our training process was constrained by a maximum sequence length of 512 tokens. During the training phase, we completed a single epoch, comprising 5000 evaluation steps.

In the context of training with hard negative examples, we incorporated 5 negative examples for every query. The selection of these examples was based on the 5 highest-ranked erroneous predictions generated by the BM25 model. To facilitate training with these challenging negatives, we followed the guidelines provided by Thakur et al. [2021], utilizing the Hardnegs template.

```
"corpus": {
    "decision_id_bge": {
        "title": "file number",
        "text": "facts and considerations of a case"
    },
    "law_id": {
        "title": "",
        "text": "text of a law"
    }
},
"queries": {
    "decision_id_bger_1": "facts of a case",
    "decision_id_bger_2": "facts"
},
"qrels": {
    "decision_id_bger_1": {"law-id": 1, "decision-id-bge": 1},
    "decision_id_bger_2": {"law-id": 1}
 }
```

Figure 3: Structure of corpus, queries and qrels for IR task

H Additional Results

H.1 Information Retrieval

Table 8: Results IR: using a subset of 100 queries and only relevant documents in the corpus resulting in an easier task

Model	**Additional**	$Rcap@1\uparrow$	$Rcap@10\uparrow$	$Rcap@100\uparrow$	$NDCG@1\uparrow$	$NDCG@10\uparrow$	$NDCG@100\uparrow$
Train + Evaluate S-BERT	sbert-legal-xlm-roberta-base	32.32	32.34	81.77	32.32	30.89	49.11
Train + Evaluate S-BERT	sbert-legal-swiss-roberta-base	**36.36**	**35.68**	76.03	**36.36**	**34.54**	49.90
Train + Evaluate S-BERT	distiluse-base-multilingual-cased-v1	22.22	30.35	84.38	22.22	25.72	48.66
Evaluate S-BERT	distiluse-base-multilingual-cased-v1	8.08	11.83	43.35	8.08	10.55	21.56
Train(HN) + Evaluate S-BERT	distiluse-base-multilingual-cased-v1	27.27	33.94	**86.81**	27.27	30.09	**52.03**
Dim Reduction	distiluse-base-multilingual-cased-v1	0.00	1.59	5.43	0.00	1.17	2.41
Cross Encoder	distiluse-base-multilingual-cased-v1	5.94	8.04	14.20	2.97	1.84	7.35
Lexical		5.94	8.04	14.20	5.94	8.52	10.64
ML Lexical	'German'	9.90	8.41	15.19	9.90	9.14	11.58

Table 9: Results IR Additional: Results IR Abbreviations: Capped **R**ecall, **N**DCG, **dist**iluse-base-multilingual-cased-v1, joelito/swiss-legal-**rob**erta-base, joelito/legal-**xlm**-roberta-base, **T**rain, **H**ard **N**egative, **E**valuate, **S**-**B**ert, **Dim** Reduction

Model		**Adaption**	$R@1\uparrow$	$R@10\uparrow$	$R@100\uparrow$	$N@1\uparrow$	$N@10\uparrow$	$N@100\uparrow$
LR			8.38	6.43	15.76	8.38	6.66	10.23
LR		S	**10.64**	7.57	16.47	**10.64**	8.04	11.33
LR		SL	7.91	**9.99**	**32.46**	9.13	**9.65**	**18.03**
MLR	'German'		8.69	6.54	15.99	8.69	6.82	10.43
MLR	'German'	S	**10.88**	7.65	16.80	**10.88**	8.14	11.53
MLR	'German'	SL	8.05	**9.94**	**32.63**	9.30	**9.70**	**18.17**
MLR	'French'		**11.37**	**7.74**	**16.54**	**11.37**	**8.34**	**11.51**
MLR	'French'	S	10.97	7.60	16.52	10.97	8.14	11.41
MLR	'Italian'		10.08	7.118	16.294	10.08	7.582	11.021
MLR	'English'		8.38	6.43	15.76	8.38	6.66	10.23
T+E SB	xlm		2.77	2.58	10.17	6.36	5.66	12.03
T+E SB	rob		3.97	3.47	12.28	9.12	7.76	15.16
E SB	dist		0.90	0.75	2.64	2.06	1.70	3.31
T+E SB	dist		4.4	3.92	12.64	10.11	8.76	16.16
T+E SB	dist	S	4.69	4.14	13.39	10.77	9.27	17.05
T+E SB	dist	SL	1.79	3.92	14.17	4.03	6.17	12.91
SB T(HN)+E	dist		3.97	4.46	13.36	9.12	9.21	16.87
SB T(HN)+E	dist	S	3.76	4.75	12.80	8.64	9.66	16.57
SB T(HN)+E	dist	SL	2.34	4.37	14.43	5.27	6.99	13.75
T+E SB	dist	DE	4.22	4.49	15.21	8.21	8.15	15.86
T+E SB	dist	DE SL	4.06	8.47	29.43	4.51	6.73	13.78
T+E SB	dist	FR	1.88	2.2	9.19	5.77	6.22	13.94
T+E SB	dist	FR SL	2.69	5.68	27.28	3.0	4.59	11.11
T+E SB	dist	IT	0.22	0.24	0.79	5.43	5.74	11.44
T+E SB	dist	IT SL	1.71	4.54	16.24	1.91	3.38	6.83
Dim	dist		0.71	0.62	2.42	1.64	1.4	2.95

For the ML Lexical Retrieval model a main language must be chosen, indicated with German, French and Italian. Dataset adaptions are indicated with: (S) stopword removal, (SL) using only single language links, (DE/FR/IT) using only queries in one language. Table 8 shows the results of the IR task on a subset of 100 queries and with only relevant documents while Table 9 shows more detailed results using all queries.

H.2 Court View Generation

Table 10 shows the results of the CVG task from both, the main and the origin dataset. Table 12 presents the CVG task results split by language, detailing scores for German, French, and Italian.

H.3 Leading Decision Summarization (LDS)

Table 13 shows all results of the LDS task. Table 14 presents the LDS task results split by language, detailing scores for German, French, and Italian.

Table 10: Results of Court View Generation task. 'In Len' denotes input length in tokens. **Bold**: best within model; underlined: best overall.

Model	In Len ↑	Main Scores ↑				Origin Scores ↑			
		BERT	**BLEU**	**MET**	**R1 / R2 / RL**	**BERT**	**BLEU**	**MET**	**R1 / R2 / RL**
$mT5_{Large}$	2048	**75.74**	**66.92**	**34.44**	**34.91** / **15.58** / **33.53**	76.24	**62.59**	32.25	34.80 / 16.11 / 33.58
$mT5_{Large}$	1024	75.56	66.68	34.02	34.26 / 14.72 / 32.87	74.99	58.35	31.06	33.35 / 14.80 / 32.16
$mT5_{Large}$	512	75.27	66.12	33.48	33.61 / 14.26 / 32.21	**76.33**	62.08	**32.92**	**36.61 / 18.17 / 34.84**
$mT5_{Base}$	2048	75.01	65.48	32.89	33.23 / 13.57 / 31.89	75.99	**63.39**	**34.15**	36.48 / **18.81** / 35.58
$mT5_{Base}$	1024	**75.15**	**65.73**	**33.15**	**33.49 / 13.96 / 32.18**	76.07	60.99	33.50	**37.68** / 18.79 / **36.58**
$mT5_{Base}$	512	74.89	65.55	32.66	32.66 / 13.16 / 31.35	**76.08**	62.21	32.80	36.40 / 17.58 / 34.98
$mT5_{Small}$	2048	**74.13**	**63.97**	**30.96**	**31.29 / 11.01 / 29.90**	75.23	56.59	30.71	34.68 / 13.64 / 33.24
$mT5_{Small}$	1024	74.00	63.70	30.68	31.05 / 10.77 / 29.64	**75.75**	58.99	31.17	34.62 / 14.25 / **33.91**
$mT5_{Small}$	512	73.92	63.83	30.57	30.58 / 10.35 / 29.20	75.63	**61.12**	**32.33**	**35.16 / 14.45** / 33.72

Table 11: Results of Court View Generation task. 'In Len' denotes input length in tokens. **Bold**: best within setup; underlined: best overall. (*) These models were fine-tuned on only 1'000 samples for 3 epochs. All models, except for the mT5 models, were evaluated on the validation set.

Model	**Setup**	**In Len ↑**	**BERT ↑**	**BLEU ↑**	**MET ↑**	**R1 / R2 / RL ↑**
$mT5_{Large}$	Fine-tuned	2048	**75.74**	**66.92**	**34.44**	**34.91 / 15.58 / 33.53**
$mT5_{Large}$	Fine-tuned	1024	75.56	66.68	34.02	34.26 / 14.72 / 32.87
$mT5_{Large}$	Fine-tuned	512	75.27	66.12	33.48	33.61 / 14.26 / 32.21
$mT5_{Base}$	Fine-tuned	2048	75.01	65.48	32.89	33.23 / 13.57 / 31.89
$mT5_{Base}$	Fine-tuned	1024	75.15	65.73	33.15	33.49 / 13.96 / 32.18
$mT5_{Base}$	Fine-tuned	512	74.89	65.55	32.66	32.66 / 13.16 / 31.35
$mT5_{Small}$	Fine-tuned	2048	74.13	63.97	30.96	31.29 / 11.01 / 29.90
$mT5_{Small}$	Fine-tuned	1024	74.00	63.70	30.68	31.05 / 10.77 / 29.64
$mT5_{Small}$	Fine-tuned	512	73.92	63.83	30.57	30.58 / 10.35 / 29.20
GPT-3.5-Turbo	Fine-tuned*	2048	72.31	62.23	28.08	26.06 / 7.19 / 24.54
LLaMA-2-13B Chat	Fine-tuned*	2048	**74.22**	**63.51**	**33.33**	**34.36 / 16.68 / 33.20**
LLaMA-2-7B Chat	Fine-tuned*	2048	73.27	62.34	31.6	32.31 / 14.47 / 31.38
GPT-4	1-shot	2048	70.39	59.69	24.63	23.87 / 4.64 / 22.32
GPT-3.5-Turbo-16K	1-shot	8192	**70.86**	59.89	**25.63**	24.97 / 5.44 / 23.50
GPT-3.5-Turbo-16K	1-shot	2048	70.73	59.92	25.55	24.95 / 5.43 / 23.49
Claude 2	1-shot	8192	69.71	60.92	24.88	24.92 / **5.55 / 23.61**
Claude 2	1-shot	2048	69.45	**61.26**	24.85	**24.98** / 5.39 / **23.61**
Claude Instant	1-shot	8192	67.5	58.18	23	23.91 / 4.32 / 22.58
Claude Instant	1-shot	2048	67.49	57.76	22.84	23.78 / 4.23 / 22.43
GPT-4	0-shot	2048	69.41	58.16	23.25	22.61 / 3.95 / 21.10
GPT-3.5-Turbo-16K	0-shot	8192	67.93	56.87	22.62	21.21 / 3.56 / 19.88
GPT-3.5-Turbo-16K	0-shot	2048	67.80	56.52	22.32	20.99 / 3.46 / 19.74
Claude 2	0-shot	8192	**70.22**	**62.38**	**25.74**	**25.66 / 6.06 / 24.27**
Claude 2	0-shot	2048	69.46	61.36	24.90	24.66 / 5.30 / 23.24
Claude Instant	0-shot	8192	66.94	59.02	22.95	23.16 / 4.04 / 21.82
Claude Instant	0-shot	2048	66.50	58.78	22.65	22.65 / 3.80 / 21.32
LLaMA-2-70B Chat	0-shot	2048	66.78	53.04	19.13	18.48 / 3.21 / 17.35
LLaMA-2-13B Chat	0-shot	2048	67.23	55.01	20.18	19.76 / 3.26 / 18.57
LLaMA-2-7B Chat	0-shot	2048	63.74	40.62	11.29	10.75 / 1.62 / 10.13

H.4 Text Classification

Table 15 shows more detailed results on the text classification datasets including standard deviations across seeds.

H.4.1 Language specific results

Table 16 shows more detailed results on the text classification datasets language specific scores.

SwissBERT, where pretraining tokens were most focused towards the dominant language German seems to have quite even results, with scores in Italian even being the highest. Models trained on CC100 (MiniLM, mDeBERTa, XLM-R and X-MOD) showed mixed results. For all models, German performance was very close to French performance. MiniLM, mDeBERTa, and X-MOD showed

Table 12: Results of the CVG task split by language. Scores are presented in the order: German, French, and Italian.

Model	**Setup**	**BERT ↑**	**BLEU ↑**	**MET ↑**	**R1 ↑**	**R2 ↑**	**RL ↑**
GPT-3.5-Turbo	Fine-tuned	71.89 / 73.01 / 71.17	62.29 / 62.15 / 62.29	29.02 / 27.43 / 25.38	25.68 / 26.97 / 23.48	7.31 / 7.49 / 4.64	24.60 / 24.94 / 21.77
LLaMA-2-13B-Chat	Fine-tuned	75.06 / 73.76 / 71.03	66.24 / 61.05 / 58.81	36.44 / 30.69 / 26.95	36.22 / 33.37 / 27.12	19.51 / 14.55 / 9.44	35.35 / 31.92 / 25.63
GPT-4	1-shot	69.73 / 71.31 / 69.22	58.83 / 60.89 / 58.23	24.20 / 25.48 / 21.76	22.40 / 25.86 / 21.85	3.56 / 6.15 / 2.88	21.26 / 23.82 / 20.35
Claude-2	1-shot	68.47 / 71.46 / 67.40	60.98 / 60.79 / 61.48	24.90 / 25.13 / 22.67	23.56 / 26.65 / 23.88	4.61 / 6.79 / 4.58	22.65 / 24.88 / 22.46
GPT-4	0-shot	69.12 / 69.87 / 68.54	57.90 / 58.58 / 57.18	23.25 / 23.42 / 21.83	21.48 / 24.14 / 21.08	2.97 / 5.26 / 2.78	20.25 / 22.32 / 19.44
Claude-2	0-shot	67.75 / 71.68 / 67.75	60.82 / 61.77 / 61.39	24.59 / 25.45 / 22.57	23.23 / 27.29 / 23.13	4.29 / 6.86 / 3.99	22.26 / 25.46 / 21.47
LLaMA-2-13B-Chat	0-shot	66.72 / 67.94 / 66.65	53.20 / 56.73 / 57.61	19.86 / 20.60 / 19.98	18.22 / 21.50 / 20.44	2.18 / 4.59 / 3.03	17.39 / 19.92 / 18.95

Table 13: Results of Leading Decision Summarization (LDS) task. 'In Len' denotes input length in tokens. **Bold**: best within setup; <u>underlined</u>: best overall. All models, except for the mT5 models, were evaluated on the validation set.

Model	**Setup**	**In Len ↑**	**BERT ↑**	**BLEU ↑**	**MET ↑**	**R1 / R2 / RL ↑**
$mT5_{Large}$	Fine-tuned	4096	*OOM*	*OOM*	*OOM*	*OOM*
$mT5_{Large}$	Fine-tuned	2048	73.10	27.21	21.88	31.47 / 12.22 / 29.94
$mT5_{Large}$	Fine-tuned	512	70.67	26.89	18.31	24.76 / 6.15 / 23.48
$mT5_{Base}$	Fine-tuned	4096	**73.33**	**30.81**	**23.50**	<u>**32.43**</u> / <u>**12.78**</u> / <u>**30.87**</u>
$mT5_{Base}$	Fine-tuned	2048	72.45	30.13	21.94	30.09 / 10.79 / 28.71
$mT5_{Base}$	Fine-tuned	512	70.60	27.10	18.31	24.72 / 6.15 / 23.55
$mT5_{Small}$	Fine-tuned	4096	72.04	28.68	21.29	29.61 / 10.31 / 28.12
$mT5_{Small}$	Fine-tuned	2048	71.38	24.64	19.28	27.88 / 9.19 / 26.54
$mT5_{Small}$	Fine-tuned	512	69.66	20.73	15.95	22.91 / 5.36 / 21.85
GPT-4	1-shot	4096	<u>**73.55**</u>	47.75	34.72	30.82 / 9.68 / 28.89
GPT-3.5-Turbo-16K	1-shot	8192	72.92	46.15	33.68	29.69 / 9.47 / 27.97
GPT-3.5-Turbo-16K	1-shot	4096	72.89	45.21	32.76	29.69 / 9.25 / 27.94
Claude 2	1-shot	16384	73.21	<u>**49.79**</u>	<u>**35.58**</u>	**31.19 / 9.76 / 29.44**
Claude 2	1-shot	4096	72.91	47.55	33.57	30.28 / 9.12 / 28.58
Claude Instant	1-shot	16384	72.23	46.29	31.05	28.26 / 8.78 / 26.42
Claude Instant	1-shot	4096	72.44	44.80	30.29	27.89 / 8.56 / 26.18
GPT-4	0-shot	4096	**71.56**	48.35	32.97	26.52 / **8.93** / 24.51
GPT-3.5-Turbo-16K	0-shot	4096	70.28	46.08	30.60	25.18 / 7.58 / 23.59
Claude 2	0-shot	16384	71.45	**49.29**	**33.51**	**28.61** / 8.84 / **26.59**
Claude 2	0-shot	4096	71.13	49.20	32.54	27.70 / 8.39 / 25.90
Claude Instant	0-shot	16384	70.64	45.19	28.92	26.23 / 7.68 / 24.28
Claude Instant	0-shot	4096	71.33	45.65	29.22	26.13 / 8.16 / 24.15

Italian underperformance whereas XLM-R showed very strong performance in Italian, especially the large variant. Even though the Legal-ch-R models are based on XLM-R, they show underperformance in Italian, but similar performance between French and German. mT5 models performed well in French, the base variant additionally also performed well on Italian. BLOOM was much better in French than in other languages, not surprising given it did not have German and Italian in the pretraining data.

Overall, there only seems to be a weak trend connecting higher percentage of a given language in the pretraining corpus leading to better downstream results in that language.

Table 14: Results of the LDS task split by language. Scores are presented in the order: German, French, and Italian.

Model	Setup	BERT ↑	BLEU ↑	MET ↑	R1 ↑	R2 ↑	RL ↑
GPT-4	1-shot	73.89 / 72.84 / 74.08	49.32 / 45.85 / 38.08	36.50 / 32.01 / 28.75	31.14 / 30.03 / 32.43	9.91 / 9.13 / 10.82	29.25 / 27.99 / 30.84
GPT-3.5	1-shot	73.26 / 72.21 / 72.50	47.44 / 41.50 / 39.08	35.23 / 28.48 / 27.37	30.64 / 28.22 / 26.11	9.46 / 8.87 / 8.96	28.90 / 26.46 / 24.34
Claude-2	1-shot	73.17 / 72.34 / 73.69	48.62 / 45.50 / 47.68	35.07 / 30.69 / 33.91	30.13 / 30.30 / 32.57	9.29 / 8.52 / 11.50	28.39 / 28.71 / 30.86
Claude-Instant	1-shot	72.97 / 71.51 / 71.44	45.66 / 43.40 / 42.06	32.02 / 27.20 / 27.40	28.40 / 27.17 / 25.58	8.80 / 8.12 / 8.28	26.71 / 25.36 / 24.20
GPT-4	0-shot	72.65 / 69.56 / 70.95	49.17 / 46.96 / 46.89	36.26 / 27.39 / 27.10	28.27 / 23.45 / 24.25	10.88 / 5.36 / 7.74	26.26 / 21.42 / 22.51
GPT-3.5	0-shot	71.25 / 68.39 / 69.84	47.39 / 43.68 / 44.17	33.31 / 25.51 / 27.91	27.11 / 21.73 / 21.86	9.22 / 4.64 / 4.75	25.47 / 20.22 / 20.48
Claude-2	0-shot	71.74 / 69.81 / 72.77	50.21 / 46.94 / 52.35	34.72 / 28.09 / 35.60	28.93 / 25.02 / 31.04	9.27 / 6.38 / 11.55	27.11 / 23.22 / 29.45
Claude-Instant	0-shot	71.99 / 70.08 / 70.81	46.82 / 43.59 / 43.33	31.41 / 25.06 / 27.55	27.43 / 23.60 / 25.63	9.17 / 6.11 / 8.48	25.48 / 21.57 / 23.63

Table 15: Configuration aggregate scores with standard deviations on the test set. The macro-F1 scores are provided.

Model	CPB-F	CPB-C	CPC-F	CPC-C	SLAP-F	SLAP-C	JP-F	JP-C	Agg.
MiniLM	54.7$_{+/-1.9}$	65.8$_{+/-1.6}$	9.8$_{+/-2.8}$	20.8$_{+/-3.0}$	59.7$_{+/-3.8}$	61.1$_{+/-3.7}$	58.1$_{+/-0.4}$	78.5$_{+/-2.3}$	32.4
DistilmBERT	56.2$_{+/-0.5}$	65.4$_{+/-1.7}$	19.6$_{+/-1.1}$	22.1$_{+/-0.4}$	63.7$_{+/-11.7}$	65.9$_{+/-6.4}$	59.9$_{+/-0.9}$	75.5$_{+/-3.3}$	42.1
mDeBERTa-v3	55.1$_{+/-2.0}$	69.8$_{+/-2.8}$	21.0$_{+/-3.6}$	17.5$_{+/-4.4}$	63.8$_{+/-6.3}$	59.3$_{+/-7.6}$	60.6$_{+/-0.9}$	77.9$_{+/-2.6}$	40.2
XLM-R$_{\text{Base}}$	57.2$_{+/-1.5}$	65.9$_{+/-3.2}$	21.3$_{+/-1.5}$	23.7$_{+/-1.9}$	67.2$_{+/-15.9}$	73.4$_{+/-2.5}$	60.9$_{+/-0.6}$	79.7$_{+/-2.5}$	44.6
XLM-R$_{\text{Large}}$	56.4$_{+/-1.8}$	67.9$_{+/-1.9}$	24.4$_{+/-7.2}$	29.1$_{+/-2.7}$	65.1$_{+/-8.5}$	78.9$_{+/-4.6}$	60.8$_{+/-0.6}$	80.9$_{+/-2.4}$	48.6
X-MOD$_{\text{Base}}$	56.6$_{+/-1.8}$	67.8$_{+/-2.9}$	20.0$_{+/-3.0}$	20.6$_{+/-3.5}$	63.9$_{+/-10.1}$	64.4$_{+/-7.0}$	60.5$_{+/-0.6}$	79.1$_{+/-2.6}$	41.9
SwissBERT$_{\text{(xlm-vocab)}}$	56.9$_{+/-0.7}$	67.3$_{+/-4.7}$	25.7$_{+/-8.3}$	23.0$_{+/-4.0}$	61.5$_{+/-9.5}$	73.2$_{+/-2.1}$	61.4$_{+/-0.6}$	79.4$_{+/-2.5}$	46.1
mT5$_{\text{Small}}$	52.2$_{+/-1.9}$	62.1$_{+/-5.2}$	13.2$_{+/-2.4}$	17.9$_{+/-1.7}$	53.1$_{+/-13.8}$	60.9$_{+/-15.9}$	58.9$_{+/-1.0}$	74.2$_{+/-3.6}$	34.4
mT5$_{\text{Base}}$	52.1$_{+/-1.6}$	61.5$_{+/-3.9}$	14.0$_{+/-2.8}$	19.7$_{+/-1.6}$	58.4$_{+/-17.2}$	61.8$_{+/-16.8}$	54.5$_{+/-1.5}$	72.0$_{+/-3.1}$	35.9
BLOOM-560m	53.0$_{+/-1.7}$	61.7$_{+/-4.1}$	10.7$_{+/-3.7}$	8.0$_{+/-3.5}$	52.6$_{+/-10.7}$	53.2$_{+/-8.5}$	60.5$_{+/-0.7}$	73.4$_{+/-7.2}$	24.9
Legal-ch-R$_{\text{Base}}$	57.7$_{+/-1.6}$	70.5$_{+/-2.3}$	16.2$_{+/-5.8}$	20.1$_{+/-5.6}$	77.0$_{+/-3.6}$	79.7$_{+/-0.9}$	64.0$_{+/-1.3}$	86.4$_{+/-1.9}$	40.9
Legal-ch-R$_{\text{Large}}$	55.9$_{+/-2.2}$	68.9$_{+/-2.1}$	25.8$_{+/-7.8}$	16.3$_{+/-8.7}$	76.9$_{+/-2.3}$	84.9$_{+/-9.7}$	62.8$_{+/-0.9}$	87.1$_{+/-2.2}$	43.3
Legal-ch-LF$_{\text{Base}}$	58.1$_{+/-2.1}$	70.8$_{+/-2.9}$	21.4$_{+/-2.9}$	17.4$_{+/-8.6}$	80.1$_{+/-12.7}$	77.1$_{+/-4.8}$	65.4$_{+/-1.7}$	86.4$_{+/-1.8}$	42.5

Table 16: Configuration aggregate scores. The macro-F1 scores from the language-specific subsets of the test set are provided.

Model	CPB-F	CPB-C	CPC-F	CPC-C	SLAP-F	SLAP-C	JP-F	JP-C	Agg.
Languages	de / fr / it	de / fr / it	de / fr / it	de / fr / it	de / fr / it	de / fr / it	de / fr / it	de / fr / it	de / fr / it
MiniLM	57.5 / 53.9 / 52.9	68.1 / 65.4 / 64.2	12.1 / 13.1 / 6.8	24.6 / 21.9 / 17.3	55.3 / 60.0 / 64.5	57.6 / 60.0 / 66.5	57.7 / 58.1 / 58.7	77.8 / 81.7 / 76.1	36.1 / 36.6 / 26.6
DistilmBERT	56.3 / 55.6 / 56.8	67.8 / 63.9 / 64.7	20.2 / 18.2 / 20.7	22.6 / 21.6 / 22.2	50.9 / 67.2 / 79.6	57.8 / 68.8 / 72.9	60.5 / 60.7 / 58.6	75.6 / 79.8 / 71.7	41.4 / 41.3 / 43.6
mDeBERTa-v3	57.6 / 55.1 / 52.7	73.9 / 68.1 / 67.7	25.4 / 22.8 / 16.8	22.1 / 21.6 / 12.6	59.7 / 60.1 / 73.3	59.4 / 69.9 / 51.3	59.5 / 61.8 / 60.4	78.8 / 80.7 / 74.5	44.8 / 43.8 / 33.9
XLM-R$_{\text{Base}}$	59.4 / 56.3 / 56.0	70.2 / 65.4 / 62.5	20.0 / 20.6 / 23.5	26.5 / 22.1 / 23.1	54.5 / 64.6 / 92.2	71.5 / 71.9 / 77.1	60.9 / 61.6 / 60.2	79.9 / 82.8 / 76.7	44.5 / 43.3 / 46.2
XLM-R$_{\text{Large}}$	58.4 / 56.8 / 54.1	70.5 / 67.3 / 66.0	22.5 / 19.7 / 36.2	26.7 / 28.2 / 33.0	65.5 / 56.1 / 77.0	73.7 / 78.8 / 84.9	60.8 / 61.6 / 60.1	81.3 / 83.7 / 77.9	46.9 / 45.1 / 54.9
X-MOD$_{\text{Base}}$	59.0 / 56.2 / 54.8	71.1 / 68.7 / 64.1	19.8 / 17.2 / 24.4	23.2 / 24.2 / 16.4	55.7 / 61.1 / 79.3	63.1 / 74.5 / 57.8	60.2 / 61.3 / 60.0	79.4 / 82.4 / 76.0	42.6 / 42.1 / 40.9
SwissBERT$_{\text{(xlm-vocab)}}$	57.6 / 55.9 / 57.3	72.4 / 69.3 / 61.2	23.8 / 20.3 / 39.4	28.5 / 24.0 / 18.7	50.0 / 66.8 / 72.4	71.2 / 72.4 / 76.1	61.1 / 62.2 / 60.9	79.8 / 82.5 / 76.3	46.7 / 44.4 / 47.3
mT5$_{\text{Small}}$	54.8 / 51.7 / 50.3	69.2 / 61.9 / 56.4	14.2 / 16.2 / 10.5	15.9 / 18.1 / 20.2	37.6 / 67.6 / 66.2	51.7 / 86.6 / 54.4	59.8 / 59.5 / 57.5	75.9 / 77.7 / 69.4	33.1 / 38.3 / 32.3
mT5$_{\text{Base}}$	54.1 / 52.1 / 50.3	66.4 / 61.9 / 56.8	10.6 / 16.3 / 16.9	18.7 / 18.7 / 22.1	40.4 / 80.8 / 70.6	47.2 / 87.9 / 62.7	56.2 / 55.0 / 52.6	73.4 / 75.3 / 67.9	31.0 / 39.0 / 38.9
BLOOM-560m	55.1 / 53.2 / 50.9	64.6 / 65.3 / 56.2	12.6 / 16.1 / 7.1	9.5 / 13.6 / 5.1	39.9 / 61.8 / 63.2	42.7 / 61.8 / 59.6	59.8 / 61.5 / 60.3	68.8 / 84.2 / 69.1	26.8 / 34.8 / 18.4
Legal-ch-R$_{\text{Base}}$	59.3 / 58.4 / 55.5	73.8 / 69.4 / 68.6	24.3 / 20.5 / 10.5	26.2 / 25.3 / 14.0	79.8 / 72.1 / 79.6	80.8 / 79.6 / 78.6	62.5 / 65.8 / 63.9	87.6 / 87.9 / 83.7	49.4 / 46.3 / 31.7
Legal-ch-R$_{\text{Large}}$	58.3 / 55.7 / 53.6	71.9 / 68.5 / 66.7	23.0 / 21.3 / 38.7	28.5 / 26.0 / 9.0	74.0 / 77.4 / 79.6	77.6 / 80.5 / 99.5	61.6 / 63.9 / 63.0	88.6 / 88.7 / 84.1	49.0 / 47.0 / 36.2
Legal-ch-LF$_{\text{Base}}$	60.7 / 58.3 / 55.5	74.8 / 70.0 / 67.8	25.3 / 21.5 / 18.4	29.2 / 26.7 / 9.9	75.9 / 70.3 / 99.7	82.2 / 72.6 / 100	63.1 / 67.1 / 66.0	87.5 / 87.9 / 84.0	51.2 / 47.2 / 31.0

Table 17: Configuration aggregate scores on the validation set. The macro-F1 scores are provided. The highest values are in bold. It is important to note that the scores presented here are calculated as the harmonic mean over multiple seeds.

Model	CPB-F	CPB-C	CPC-F	CPC-C	SLAP-F	SLAP-C	JP-F	JP-C	Agg.
MiniLM	59.1	71.0	14.9	36.9	73.8	78.9	60.9	81.4	44.4
DistilmBERT	59.6	70.1	26.3	35.8	74.1	90.3	60.8	78.8	53.1
mDeBERTa-v3	60.1	73.0	30.4	36.0	77.4	82.0	63.3	81.1	55.5
XLM-R$_{\text{Base}}$	60.1	70.5	26.9	38.5	78.7	92.2	62.9	82.5	55.0
XLM-R$_{\text{Large}}$	60.5	71.7	27.2	39.7	74.0	96.2	63.2	83.1	55.5
X-MOD$_{\text{Base}}$	57.1	71.0	27.0	33.4	81.3	94.4	62.5	82.1	53.5
SwissBERT$_{\text{(xlm-vocab)}}$	59.0	72.1	29.4	38.8	85.6	95.5	62.6	82.3	56.8
mT5$_{\text{Small}}$	54.8	66.1	26.3	32.5	84.7	88.0	60.1	77.3	51.6
mT5$_{\text{Base}}$	55.7	64.4	24.3	29.3	83.0	82.6	47.7	66.1	47.3
BLOOM-560m	52.2	64.3	20.1	21.8	78.5	82.4	60.5	76.7	43.3
Legal-ch-R$_{\text{Base}}$	61.2	**73.6**	27.7	41.0	99.0	96.1	65.3	88.8	58.2
Legal-ch-R$_{\text{Large}}$	**61.8**	73.5	29.8	32.0	**99.3**	**98.8**	65.1	**89.7**	56.6
Legal-ch-LF$_{\text{Base}}$	59.4	72.7	**32.2**	**42.5**	99.1	98.1	**67.0**	89.1	**60.8**

I Detailed Data Description

In this section, we provide additional information about the datasets. Table 18 provides additional information about general dataset metadata.

Table 18: Listing of cantons, courts, chambers, law-areas

Metadata	Number	Examples
Cantons	26 (+1)	Aargau (AG), Bern (BE), Basel-Stadt (BS), Solothurn (SO), Ticino (Ti), Vaud (VD),... (+ Federation (CH))
Courts	184	Cantonal Bar Supervisory Authority, Supreme Court, administrative authorities, Tax Appeals Commission, Cantonal Court, Federal Administrative Court, ...
Chambers	456	GR-UPL0-01, AG-VB-002, CH-BGer-011, ZH-OG-001, ZG-VG-004, VS-BZG-009, VD-TC-002, TI-TE-001, ...
Law-Areas	4	Civil, Criminal, Public, Social
Languages	5	German, French, Italian, Romansh, English

I.1 Pre-training

I.1.1 Rulings and Legislation

Figures 6 and 8 provide an overview of the distribution of languages and cantons in the rulings dataset respectively. Figure 4 shows the length distribution of the cases.

Figures 7 and 9 provide an overview of the distribution of languages and cantons in the legislation dataset respectively. Figure 5 shows the length distribution of the legislation texts

I.2 Leading Decisions

Figures 10 and 11 show the length distributions for the facts and considerations of the Leading Decisions dataset.

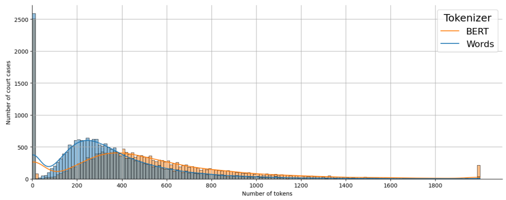


Figure 10: Leading Decisions facts length distribution

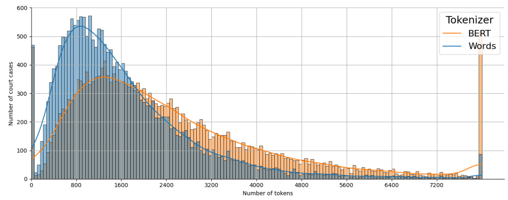


Figure 11: Leading Decisions considerations length distribution

I.3 Law Area Prediction

Figure 12 shows the length distribution for the facts of the LAP dataset.

I.4 Criticality Prediction

Figures 13 and 14 show the length distributions for the facts and the considerations of the acCP dataset respectively.

I.5 Judgment Prediction

Figure 15 shows the length distribution for the facts of the JP dataset.

I.6 Court View Generation

Figures 16 and 17 show the length distributions for the facts and the considerations of the CVG dataset respectively.

I.7 Leading Decision Summarization

Figures 18 and 19 show the length distributions for the input text and the summary of the LDS dataset respectively

I.8 Information Retrieval

Figure 20 shows the length distribution for the facts of the IR dataset. Figure 3 shows the structure of the corpus, queries and qrels for the IR task.

I.9 Citation Extraction

Table 26 shows an illustration of the Citation Extraction (CE) task.

Table 19: Illustration of the pre-training corpora

Motivation: Pre-training Corpora

A large corpus of high quality domain specific text is crucial for training LLMs capable of performing tasks in a given domain. This dataset collects a large part of publicly available legal text relevant for Switzerland.

Legislation text:

Der Grosse Rat des Kantons Aargau, gestützt auf die §§ 72 Abs. 3 und 78 Abs. 1 der Kantonsverfassung, beschliesst:
1. Allgemeine Bestimmungen
§ Gegenstand und Zweck
1 Dieses Gesetz regelt a) die amtliche Information der Öffentlichkeit und den Zugang zu amtlichen Dokumenten [...]
§ 15 Bekanntgabe an Private
Öffentliche Organe geben Privaten Personendaten nur bekannt, wenn
a) sie dazu gesetzlich verpflichtet sind, oder
b) die Bekanntgabe nötig ist, um eine gesetzliche Aufgabe erfüllen zu können [...]

Metadata:

UUID: *58450ad4-108d-4e10-b559-a7efece689d7*
Year: *2015*, Language: *German*, Canton: *AG*
Title: *Gesetz über die Information der Öffentlichkeit, den Datenschutz und das Archivwesen*
Abbreviation: *IDAG*, SR Number: *150.700*

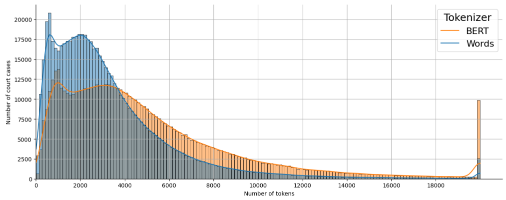


Figure 4: Rulings text length distribution

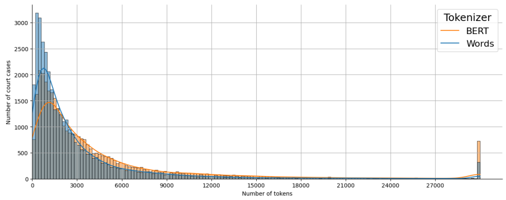


Figure 5: Legislation text length distribution

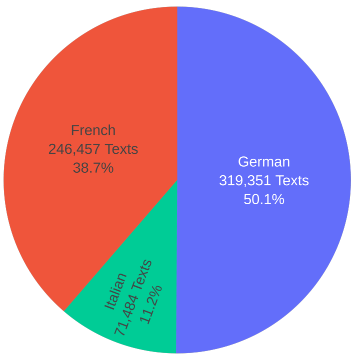


Figure 6: Language distribution of rulings texts

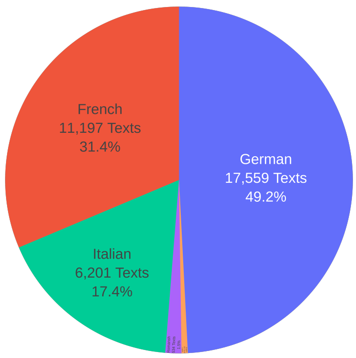


Figure 7: Language distribution of legislation texts

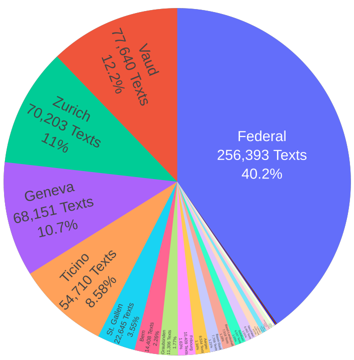


Figure 8: Cantonal distribution of rulings texts

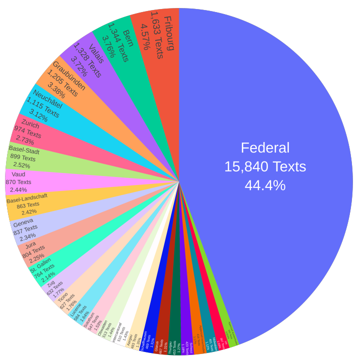


Figure 9: Cantonal distribution of legislation texts

Table 20: Illustration of the Sub Law Area Prediction (SLAP) task

Motivation: Sub Law Area Prediction (SLAP)

Before the judge even sees a complaint, it is first handled by the court's administrative staff, deciding to which chamber (suborganisation inside the court hearing matters in a specific subpart of the law) the complaints should be routed. For this task, models trained on a dataset like ours could assist by providing a suggestion.

Input	**Target**
[Facts]: I. Faits 1. En date du 11 novembre 2013, l'intimé a déposé à la Commune de Corcelles une demande de permis de construire pour la pose d'un revêtement bitumineux sur l'accès à son immeuble, le prolongement d'un chemin existant et l'installation d'une piscine sur les parcelles n° C._ et D._ du registre foncier de la commune de Corcelles. Les parcelles se situent en zone agricole. Le recourant a formé opposition contre ce projet de construction. Dans sa décision globale du 1er décembre 2014, la Préfecture du Jura bernois a accepté la demande d'octroi du permis de construire. 2. Le 31 décembre 2014, le recourant a déposé un recours contre cette décision auprès de la Direction des travaux publics, des transports et de l'énergie du canton de Berne (TTE). Il fait valoir, en substance, que différentes conditions de la décision globale du 1er décembre 2014 n'auraient pas été respectées. Il fait également valoir que l'intimé aurait réalisé sur son immeuble différents travaux sans autorisation. 3. L'Office juridique, qui dirige les procédures de recours pour la TTE1, a requis le dossier préliminaire et dirigé l'échange des mémoires. Les prises de position de l'intimé, de l'instance précédente et de la commune de Corcelles ont été envoyées le 15 janvier 2015, le 4 février 2015 et le 6 février 2015. Le recourant a déposé deux autres prises de position, le 9 janvier 2015 et le 11 mars 2015. Dans la mesure où cela est important pour la décision, il sera fait référence aux mémoires dans les considérants ci-dessous. L'intimé a vendu son immeuble entre-temps. La présente décision est envoyée au nouveau propriétaire pour information.	Urban Planning and Environmental

Possible Labels:

Tax, Urban Planning and Environmental, Expropriation, Public Administration, Other Fiscal, Rental and Lease, Employment Contract, Bankruptcy, Family, Competition and Antitrust, Intellectual Property, Substantive Criminal, Criminal Procedure

Metadata:

Decision ID: *519d0350-6e0e-5551-9bc9-1df033382168*
Year: *2015*, Language: *French*, Law Area: *Public*, Law Sub Area: *Urban Planning and Environmental*
Court: *BE_VB*, Chamber: *BE_VB_001,* Canton: *BE*, Region: *Espace Mittelland*

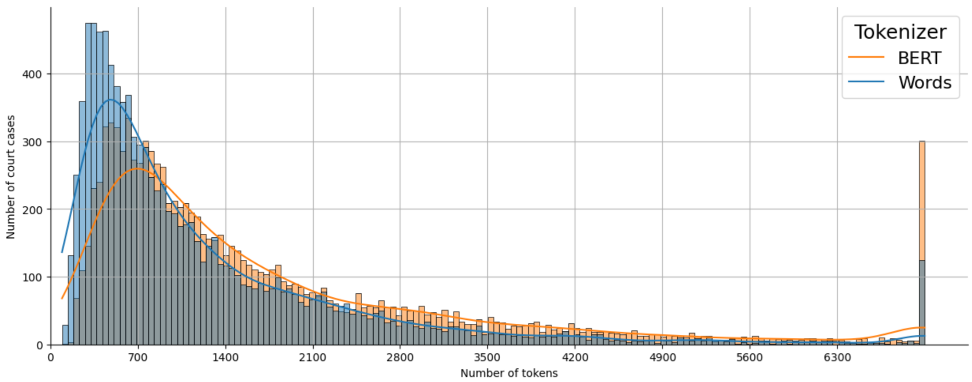


Figure 12: Law Area Prediction facts length distribution

Table 21: Illustration of the Criticality Prediction (CP) task

Motivation: Criticality Prediction (CP)	
We see two potential applications of the Criticality Prediction task: Prioritization and Classification. The prioritization task takes as input the facts as a proxy for a complaint and produces a prioritization score judging how critical/important this case is. This prioritization might help decide which cases should be heard earlier or by more experienced judges. In a futuristic scenario where automatic judgment prediction is accepted, cases with low priority could be sent to automated solutions, while high priority cases would be sent to human judges. The classification task takes as input the considerations or the entire ruling and performes a post-hoc analysis comparing it to prior caselaw and judging its potential impact on future jurisprudence. It can be used to designate certain seminal cases that are likely most influential and for future cases.	
Input	**Target**
[Consideraions]: Erwägungen: 1. Angefochten ist der in einem kantonal letztinstanzlichen Scheidungsurteil festgesetzte nacheheliche Unterhalt in einem Fr. 30'000.– übersteigenden Umfang; auf die Beschwerde ist somit einzutreten (Art. 72 Abs. 1, Art. 74 Abs. 1 lit. b, Art. 75 Abs. 1 und Art. 90 BGG). 2. Die Parteien pflegten eine klassische Rollenteilung, bei der die Ehefrau die Kinder grosszog und sich um den Haushalt kümmerte. Infolge der Trennung nahm sie im November 2005 wieder eine Arbeitstätigkeit auf und erzielt mit einem 80%-Pensum Fr. 2'955.– netto pro Monat. Beide kantonalen Instanzen haben ihr jedoch auf der Basis einer Vollzeitstelle ein hypothetisches Einkommen von Fr. 3'690.– angerechnet. Das Obergericht hat zwar festgehalten, der Ehefrau sei eine Ausdehnung der Arbeitstätigkeit kaum möglich, gleichzeitig aber erwogen, es sei nicht ersichtlich, weshalb sie nicht einer Vollzeitbeschäftigung nachgehen könne. Ungeachtet dieses Widerspruches wird das Einkommen von Fr. 3'690.– von der Ehefrau ausdrücklich anerkannt, weshalb den nachfolgenden rechtlichen Ausführungen dieser Betrag zugrunde zu legen ist. Der Ehemann verdient unbestrittenermassen Fr. 5'334.– netto pro Monat. [...]	BGE label: critical Citation label: critical-1
Possible Labels: BGE label: critical, non-critical Citation label: critical-1, critical-2, critical-3, critical-4	
Metadata: Decision ID: *65aad3f6-33c2-4de2-91c7-436e8143d6ea* Year: *2007*, Language: *German*, Law Area: *Civil*, BGE Label: *Critical*, Citation Label: *Citation-1* Court: *CH_BGer*, Chamber: *CH_BGer_005,* Canton: *CH*, Region: *Federation*	

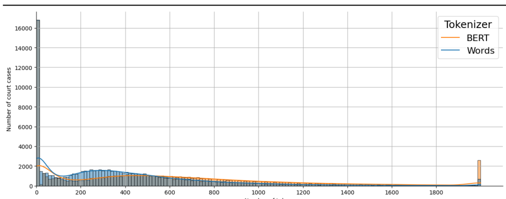


Figure 13: Criticality Prediction facts length distribution

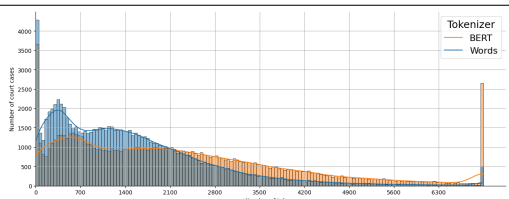


Figure 14: Criticality Prediction considerations length distribution

Table 22: Illustration of the Judgment Prediction (JP) task

Motivation: Judgment Prediction (JP)	
Judgment Prediction might be used in the future in jurisdictions that are experiencing extremely high case loads such as US immigration. The legal maxim "justice delayed is justice denied" may provide motivation for judgment prediction being applied in such highly overloaded jurisdictions by giving affected people the opportunity to have their case heard much earlier (in some jurisdictions wait times are years). For example, consider a scenario where a person is held in pretrial detention awaiting their case to be heard. With Judgment Prediction, it may be possible to identify cases where there is a high likelihood of the individual being not guilty. These cases can then be prioritized for judicial review, potentially reducing the time innocent individuals spend in detention due to system backlogs.	
Input	**Target**
[Facts]: En fait : A. Le 5 février 2015 à 21 h 30, à [...],A.H._ a été appréhendé par la police, qui l'a entendu le lendemain vers 0 h 45 comme prévenu notamment d'infraction à la LStup (Loi fédérale sur les stupéfiants ; RS 812.121). L'intéressé a déclaré qu'alors qu'il se trouvait dans un bar à [...], un homme s'était assis à côté de lui et lui avait demandé de la cocaïne. Le prévenu s'était rendu dans l'appartement occupé notamment par B.H._, un compatriote qui l'hébergeait à l'occasion, pour prendre une boulette de cocaïne, qu'il avait ensuite vendue à l'inconnu pour 100 francs. Les policiers lui avaient finalement indiqué que le client en question était en réalité un agent de police en civil. Sur la base d'indications fournies par A.H._ au client lors de cette transaction, l'appartement occupé par B.H._, rue de [...] à [...], avait fait l'objet, la veille vers 22 h 45, d'une perquisition qui avait amené la découverte de 29.8 g de cocaïne et de plusieurs téléphones portables. Le 6 février 2015 à 13 h 50, le Procureur cantonal Strada a procédé à l'audition d'arrestation de A.H._, lequel a confirmé ses déclarations à la police, en particulier la vente d'une boulette de cocaïne la veille au soir. L'intéressé a également indiqué qu'il avait aidé B.H._ à confectionner des boulettes de cocaïne trois jours plus tôt, qu'il consommait de la cocaïne depuis décembre 2014 et qu'il n'avait pas le droit de demeurer en Suisse, où il était revenu le 23 janvier 2015. B. Par ordonnance du 7 février 2015, le Tribunal des mesures de contrainte, faisant droit à la requête du Ministère public, a ordonné, en raison du risque de fuite et du risque de collusion, la détention provisoire de A.H._ pour une durée maximale de trois mois, soit au plus tard jusqu'au 5 mai 2015. C. Par acte du 17 février 2015, A.H._ a interjeté recours devant la Chambre des recours pénale contre cette ordonnance, en concluant, avec suite de frais et de dépens, à sa réforme principalement en ce sens que la demande de détention provisoire soit refusée et la mise en liberté provisoire ordonnée, et à ce que la procédure dirigée contre lui soit classée. Il n'a pas été ordonné d'échanges d'écritures.	Dismissal
Possible Labels:	
Approval, Dismissal	
Metadata:	
Decision ID: *0dd2f9f7-872e-4200-9f9c-f1c12520c267* Year: *2015*, Language: *French*, Law Area: *Penal*, Judgment: *Dismissal* Court: *VD_TC*, Chamber: *VD_TC_013,* Canton: *VD*, Region: *Région lémanique*	

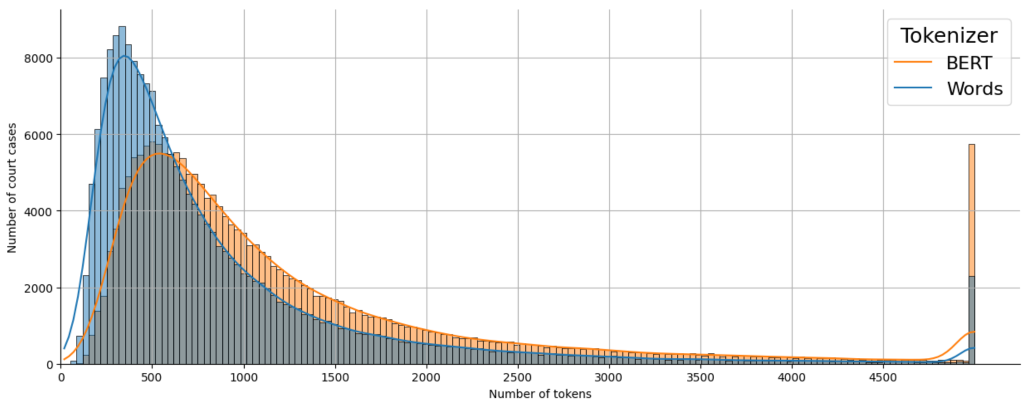


Figure 15: Judgment Prediction facts length distribution

Table 23: Illustration of the Court View Generation (CVG) task

Motivation: Court View Generation (CVG)

Court view generation is arguably one of the most difficult NLP tasks. It requires extensive legal reasoning capabilities, significant experience, and good knowledge of the specific law area a judge is operating in. Machines may assist judges and clerks by suggesting word or sentence continuations while they are typing or even by setting up complete drafts. To solve this task well, the following ingredients are likely necessary: First, a strong retrieval system capable of providing the necessary legal context based on legislation and influential previous rulings. Second, a strong legal reasoning system capable of analyzing the facts, any lower court decisions, and the retrieved documents. This matter is complicated further in our dataset due to lengthy documents in multiple languages.

Input	**Target**
[Facts]: Zum Sachverhalt: 1. Am Dienstag, 23. August 2005, um 12.16 Uhr, lenkte X seinen Personenwagen von Bronschhofen her kommend auf der Hauptstrasse in Richtung Wil. Auf der Höhe Hauptstrasse 64 wurde er von der Kantonspolizei St. Gallen anlässlich einer Geschwindigkeitskontrolle innerorts mit einer Geschwindigkeit von 80 km/h gemessen. Nach Abzug der technisch bedingten Sicherheitsmarge von 5 km/h resultierte eine rechtlich relevante Geschwindigkeit von 75 km/h. 2. Mit Strafbescheid des Untersuchungsamtes Gossau wurde der Angeklagte am 10. Mai 2006 wegen grober Verletzung der Verkehrsregeln zu einer Busse von Fr. 610.00 verurteilt. Dagegen erhob er Einsprache. Der Einzelrichter des Kreisgerichtes Alttoggenburg-Wil verurteilte ihn mit Urteil vom 14. September 2006 wegen grober Verkehrsregelverletzung und fällte eine Busse von Fr. 600.00 aus. Für die Löschung im Strafregister wurde eine Probezeit von zwei Jahren angesetzt, die Kosten des Verfahrens wurden dem Angeklagten auferlegt. 3. Dagegen erklärte der Verteidiger fristgerecht Berufung. Er verlangte einen Freispruch von der groben Verkehrsregelverletzung (Art. 90 Ziff. 2 SVG) [...] Die Staatsanwaltschaft trug auf Abweisung der Berufung an.	[Considerations]: Aus den Erwägungen: 1. Nach Art. 90 Ziff. 2 SVG wird mit Freiheitsstrafe bis zu drei Jahren oder Geldstrafe bestraft, wer [...] Der subjektive Tatbestand der groben Verkehrsregelverletzung ist hier deshalb regelmässig zu bejahen. Eine Ausnahme kommt etwa da in Betracht, wo [...] 2. Der Angeklagte bringt vor, die Vorinstanz habe den Grundsatz in dubio pro reo verletzt, wenn [...] Indem der Angeklagte innerorts mit mindestens 25 km/h zu schnell gefahren ist, hat er den objektiven Tatbestand der groben Verkehrsregelverletzung erfüllt. [. . .] Aus dem gleichen Grund ist auch der Beweisantrag zur Vornahme eines Augenscheins abzuweisen. III. 1. Der Angeklagte hat eine grobe Verkehrsregelverletzung begangen. Sein Verschulden wiegt schon deshalb nicht mehr leicht, weil [. . .], so erscheint eine Geldstrafe von 4 Tagessätzen angemessen (Art. 34 i.V.m. Art. 47 StGB). [. . .] Die Voraussetzungen für den bedingten Strafvollzug sind fraglos erfüllt (Art. 42 StGB). [. . .] 2. Der Vollzug der Geldstrafe wird unter Ansetzung einer Probezeit von zwei Jahren bedingt aufgeschoben. Bewährt sich der Angeklagte während der Probezeit nicht, so muss die Prognose seines künftigen Legalverhaltens neu gestellt werden und der Angeklagte müsste mit dem Vollzug der Geldstrafe rechnen. [. . .]

Possible Labels:

Text

Metadata:

Decision ID: *0f86bb1e-ed24-52a1-bec7-e04451485a7f*
Year: *2007*, Language: *German*, Law Area: *Penal*
Court: *SG_KG*, Chamber: *SG_KG_001,* Canton: *SG*, Region: *Eastern Switzerland*

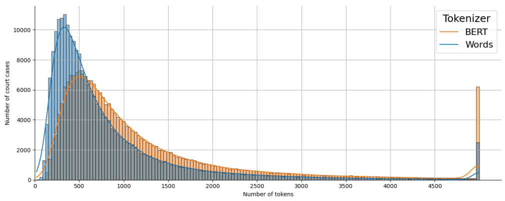


Figure 16: Court View Generation facts length distribution

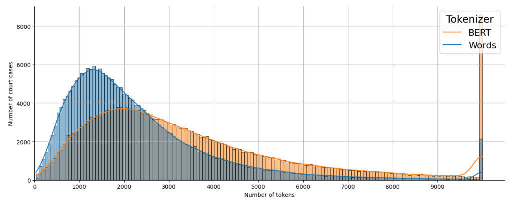


Figure 17: Court View Generation considerations length distribution

Table 24: Illustration of the Leading Decision Summarization (LDS) task

Motivation: Leading Decision Summarization (LDS)

Summarizing cases is very important for lawyers to absorb the most relevant information in less time. Lawyers need to read many cases during their research. Reducing the time needed to comprehend the gist of a case brings direct economic value.

Input	**Target**
[Case Text]: BGE 141 IV 201 S. 201 Dai considerandi: 8. 8.2.1 È stato accertato, senza arbitrio, che la ricorrente ha più volte chiesto a F. di trovare, nel senso di contattare e ingaggiare (avendo precisato che aveva i soldi per pagare), qualcuno che potesse uccidere il marito e che egli rifiutò di fare quello che gli si domandava. 8.2.2 Contrariamente a quanto sostenuto nel gravame, la contestata richiesta risulta tutt'altro che generica: permetteva di ben comprendere sia il genere di infrazione finale prospettata (reato contro la vita) sia la vittima designata sia il comportamento da assumere, ossia reperire e ingaggiare qualcuno allo scopo, atteso che vi era a disposizione denaro. F. non si è risolto a commettere alcunché, motivo per cui si è di fronte solo a un tentativo di istigazione e la questione del nesso causale tra l'atto di persuasione e la decisione dell'istigato di commettere il reato neppure si pone. [...] È piuttosto nell'ambito della commisurazione della pena che occorre considerare la gravità reale del tentativo di istigazione, le conseguenze concrete dell'atto commesso e la prossimità del risultato (v. sentenza 6S.44/2007 del 6 giugno 2007 consid. 4.5.5). Nella fattispecie la Corte cantonale ha effettivamente considerato tali aspetti al momento di commisurare la pena. Sicché su questo punto la condanna della ricorrente non viola l' art. 24 cpv. 2 CP ed è conforme al diritto federale.	[Regeste]: Regeste Art. 24 Abs. 2 StGB; indirekte Anstiftung (Kettenanstiftung), Versuch. Auch die versuchte indirekte Anstiftung (Kettenanstiftung) zu einem Verbrechen ist strafbar (E. 8.2.2).

Possible Labels:

Text

Metadata:

Decision ID: *91ae0d9f-9aec-4b2b-a7ee-042abc42adaa*
Year: *2015*, Language: *Italian*
Court: *CH_BGE*, Chamber: *CH_BGE_006*, Canton: *CH*, Region: *Federation*

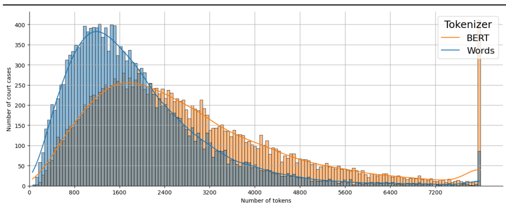


Figure 18: Leading Decision Summarization input length distribution

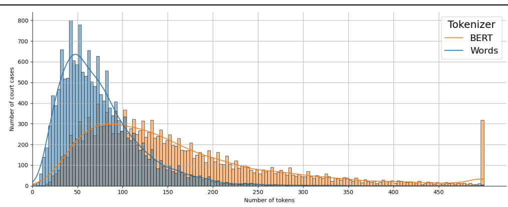


Figure 19: Leading Decision Summarization summary length distribution

Table 25: Illustration of the Multilingual Information Retrieval (MLIR) task

Motivation: Multilingual Information Retrieval (MLIR)

Information retrieval is at the heart of the daily work of lawyers. Much like in scientific writing, lawyers base their arguments on prior caselaw, relevant legislation, and legal analyses. Thus, they spend a large part of their work searching for these documents, motivating the importance of legal IR. Annotating these data at scale is very costly, which is why this dataset is based on the citation graph of Swiss Supreme Court cases. In Switzerland lawyers operate in a trilingual jurisdiction with legislation and caselaw appearing in up to three official languages German, French and Italian. This means that for a complaint written in German, a case written in French might be relevant, leading to Multilingual IR, further complicating the task.

Input	**Target**
[Facts]: Fatti: A. Il 9 novembre 2006 G._, nata P._ (1959), ha contratto matrimonio con F._. Dall'unione non sono nati figli. Per contro il marito ha avuto figli (ormai adulti) dal primo matrimonio i quali non hanno però vissuto in economia domestica con G._. Il 26 settembre 2011 è deceduto F._. Con domanda del 4 ottobre 2011 G._ ha chiesto alla cassa di compensazione Medisuisse l'erogazione di una rendita vedovile. Con decisione del 27 ottobre 2011, sostanzialmente confermata il 22 dicembre successivo in seguito all'opposizione dell'interessata, la cassa di compensazione ha respinto la richiesta di prestazione per il motivo che la richiedente non era stata sposata almeno cinque anni con il defunto marito, come invece prescritto dalla legge, bensì "solo" 4 anni 10 mesi e 18 giorni. B. Osservando che il termine di cinque anni non era adempiuto per soli pochi giorni e invocando di conseguenza una applicazione della legge secondo un giudizio di giustizia ed equità, G._ si è aggravata al Tribunale delle assicurazioni del Cantone Ticino e ha chiesto il riconoscimento della rendita. Statuendo per giudice unico, la Corte cantonale ha respinto il ricorso per pronuncia del 29 febbraio 2012. C. L'interessata ha presentato ricorso al Tribunale federale al quale ribadisce la richiesta di prima sede. Dei motivi si dirà, per quanto occorra, nei considerandi. Non sono state chieste osservazioni al gravame.	Laws: 75488867-c001-4eb9-93b9-04264ea91f55 e10ed709-8b11-47e3-8006-88b26d86e498 e6b06567-1236-4210-adb3-e11c26e497d5 2ef9b20e-bb7c-491f-9391-59ac4f74e3c9 b8d4aeef-a8ef-40d9-92a1-090a37538008 1af9b596-92d7-4f80-a38b-876ed88ccfe5 53be6a03-1fd8-4980-aa5c-bd81e9a54d5e 4b5a2135-fee2-4e3b-811e-15ce1c71bddf 6ab38922-6309-4021-83cc-56d776d7a332 Cited Rulings: 54df6482-97d7-47eb-afb1-1ccb9369cb89 921a799a-9077-4057-8e46-4919fd4f3263

Possible Labels:

10K laws and rulings (see Section 4.5 for more information)

Metadata:

Decision ID: *6856ac58-5d12-48c4-acef-831d50c79886*
Year: *2012*, Language: *Italian* Law Area: *Social*
Court: *BE_VB*, Chamber: *CH_BGer_009,* Canton: *BE*, Region: *Federation*

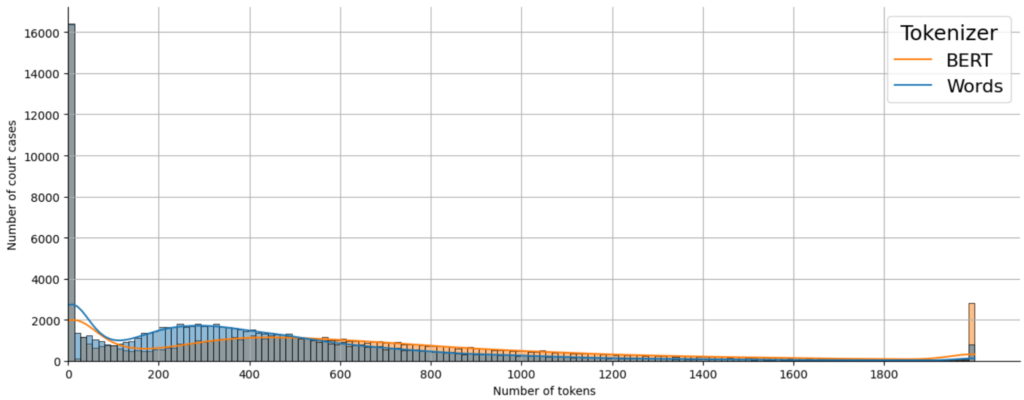


Figure 20: Information Retrieval facts length distribution

Table 26: Illustration of the Citation Extraction (CE) task

Motivation: Citation Extraction (CE)

Citation extraction is an important preprocessing step to collect information from legal documents. It enables easy semantic linking to relevant legislation, caselaw and analyses. Due to extensive rulebooks, simple regexes are often insufficient for accurate extraction of legal citations, motivating the need for more complex approaches.

Input	**Target**
Considerations: ['ergangen', 'ist', 'und', 'sich', 'das', 'Verfahren', 'daher', 'noch', 'nach', 'dem', 'Bundesgesetz', 'über', 'die', 'Organisation', 'der', 'Bundesrechtspflege', '(', 'OG', ')', 'vom', '16', '.', 'Dezember', '1943', 'richtet', '(', 'vgl', '.', 'Art', '.', '132', 'Abs', '.', '1', 'BGG', ';', 'BGE', '132', 'V', '393', 'E', '.', '1', '.', '2', 'S', '.', '395', ')', 'dass', 'das', 'Verfahren', 'nicht', 'die', 'Bewilligung', 'oder', 'Verweigerung', 'von', 'Versicherungsleistungen', 'zum', 'Gegenstand', 'hat', 'und', 'deshalb', 'gemäss', 'Art', '.', '134', 'Satz', '1', 'OG', '[', 'in', 'der', 'von', '1']	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 4, 4, 0, 0, 0, 0, 0, 1, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 4, 4, 0, 0, 0, 0, 0, 0, 0, 0]

Possible Labels:

0: O
1: B-CITATION
2: I-CITATION
3: B-LAW
4: I-LAW

Metadata:

Decision ID: *1572342e-a20d-4137-9593-47fc43b98af3*
Year: *2007*, Language: *German*, Law Area: *Social*
Court: *CH_BGer*, Chamber: *CH_BGer_009,* Canton: *CH*, Region: *Federation*

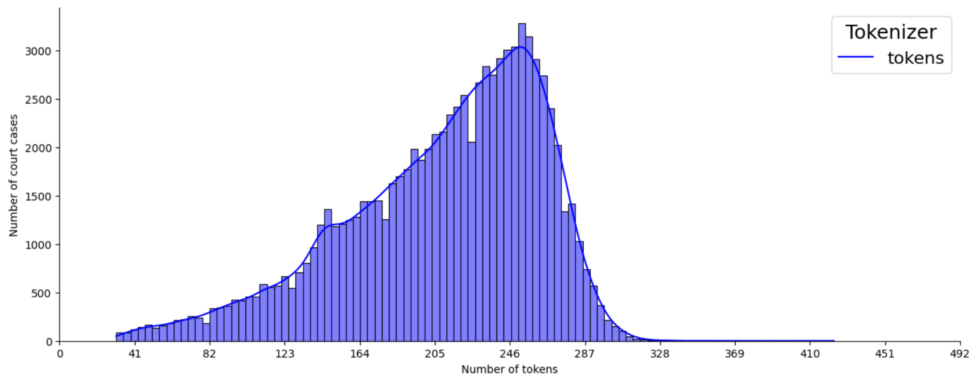


Figure 21: Citation Extraction considerations length distribution

J Error Analysis

J.1 Court View Generation (CVG)

J.1.1 German

Table 27 shows a comparison of generated text in the German language from GPT-3.5-Turbo-16K, GPT-4, Claude Instant and Claude 2. Generally, specific textual constructs, such as those related to asylum applications, appear frequently in a similar manner so fine-tuned models are able to proficiently predict these patterns, whereas zero-shot models face challenges. *LLaMA-2-13B-Chat (0-shot)* switches to English after a few sentences, and the German segments containing linguistic and grammatical errors. A possible reason could be that LLaMA-2 was predominantly trained in English rather than other languages. It also made an unsupported claim regarding the appellant's age, which wasn't mentioned in the input. Zero-shot models tend to center more on the primary content, while fine-tuned models are tailored to predict formalities, mirroring the target. The fine-tuned *LLaMA-2-13B-Chat* references 'BFM' instead of the 'SEM' (State Secretariat for Migration), deviating from both *GPT-3.5*'s outputs and the original input. To note, SEM emerged in 2004 from a merger between the BFM (Federal Office for Refugees) and the IMES (Federal Office for Immigration, Integration and Emigration), nevertheless our sample was from 2016. As outlined in Section 5.3, constraints in model availability and computational resources led to truncation of the target output during fine-tuning and experimentation, thus accentuating formal aspects like court jurisdiction and the legitimacy of appeals. This causes zero-shot iterations to receive lower scores, even when they remain contextually accurate. In terms of content, both zero-shot models were on the right track, tending more towards rejection. Furthermore, zero-shot predictions provide clearer insights by minimizing emphasis on formal nuances.

J.1.2 French

Table 28 shows a comparison of generated text in the French language from GPT-3.5-Turbo-16K, GPT-4, Claude Instant and Claude 2. Remarkably, the LLaMA zero-shot model doesn't revert to English when handling the French text, unlike its behavior in the German scenario.

J.2 Leading Decision Summarization (LDS)

Table 29 shows a comparison of generated text from GPT-3.5-Turbo-16K, GPT-4, Claude Instant and Claude 2.

In the side-by-side analysis, all models demonstrated an authentic stylistic representation. GPT-3.5-Turbo 16K provides a more narrative-style summary, while the other models stick more closely to a classic 'Regeste' format. However, when it came to the factual accuracy, GPT-3.5-Turbo-16K faltered by negating a crucial element (*"wurde entschieden, dass 'kein' hypothetisches Einkommen.."*). Claude-2 cited an outdated legal provision, Article 137 of the ZGB, which has been inactive since 2011, even though the case decision was from 2017 and no such reference was present in the input. Despite this, the article cited remains relevant to the context. On the other hand, GPT-4's reference to BGE 137 III 118 and GPT-3.5's mention of BGE 143 III 233 S. 235 were accurate and existed in the provided input.

Table 27: Comparison of the generated considerations (CVG task) from four LLMs in different settings. The scores of each example is stated below in the following order: BERTScore, BLEU-Score, METEOR, ROUGE-1 / ROUGE-2 / ROUGE-L. The first row contains the facts of the case which were given as input, the second row contains the the considerations, the target of this task.

Generated Considerations - German

Sachverhalt: A. Der Beschwerdeführer ersuchte am 14. August 2015 in der Schweiz um Asyl. Am 27. August 2015 wurde er zur Person befragt (BzP) und am 10. September 2015 fand – im Beisein einer Vertrauensperson – die Anhö- rung zu seinen Asylgründen statt. B. Der aus Albanien stammende Beschwerdeführer gab an, er sei am (...) geboren und somit noch minderjährig. Zur Begründung seines Asylge- suchs führte er aus, er habe eine Beziehung zu einer jungen Frau gehabt, worauf deren Vater, nachdem er von der Beziehung mit seiner Tochter er- fahren gehabt habe, von ihm verlangt habe, diese zu heiraten. Da er aber das Mädchen nicht liebe, habe er weder eine Ehe mit ihr eingehen noch eine Beziehung mit ihr weiterführen wollen. Der Vater der jungen Frau habe sich mit seinen Eltern in Verbindung gesetzt und ihnen erklärt, dass ihr Sohn seine Tochter zu heiraten habe, da er mit ihr eine Beziehung geführt habe. Seine Eltern seien damit einverstanden gewesen. Seither habe ihn sein Vater täglich ermahnt, dass er dieses Mädchen nun heiraten müsse. Weil er sich jeden Tag dasselbe von seinem Vater habe anhören müssen und er dies nicht mehr ertragen habe, habe er sich zum Verlassen seines Heimatlandes entschlossen. Bei einer Rückkehr nach Albanien hätte er mit Schwierigkeiten zu rechnen, da man ihn zur Heirat zwingen würde. Zum Beleg seiner Personalien reichte der Beschwerdeführer eine Identi- tätskarte ein. C. Mit Verfügung vom 5. Oktober 2015 – dem Beschwerdeführer und seiner Vertrauensperson eröffnet am darauffolgenden Tag – stellte die Vorinstanz fest, der Beschwerdeführer erfülle die Flüchtlingseigenschaft nicht, lehnte das Asylgesuch ab und verfügte die Wegweisung aus der Schweiz. Den zuständigen Kanton beauftragte sie mit dem Vollzug der Wegweisung. D. Mit Eingabe vom 13. Oktober 2015 (Poststempel) erhob der Beschwerde- führer gegen diesen Entscheid beim Bundesverwaltungsgericht Be- schwerde und beantragte die Aufhebung der angefochtenen Verfügung, sodann sei ihm Asyl zu gewähren, eventualiter sei die Unzumutbarkeit des Wegweisungsvollzugs festzustellen und ihm die vorläufige Aufnahme zu D-6530/2015 Seite 3 gewähren, subeventualiter sei die Sache an die Vorinstanz zurückzuwei- sen und diese anzuweisen, eine allfällige Wegweisung des Beschwerde- führers pflichtgemäss abzuklären, dies unter Würdigung seiner Minderjäh- rigkeit. In prozessualer Hinsicht beantragte er die unentgeltliche Prozess- führung im Sinne von Art. 65 Abs. 1 VwVG sowie den Verzicht auf die Er- hebung eines Kostenvorschusses.

Das Bundesverwaltungsgericht zieht in Erwägung: 1. 1.1. Gemäss Art. 31 VGG beurteilt das Bundesverwaltungsgericht Be- schwerden gegen Verfügungen nach Art. 5 VwVG. Das SEM gehört zu den Behörden nach Art. 33 VGG und ist daher eine Vorinstanz des Bundesver- waltungsgerichts. Eine das Sachgebiet betreffende Ausnahme im Sinne von Art. 32 VGG liegt nicht vor. Das Bundesverwaltungsgericht ist daher zuständig für die Beurteilung der vorliegenden Beschwerde und entschei- det auf dem Gebiet des Asyls in der Regel – so auch vorliegend – endgültig (Art. 105 AsylG [SR 142.31]; Art. 83 Bst. d Ziff. 1 BGG). 1.2. Das Verfahren richtet sich nach dem VwVG, dem VGG und dem BGG, soweit das AsylG nichts anderes bestimmt (Art. 37 VGG und Art. 6 AsylG). 2. Die Kognition des Bundesverwaltungsgerichts und die zulässigen Rü- gen richten sich im Asylbereich nach Art. 106 Abs. 1 AsylG und im Bereich des Ausländerrechts nach Art. 49 VwVG (vgl. BVGE 2014/26 E. 5), 2.1. Mit Beschwerde kann die Verletzung von Bundesrecht (einschliesslich Missbrauch und Überschreiten des Ermessens) sowie die unrichtige und unvollständige Feststellung des rechtserheblichen Sachverhalts gerügt werden (Art. 106 Abs. 1 AsylG). 3. 3.1. Vorab ist die Frage der Prozessfähigkeit als Sachurteilsvoraussetzung von Amtes wegen zu prüfen (FRITZ GYGI, Bundesverwaltungsrechtspflege, 2. Aufl. 1983, S. 73). Als verfahrensrechtliches Korrelat der Handlungsfä- higkeit ist sie nach den einsch

Decision ID: *26c65bde-9976-569d-920e-a3781b01536f*, Year: *2016*, Language: German, Court: *CH_BVGE*

GPT-3.5-Turbo (Fine-tuned)	**LLaMA-2-13B Chat** (Fine-tuned)
Das Bundesverwaltungsgericht zieht in Erwägung: 1. 1.1 Gemäss Art. 31 VGG beurteilt das Bundesverwaltungsgericht Be- schwerden gegen Verfügungen nach Art. 5 VwVG. Das SEM gehört zu den Behörden nach Art. 33 VGG und ist daher eine Vorinstanz des Bundesver- waltungsgerichts. Eine das Sachgebiet betreffende Ausnahme im Sinne von Art. 32 VGG liegt nicht vor. Das Bundesverwaltungsgericht ist daher zuständig für die Beurteilung der vorliegenden Beschwerde, soweit diese auf dem Gebiet des Asyls (Art. 105 AsylG) zulässig (Art. 83 BGG) ist. 1.2 Beurteilt das Bundesverwaltungsgericht Beschwerden nach Art. 31 VGG, so wendet es – unter Vorbehalt der Anwendung von staatsvertra- glichen Rechten – materielles Bundesrecht an (Art. 37 Abs. 1 VGG). 1.3 Mit Beschwerde kann die Verletzung von Bundesrecht (einschliesslich die unrichtige oder unvollständige Feststellung des rechtserheblichen Sach- verhalts [Art. 97 Abs. 1 BGG] und die Unangemessenheit [Art. 105 Abs. 2 BGG]), die Rechtsverletzung nach Art. 95 BGG und die unrichtige oder un- zweckmässige Sachverhaltsfeststellung (Art. 97 Abs. 1 BGG) gerügt werden. Das Bundesverwaltungsgericht wendet nunmehr auch das Beschwerde- recht nach Art. 105 AsylG sowie die Staatsvertragsvorbehalte an, denen Rechnung zu tragen ist (Art. 37 Abs. 3 VGG). D-6530/2015 Seite 4 2. 2.1 Zur Beschwerde gegen Entscheide über Asylgesuche ist berechtigt, wer durch den angefochtenen Entscheid besonders berührt ist und ein rechtlich geschütztes Inter	[Erwägungen]: 1. 1.1 Gemäss Art. 31 VGG beurteilt das Bundesverwaltungsgericht Be- schwerden gegen Verfügungen nach Art. 5 VwVG. Das BFM gehört zu den Behörden nach Art. 33 VGG und ist daher eine Vorinstanz des Bundesver- wal- tungsgerichts. Eine das Sachgebiet betreffende Ausnahme im Sinne von Art. 32 VGG liegt nicht vor. Das Bundesverwaltungsgericht ist daher zuständig für die Beurteilung der vorliegenden Beschwerde und entschei- det auf dem Gebiet des Asyls endgültig, ausser bei Vorliegen eines Auslie- ferungsersuchens des Staates, vor welchem die beschwerdeführende Per- son Schutz sucht (Art. 105 AsylG [SR 142.31]; Art. 83 Bst. d Ziff. 1 BGG). Eine solche Ausnahme im Sinne von Art. 83 Bst. d Ziff. 1 BGG liegt nicht vor, weshalb das Bundesverwaltungsgericht endgültig entscheidet. 1.2 Das Verfahren richtet sich nach dem VwVG, dem VGG und dem BGG, soweit das AsylG nichts anderes bestimmt (Art. 37 VGG und Art. 6 AsylG). 1.3 Die Beschwerde ist frist- und formgerecht eingereicht. Der Beschwerdeführer hat am Verfahren vor der Vorinstanz teilgenommen, ist durch die angefochtene Verfügung besonders berührt und hat ein schutzwürdiges Interesse an deren Aufhe- bung beziehungsweise Ände- rung. Er ist daher zur Einreichung der Beschwerde legitimiert (Art. 105 und Art. 108 Abs. 1 AsylG; Art. 48 Abs. 1 sowie Art. 52 Abs. 1 VwVG). Auf die Beschwerde ist einzutreten. 2. Mit Beschwerde kann die Verletzung von Bundesrecht
85.15, 80.08, 50.36, 57.45/41.46/57.45	85.07, 78.41, 47.00, 60.56/51.46/60.56
Claude 2 (0-shot)	**LLaMA-2-13B Chat** (0-shot)
Hier meine Erwägungen zum vorliegenden Sachverhalt: 1. Prozessvoraussetzungen Die sachliche Zuständigkeit des Bundesverwaltungsgerichts ergibt sich aus Art. 31 VGG. Gemäss Art. 31 Abs. 1 VGG beurteilt das Bundesverwaltungsgericht Beschw- erden gegen Verfügungen nach Art. 5 AsylG. Die Vorinstanz hat am 5. Oktober 2015 eine Verfügung nach Art. 5 AsylG erlassen. Gegen diese Verfügung wurde frist- und formgerecht Beschwerde erhoben. Die Prozessvoraussetzungen sind damit erfüllt. 2. Materielle Beurteilung 2.1 Flüchtlingseigenschaft Gemäss Art. 3 AsylG gilt als Flüchtling eine Person, die in ihrem Heimatstaat wegen ihrer Rasse, Religion, Nationalität, Zugehörigkeit zu einer bestimmten sozialen Gruppe oder wegen ihrer politischen Anschauungen ernsthaften Nachteilen ausgesetzt ist oder begründete Furcht hat, solchen Nachteilen ausgesetzt zu werden. Der Beschwerdeführer macht geltend, er werde in seinem Heimatland Albanien zur Ehe mit einer Frau gezwungen, die er nicht heiraten möchte. Eine Zwangsheirat kann eine Verletzung der Men- schenrechte darstellen. Allerdings ist im vorliegenden Fall nicht ersichtlich, dass der Beschwerdeführer aufgrund eines in Art. 3 AsylG genannten Merkmals von Zwangsheirat bedroht wäre. Vielmehr scheint es sich um einen familiären Konflikt zwischen den Eltern des Beschwerdeführers und der Familie der jungen Frau zu handeln. Aus den Akten ergeben sich keine Hinweise, dass der Beschwerdeführer bei einer Rückkehr nach Albanien staatlichen Verfolgungshandlungen ausgesetzt wäre. Auch liegen keine Beweise dafür vor, dass er durch nichtstaat	Erwägungen Das Bundesverwaltungsgericht hat den Beschwerdeführer auf seine Er- wägungen hin gehört und bewertet die behaupteten Asylgründe. 1. Minderjährigkeit Der Beschwerdeführer gibt an, geboren zu sein und minderjährig zu sein. Die Vorin- stanz hat jedoch festgestellt, dass er nicht minderjährig ist. Es ist jedoch unumstritten, dass der Beschwerdeführer eine Beziehung zu einer jungen Frau hatte, die von ihrem Vater verlangt wurde, diese zu heiraten. Der Beschwerdeführer hat jedoch keine Ehe mit ihr eingegangen und will nicht weitere Beziehung mit ihr haben. 2. Asylgründe Der Beschwerdeführer behauptet, dass er aus Albanien geflohen ist, weil er sich dort Schwierigkeiten because of his refusal to marry the young woman would have faced. However, the evidence presented does not support this claim. There is no indication that the young woman's father would have subjected the applicant to persecution or violence if he had not married his daughter. 3. Flüchtlingseigenschaft The applicant does not meet the criteria for refugee status under the Geneva Convention. He has not demonstrated that he is fleeing persecution or that his life or freedom would be threatened if he were to return to Albania. 4. Personalien The applicant has provided an identity card as proof of his personal details. However, there is no independent evidence to support his claims. The Vorinstanz has found that the applicant does not meet the criteria for refugee status, and therefore, his asylum application should be rejected. 5. Unzumutbarkeit des Wegweisungsvollzugs The applicant has requested that the court consider the unreasonableness of the deportation order. However, the court has found that the order is reasonable and in accordance with the law. 6. Allfäl- lige Wegweisung The court has considered the possibility of a conditional release of the applicant. However, given the lack of evidence supporting the applicant's claims and the fact that he does not meet the criteria for refugee status, the court has decided not to grant conditional release. 7
66.96, 60.71, 26.59, 27.80/6.31/27.12	62.23, 38.56, 11.97, 14.33/0.42/13.71

Table 28: Comparison of the generated considerations (CVG task) from four LLMs in different settings. The scores of each example is stated below in the following order: BERTScore, BLEU-Score, METEOR, ROUGE-1 / ROUGE-2 / ROUGE-L. The first row contains the facts of the case which were given as input, the second row contains the the considerations, the target of this task.

Generated Considerations - French

considérant en fait A. A._ et l'entreprise individuelle C._ ont conclu un contrat de travail d'une durée indéterminée, ayant pour objet l'engagement de A._ en qualité de collaboratrice administrative, à un taux d'activité de 40%, dès le 10 février 2014. Financée principalement par l'assurance chômage, respectivement par le SECO, C._ a pour but d'apporter son soutien aux jeunes en difficulté d'insertion professionnelle au moyen de cours, d'ateliers et de coaching et a notamment été mandatée par l'association D._ pour apporter son soutien aux adolescents de langue allemande. Dès le 1er mars 2014, A._ était occupée à 40% en tant que collaboratrice administrative et à 10% en tant qu'enseignante. Le contrat de travail, signé le 10 juin 2014, prévoyait un salaire mensuel brut de CHF 2'733.80 pour 16 heures d'activité administrative hebdomadaire ainsi qu'un salaire horaire brut de CHF 50.20 pour 4 heures d'enseignement par semaine. Dès le 1er septembre 2014, A._ a augmenté son taux d'enseignement à 40%, à l'essai. En août 2014, C._ a été transformée en la société B._ GmbH, dont E._ est l'associée gérante avec signature individuelle. B._ GmbH a repris tous les contrats de travail et a été inscrite au registre du commerce le 7 août 2014. B. Le 13 octobre 2014 a eu lieu une séance entre E._, A._, l'enseignante F._ et l'apprenti G._. Au cours de celle-ci, E._ a signalé que A._ était manifestement débordée par l'activité supplémentaire d'enseignement, que l'essai n'avait ainsi pas été probant, qu'il fallait donc l'arrêter et qu'il fallait engager quelqu'un pour assumer ce volet de travail. À partir du 15 octobre 2014, A._ a été mise en arrêt de travail à 100 % jusqu'au 31 octobre 2014 pour des raisons de santé. [...] En date du 30 octobre 2014, A._ a informé E._ qu'elle reprendrait son activité le 3 novembre 2014. Néanmoins, le 2 novembre 2014, A._ a été hospitalisée pour des douleurs dorsales. Elle a alors été mise en arrêt de travail à 100% du 2 au 4 novembre 2014. Le 4 novembre 2014, A._ a indiqué à son employeur qu'elle se présenterait à son travail le lendemain, mais qu'un rendez-vous chez son médecin traitant était prévu le même jour. De plus, elle a demandé à E._ de lui consacrer un peu de temps pour un entretien. Tribunal cantonal TC Page 3 de 9 E._ et A._ se sont entretenues le 5 novembre 2014. Au cours de cet entretien, A._ a déclaré son souhait de poursuivre son activité au sein de l'entreprise. [...]

en droit 1. a) La décision attaquée constitue une décision finale de première instance au sens des art. 308 al. 1 et 236 CPC. La voie de droit ouverte contre une telle décision est l'appel (art. 308 al. 1 let. a CPC), sauf si la valeur litigieuse est inférieure à CHF 10'000.- (art. 308 al. 2 CPC), auquel cas la décision ne peut faire l'objet que d'un recours (art. 319 let. a CPC). Dans le cas d'espèce, la valeur litigieuse au dernier état des conclusions est supérieure à CHF 10'000.- (cf. demande, ch. 1 et 2; réponse, ad. 1), de sorte que la voie de l'appel est ouverte. La valeur litigieuse devant la Cour est néanmoins inférieure à CHF 15'000.- (art. 74 al. 1 let. a et art. 51 al. 1 let. a LTF). Par conséquent, cet arrêt ne peut faire l'objet que d'un recours constitutionnel subsidiaire (art. 113 ss LTF). b) L'appel du 21 octobre 2016 a été déposé en temps utile, soit dans le délai légal de 30 jours à compter de la notification de la décision attaquée survenue le 21 septembre 2016 (art. 311 al. 1 CPC). Il en va de même de l'appel joint déposé le 13 décembre 2016, soit dans le délai légal de 30 jours à compter de la notification de l'appel survenue le 14 novembre 2016 (art. 312 al. 2 CPC). c) Dûment motivés et dotés de conclusions, l'appel et l'appel joint sont pour le surplus recevables en la forme. d) La cognition de la Cour est pleine et entière en fait comme en droit (art. 310 CPC). 2. Dans son appel, l'appelante reproche aux premiers juges de ne pas avoir reconnu le caractère abusif de la résiliation intervenue le 15 décembre 201

Decision ID: *8bb15591-be0b-5bc3-933a-e6f856a46fec*, Year: *2017*, Language: French, Court: *FR_TC*

GPT-3.5-Turbo (Fine-tuned)

en droit 1. En vertu de l'art. 57C al. 3 let. b de la loi du 11 avril 1991 sur l'organisation judiciaire (LOJ ; RS 173.01), le Tribunal cantonal connaît seul des causes portant sur la légalité de certaines décisions des tribunaux cantonaux. Dans le cas particulier, interjeté contre une décision finale prise par la juridiction inférieure sur la base de l'art. 64A al. 1 let. b de la loi sur la procédure civile du 19 novembre 2010 (CPC ; RS 270.11), le recours, déposé dans le délai (art. 48A al. 3 CPC) et la forme prévus par la loi (art. 79 al. 1 LOJ et 321-327 CPC), n'est en principe pas recevable, en vertu de l'art. 57C al. 2 LOJ, sauf s'il doit être traité comme un recours de droit public (cf. art. 57D LOJ). 2. Déposé dans le délai et les formes prescrits par la loi à l'encontre d'une décision finale prise par un tribunal cantonal, le recours doit être traité comme un recours de droit public, étant dès lors admissible sur le principe (art. 57C al. 4 LOJ). 3. En droit public fribourgeois, le TRPC apprécie librement la recevabilité et la wellfondation des recours de droit public cantonal, interprète et applique le droit cantonal de manière indépendante. En cas de violation du droit fédéral, le recours est, quant à lui, soumis au Tribunal fédéral. Celui-ci contrôle la violation du droit fédéral (selon les principes applicables au recours en matière de droit public, cf. à ce propos les art. 82 ss LTF), le droit constitutionnel et, le cas échéant, l'usage du pouvoir d'appréciation cantonal selon la maxime de l'arbitraire (Art. 49 al. 1 Cst. et 106 al. 2 LTF

73.33, 64.17, 28.71, 24.58/6.34/24.58

LLaMA-2-13B Chat (Fine-tuned)

[Considérations]: En droit, les premiers juges ont considéré qu'au vu des faits retenus, il fallait admettre le licenciement abusif. En effet, le lien de confiance qui unissait les parties avait été irrémédiablement rompu dès le 5 novembre 2014. A._ avait refusé de signer le avertissement et la lettre de résiliation de son contrat de travail, ce qui avait pour conséquence qu'elle ne pouvait plus être tenue de respecter les conditions de la résiliation. Par ailleurs, le fait que la demanderesse ait été mise en arrêt de travail à 100 % du 2 au 11 novembre 2014, puis à 100 % dès le 12 novembre 2014 jusqu'au 30 novembre 2014, puis à 70 % du 1er au 14 décembre 2014, puis à 50 % du 15 décembre 2014 au 15 janvier 2015 et enfin à 25 % du 16 au 29 janvier 2015 ne pouvait être interprété comme une résiliation. En effet, la demanderesse ne pouvait pas, en tant que collaboratrice à 40%, assumer son travail à 100% dès le 2 novembre 2014, ce qu'elle aurait dû faire en tant qu'enseignante si l'on tient compte de la situation médicale de cette dernière. Par ailleurs, le fait que l'employeur ait licencié la demanderesse le 7 novembre 2014, soit durant une période de protection, ne pouvait pas être considéré comme un licenciement abusif. En effet, l'employeur n'avait pas connaissance du fait que la demanderesse était en arrêt de travail à 100 % du 2 au 11 novembre 2014. Enfin, l'employeur avait mis un terme au contrat de travail par courrier du 10 décembre 2014, soit à la fin de la période de protection, ce qui était régulier.

63.70, 57.75, 22.96, 23.57/2.20/22.14

Claude 2 (0-shot)

I. Conditions de procédure
1. Selon l'art. 343 al. 1 ch. 1 CPC, le tribunal établit sa compétence d'office et examine s'il a la compétence matérielle, fonctionnelle et locale pour connaître de la demande.
En l'espèce, le litige porte sur une demande en paiement pour licenciement abusif fondée sur un contrat de travail. Il s'agit donc d'un litige relevant du droit du travail, pour lequel le Tribunal des prud'hommes de l'arrondissement de la Sarine est compétent ratione materiae (art. 8 CPC). De plus, la compétence territoriale est établie à l'art. 34 CPC. La compétence fonctionnelle n'est pas contestée.
Les conditions de procédure sont ainsi remplies.
II. En droit 2. La demanderesse a été engagée par l'entreprise individuelle C._ à un taux d'activité de 40% dès le 10 février 2014. Suite à la transformation de C._ en la société B._ GmbH en août 2014, cette dernière a repris le contrat de travail. La demanderesse a augmenté son taux d'activité à 50% dès le 1er septembre 2014, à l'essai.
3. Le 15 octobre 2014, la demanderesse a été mise en arrêt maladie jusqu'au 31 octobre 2014. Le même jour, elle a adressé un courrier au Service public de l'emploi dans lequel elle mettait en cause le professionnalisme de l'intimée.
4. Le 30 octobre 2014, la demanderesse a informé l'intimée de sa reprise de travail le 3 novembre 2014. Toutefois, elle a été hospitalisée du 2 au 4 novembre 2014 et mise à nouveau en arrêt maladie. Le 5 novembre 2014, l'intimée lui a signifié un avertissement ainsi qu'une résiliation ordinaire de son

70.04, 61.20, 23.83, 24.75/2.92/24.75

LLaMA-2-13B Chat (0-shot)

Considérant les faits allégués et les demandes des parties, le tribunal est appelé à évaluer les dispositions juridiques pertinentes et à apprécier les preuves pour déterminer si le licenciement d'A._ a été abusif.
Tout d'abord, il est important de noter que le contrat de travail entre A._ et B._ GmbH a été résilié pendant une période de protection, ce qui signifie que la résiliation était nulle et non avenue. A._ a également signalé des manquements de l'entreprise et a exprimé des critiques quant à la gestion d'E._, ce qui peut être considéré comme une raison valable pour mettre fin au contrat. Cependant, il est important de noter que ces manquements et ces critiques n'ont pas été formellement étayés par des preuves solides.
Ensuite, il est important de considérer le lien de confiance qui existe entre A._ et E._. Bien que E._ ait signalé que le lien de confiance avait été rompu en raison des critiques formulées par A._, il est possible de considérer que le lien de confiance n'a pas été rompu de manière formelle et que les rapports de travail pouvaient continuer.
En outre, il est important de prendre en compte la situation de santé d'A._, qui a été hospitalisée pour des douleurs dorsales et a été mise en arrêt de travail à plusieurs reprises. Bien que les préjudices pour la santé d'A._ soient acknowledgeables, il est important de noter que ces préjudices n'ont pas été directement causés par le licenciement d'A._.
Enfin, il est important de considérer les dispositions juridiques pertinentes. Le Code des obligations suisse (CO) prévoit que le licenciement d'un employé doit être justifié et que le employeur doit fournir des preuves solides pour étayer la résiliation. Le CO également prévoit que le licenciement abusif est une ground for dismissal.
En résumé, le tribunal doit évaluer les dispositions jur

63.92, 49.30, 13.62, 14.63/1.26/13.94

Table 29: Comparison of the generated regeste (LDS task) from four LLMs in different settings. The scores of each example is stated below in the following order: BERTScore, BLEU-Score, METEOR, ROUGE-1 / ROUGE-2 / ROUGE-L. Due to the size of the input, only the target is displayed.

Generated Regeste

Regeste Abänderung des Unterhaltsbeitrages für die Dauer des Scheidungsverfahrens infolge veränderter Verhältnisse; Rechtsmissbrauch (Art. 179 Abs. 1, Art. 2 Abs. 2 ZGB).
Vermindert der Unterhaltspflichtige sein Einkommen in Schädigungsabsicht, ist eine Abänderung der Unterhaltsleistung selbst dann ausgeschlossen, wenn die Einkommensreduktion nicht mehr rückgängig gemacht werden kann (Änderung der Rechtsprechung; E. 3).

Decision ID: *3e70603c-bbcd-47de-a066-23124945fcc1*, Year: *2017*, Language: German, Court: *CH_BGE*

GPT-4 (1-shot)	**Claude 2** (1-shot)
Regeste Abänderung von Unterhaltsbeiträgen bei Reduktion des Einkommens (BGE 137 III 118). Bei der Bemessung des Unterhaltsbeitrages ist grundsätzlich vom tatsächlich erzielten Einkommen des Unterhaltspflichtigen auszugehen. Eine Abänderung der Unterhaltsleistung ist jedoch auszuschliessen, wenn der Unterhaltspflichtige sein Einkommen in Schädigungsabsicht vermindert hat, selbst dann wenn die Einkommensverminderung nicht mehr rückgängig gemacht werden kann (E. 3.4).	Regeste Art. 137 ZGB (Unterhaltsabänderung). Abänderung des Ehegattenunterhalts bei veränderter Einkommenslage. Bei freiwilliger Schädigung ist eine Abänderung selbst dann ausgeschlossen, wenn die Einkommensverminderung nicht mehr rückgängig gemacht werden kann (E. 3).
79.82, 68.93, 55.47, 45.16/28.04/43.01	83.88, 47.53, 42.44, 53.33/29.27/53.33

GPT-3.5-Turbo-16K (0-shot)	**Claude Instant** (0-shot)
BGE 143 III 233 S. 235 - Ehegattenunterhalt: Abänderung aufgrund veränderter Einkommenslage des Unterhaltspflichtigen; Anrechnung eines hypothetischen Einkommens; Rechtsmissbräuchliches Verhalten des Unterhaltspflichtigen. Der unterhaltspflichtige Ehemann beantragte eine Abänderung des Ehegattenunterhalts aufgrund seiner veränderten Einkommenslage. Das Gericht stellte fest, dass eine Abänderung ausgeschlossen ist, wenn die veränderte Einkommenslage durch eigenmächtiges, widerrechtliches oder rechtsmissbräuchliches Verhalten des Unterhaltspflichtigen herbeigeführt wurde. Es wurde entschieden, dass in diesem Fall kein hypothetisches Einkommen angerechnet werden kann, selbst wenn die Einkommensverminderung nicht mehr rückgängig gemacht werden kann.	Regeste: a) Art. 107ff ZGB, 128 III 4 E. 4 b) Ehegattenunterhalt, Abänderungsverfahren, hypothetisches Einkommen c) Das Bundesgericht hielt fest, dass dem Unterhaltspflichtigen, der sein Einkommen in Schädigungsabsicht vermindert hat, selbst dann ein hypothetisches Einkommen anzurechnen ist, wenn die Einkommensverminderung nicht mehr rückgängig gemacht werden kann. Somit ist eine Abänderung der Unterhaltsleistung in einem solchen Fall ausgeschlossen.
72.21, 43.71, 44.39, 27.52/12.60/23.85	76.91, 67.20, 49.20, 41.67/24.30/41.67

K Prompts

In the following, we provide the prompts that we used for experiments with LLMs over 2B parameters.

K.1 Text Classification

We created our prompts in a style similar to that used in Chalkidis [2023]. First, we introduce the document, e.g., the court decision. Then, we provide an explanation of the labels and the task in general. Finally, we present the labels from which the model can choose. To limit the model from producing excessive output, we conclude the prompt with a cloze-style sentence, such as 'The relevant label in this case is:', which the model simply needs to complete.

K.1.1 Criticality Prediction

Criticality Prediction (CP) BGE Facts
Given the facts from the following Swiss Federal Supreme Court Decision:
{INPUT FROM THE VALIDATION SET}
Federal Supreme Court Decisions in Switzerland that are published additionally get the label critical, those Federal Supreme Court Decisions that are not published additionally, get the label non-critical.
Therefore, there are two labels to choose from:
- critical
- non-critical

The relevant label in this case is:

Criticality Prediction (CP) BGE Considerations
Given the considerations from the following Swiss Federal Supreme Court Decision:
{INPUT FROM THE VALIDATION SET}
Federal Supreme Court Decisions in Switzerland that are published additionally get the label critical, those Federal Supreme Court Decisions that are not published additionally, get the label non-critical.
Therefore, there are two labels to choose from:
- critical
- non-critical

The relevant label in this case is:

Criticality Prediction (CP) Citation Facts
Given the facts from the following Swiss Federal Supreme Court Decision:
{INPUT FROM THE VALIDATION SET}
How likely is it that this Swiss Federal Supreme Court Decision gets cited. Choose between one of the following labels (a bigger number in the label means that the court decision is more likely to be cited):
- critical-1
- critical-2
- critical-3
- critical-4

The relevant label in this case is:

Criticality Prediction (CP) Citation Considerations
Given the considerations from the following Swiss Federal Supreme Court Decision:
{INPUT FROM THE VALIDATION SET}
How likely is it that this Swiss Federal Supreme Court Decision gets cited. Choose between one of the following labels (a bigger number in the label means that the court decision is more likely to be cited): - critical-1
- critical-2
- critical-3
- critical-4

The relevant label in this case is:

K.1.2 Judgment Prediction

Judgment Prediction (JP) Facts
Given the facts from the following court decision:

{INPUT FROM THE VALIDATION SET}
Will this court decision get approved or dismissed? There are two labels to choose from:
- dismissal
- approval
The relevant label in this case is:

Judgment Prediction (JP) Considerations
Given the considerations from the following court decision:
{INPUT FROM THE VALIDATION SET}
Will this court decision get approved or dismissed? There are two labels to choose from:
- dismissal
- approval
The relevant label in this case is:

K.1.3 (Sub) Law Area Prediction

Law Area Prediction (LAP) Facts
Given the facts from the following court decision:
{INPUT FROM THE VALIDATION SET}
Which topic/law area is relevant out of the following options:
- Civil
- Public
- Criminal
- Social
The relevant option is:

Law Area Prediction (LAP) Considerations
Given the considerations from the following court decision:
{INPUT FROM THE VALIDATION SET}
Which topic/law area is relevant out of the following options:
- Civil
- Public
- Criminal
- Social
The relevant option is:

Sub Law Area Prediction (SLAP) Facts
Given the facts from the following court decision:
{INPUT FROM THE VALIDATION SET}
Which topic/law area is relevant out of the following options:
- Rental and Lease
- Employment Contract
- Bankruptcy
- Family
- Competition and Antitrust
- Intellectual Property
- Substantive Criminal
- Criminal Procedure
- Tax
- Urban Planning and Environmental
- Expropriation
- Public Administration
- Other Fiscal
The relevant option is:

Sub Law Area Prediction (SLAP) Considerations
Given the considerations from the following court decision:
{INPUT FROM THE VALIDATION SET}
Which topic/law area is relevant out of the following options:
- Rental and Lease
- Employment Contract

- Bankruptcy
- Family
- Competition and Antitrust
- Intellectual Property
- Substantive Criminal
- Criminal Procedure
- Tax
- Urban Planning and Environmental
- Expropriation
- Public Administration
- Other Fiscal

The relevant option is:

K.2 Text Generation

In the subsequent sections, we detail the prompts employed for the 0-shot setup. For the 1-shot experiments, we consistently appended the same example in the respective language to the 0-shot instruction.

K.2.1 Court View Generation (CVG)

Court View Generation (CVG) in German
'Ziel: Generiere Erwägungen basierend auf dem gegebenen Sachverhalt eines Schweizer Gerichtsurteils.
Hintergrund: Ein Schweizer Gerichtsurteil besteht aus Rubrum, Sachverhalt, Erwägungen, Dispositiv (Urteilsformel) und Unterschrift. Die Erwägungen sind die rechtliche Würdigung des Geschehens durch das Gericht.
Anweisung:
-Sachverhalt Verstehen: Der gegebene Sachverhalt enthält bestrittene und unbestrittene Fakten, die Begehren der Parteien, das Beweisverfahren und die Prozessgeschichte.
-Beginne mit Prozessvoraussetzungen: Prüfe zunächst, ob die Prozessvoraussetzungen (z.B. Zuständigkeit des Gerichts) erfüllt sind. Wenn nicht strittig, reicht es aus zu bestätigen, dass die Voraussetzungen erfüllt sind.
-Rechtliche Würdigung: Eruiere relevante Rechtssätze basierend auf den behaupteten und rechtlich relevanten Tatsachen.
-Setze dich mit den rechtlichen Standpunkten der Parteien auseinander.
-Beachte die Beweislastverteilung und würdige die Beweise frei, aber berücksichtige relevante gesetzliche Beweisregeln.
-Iura novit curia: Deine rechtliche Würdigung muss nicht zwangsläufig dem rechtlichen Vorbringen der Parteien entsprechen. Berücksichtige andere mögliche Argumentationslinien.
-Zusammenfassung: Fasse am Ende deine Erwägungen, das Ergebnis Ihrer rechtlichen Würdigung, zusammen.
-Output: Der generierte Text sollte strukturiert, klar und in der Form von typischen Erwägungen eines Schweizer Gerichtsurteils sein.

{Sachverhalt des Schweizer Gerichtsurteils}:
{INPUT FROM THE VALIDATION SET}

{Erwägungen}:

Court View Generation (CVG) in French
But: Génère des considérations basées sur les faits donnés d'un jugement suisse.
Contexte: Un jugement suisse est composé du rubrum, des faits, des considérations, du dispositif (formule du jugement) et de la signature. Les considérations sont l'appréciation juridique des événements par le tribunal.
Instructions:
- Comprends les faits: Les faits donnés contiennent des faits contestés et non contestés, les demandes des parties, la procédure de preuve et l'historique du procès.
- Commence par les conditions de procédure: Vérifie d'abord si les conditions de procédure (par

exemple, la compétence du tribunal) sont remplies. Si cela n'est pas contesté, il suffit de confirmer que les conditions sont remplies.
- Appréciation juridique: Évalue les dispositions juridiques pertinentes basées sur les faits allégués et juridiquement pertinents.
- Confronte-toi aux points de vue juridiques des parties.
- Tiens compte de la répartition de la charge de la preuve et évalue les preuves librement, mais tiens compte des règles légales de preuve pertinentes.
- Iura novit curia: Ton appréciation juridique ne doit pas nécessairement correspondre aux arguments juridiques présentés par les parties. Considère d'autres lignes d'argumentation possibles.
- Résumé: Résume à la fin de tes considérations le résultat de ton appréciation juridique.
- Résultat: Le texte généré devrait être structuré, clair et sous la forme de considérations typiques d'un jugement suisse.

{Faits du jugement suisse}:
{INPUT FROM THE VALIDATION SET}

{Considérations}:

Court View Generation (CVG) in Italian
Obiettivo: Genera considerazioni basate sui fatti presentati in una sentenza svizzera.
Contesto: Una sentenza svizzera si compone di rubrum, fatti, considerazioni, dispositivo (formula della sentenza) e firma. Le considerazioni rappresentano la valutazione giuridica degli eventi da parte del tribunale.
Istruzioni:
- Comprendi i fatti: I fatti presentati includono fatti contestati e non contestati, le richieste delle parti, la procedura probatoria e la storia del processo.
- Inizia con le condizioni processuali: Verifica prima di tutto se le condizioni processuali (ad es. la competenza del tribunale) sono soddisfatte. Se non contestate, basta confermare che le condizioni sono state soddisfatte.
- Valutazione giuridica: Valuta le norme giuridiche rilevanti in base ai fatti affermati e giuridicamente rilevanti.
- Confrontati con i punti di vista giuridici delle parti.
- Tieni conto della distribuzione dell'onere della prova e valuta le prove liberamente, ma considera le regole di prova legalmente rilevanti. - Iura novit curia: La tua valutazione giuridica non deve necessariamente corrispondere alle argomentazioni giuridiche delle parti. Considera altre possibili linee di argomentazione.
- Riassunto: Riassumi alla fine delle tue considerazioni il risultato della tua valutazione giuridica.
- Risultato: Il testo generato dovrebbe essere strutturato, chiaro e nella forma di considerazioni tipiche di una sentenza svizzera."

{Fatti della sentenza svizzera}:
{INPUT FROM THE VALIDATION SET}

{Considerazioni}:

K.2.2 Leading Decision Summarization (LDS)

In the LDS task, we used only one German prompt because the output (the regeste) is always in German in our dataset. Despite the input being multilingual, models tend to generate in the prompt's language, regardless of the input's language.

Leading Decision Summarization (LDS)
Ziel: Generiere eine Regeste basierend auf einem Schweizer Gerichtsurteils.
Hintergrund: Ein Schweizer Gerichtsurteil setzt sich aus Sachverhalt, Erwägungen und Dispositiv zusammen. Die Regeste dient als Kurzzusammenfassung und beinhaltet Leitsätze des Urteils. Nur Leitentscheide haben eine Regeste.
Anweisung:
1. Sachverhalt: Lies und verstehe den gegebenen Sachverhalt.
2. Erwägungen: Analysiere die Erwägungen, um die Hauptargumente und Gründe zu identifizieren.
3. Dispositiv: Beachte das Dispositiv, da es das endgültige Urteil enthält.

4. Erstelle die Regeste: Die Regeste sollte aus drei sehr kurzen Teilen bestehen: a. Zitiere die wichtigsten relevanten Artikelziffern (ohne den Artikeltitel). b. Nenne kurze, relevante, deskriptive Keywords, über die Thematik des Falls. c. Formuliere einen sehr kurzen Fliesstext, der die wichtigsten Erwägungen zitiert und kurz zusammenfasst.
Output: Die Regeste sollte eine klare und strukturierte Kurzzusammenfassung des Urteils bieten, die aus zitierten Artikeln, Keywords und einem sehr kurzen Fliesstext besteht.

{Gegebener Sachverhalt, Erwägungen und Dispositiv}:
{INPUT FROM THE VALIDATION SET}

{Regeste auf Deutsch}:

L Example Generations

Tables 30 and 31 show excerpts of examples produced by the best model for the CVG and LDS tasks, respectively.

Table 30: Nine examples of generated considerations by $mT5_{Large}$ with input length 2048 across three languages, showcasing high, average, and low scored outputs in CVG Task

Target	considerations: Erwägungen: 1. 1.1 Der angefochtene Entscheid ist in Anwendung von Sozialversicherungsrecht ergangen. Die Sozialversicherungsrechtliche Abteilung des Verwaltungsgerichts beurteilt gemäss Art. 57 des Bundesgesetzes vom 6. Oktober 2000 über den Allgemeinen Teil des Sozialversicherungsrechts (ATSG; SR 830.1) i.V.m. Art. 54 Abs. 1 lit. a des kantonalen Gesetzes vom 11. Juni 2009 über die Organisation der Gerichtsbehörden und der Staatsanwaltschaft (GSOG; BSG 161.1) Beschwerden gegen solche Entscheide. Die Beschwerdeführerin ist im vorinstanzlichen Verfahren mit ihren Anträgen nicht durchgedrungen, durch den angefochtenen Entscheid berührt und hat ein schutzwürdiges Interesse an dessen Aufhebung, weshalb sie zur Beschwerde befugt ist (Art. 59 ATSG). Die örtliche Zuständigkeit ist gegeben (Art. 58 ATSG). Da auch die Bestimmungen über Frist (Art. 60 ATSG) sowie Form (Art. 61 lit. b ATSG; Art. 81 Abs. 1 i.V.m. Art. 32 des kantonalen Gesetzes vom 23. Mai 1989 Urteil des Verwaltungsgerichts des Kantons Bern vom 6. Dez. 2022, UV/21/844, Seite 4 über die Verwaltungsrechtspflege [VRPG; BSG 155.21]) eingehalten sind, ist auf die Beschwerde einzutreten. 1.2 Anfechtungsobjekt bildet der Einspracheentscheid vom 5. November 2021 (act. II 45). Streitig und zu prüfen ist, ob die Beschwerdegegnerin ihre Leistungen im Zusammenhang mit dem Ereignis vom 19. Juni 2020 zu Recht per 30. September 2020 eingestellt hat. 1.3 Die Abteilungen urteilen gewöhnlich in einer Kammer bestehend aus drei Richterinnen oder Richtern (Art. 56 Abs. 1 GSOG). 1.4 Das Gericht überprüft den angefochtenen Entscheid frei und ist an die Begehren der Parteien nicht gebunden (Art. 61 lit. c und d ATSG; Art. 80 lit. c Ziff. 1 und Art. 84 Abs. 3 VRPG). 2. 2.1 Die Zusprechung von Leistungen der obligatorischen Unfallversicherung setzt grundsätzlich das Vorliegen eines Berufsunfalles, eines Nichtberufsunfalles oder einer Berufskrankheit (Art. 6 Abs
Predicted	considerations: Erwägungen: 1. 1.1 Der angefochtene Entscheid ist in Anwendung von Sozialversicherungsrecht ergangen. Die Sozialversicherungsrechtliche Abteilung des Verwaltungsgerichts beurteilt gemäss Art. 57 des Bundesgesetzes über den Allgemeinen Teil des Sozialversicherungsrechts vom 6. Oktober 2000 (ATSG; SR 830.1) i.V.m. Art. 54 Abs. 1 lit. a des kantonalen Gesetzes über die Organisation der Gerichtsbehörden und der Staatsanwaltschaft vom 11. Juni 2009 (GSOG; BSG 161.1) Beschwerden gegen solche Entscheide. Die Beschwerdeführerin ist im vorinstanzlichen Verfahren mit ihren Anträgen nicht durchgedrungen, durch den angefochtenen Entscheid berührt und hat ein schutzwürdiges Interesse an dessen Aufhebung, weshalb sie zur Beschwerde befugt ist (Art. 59 ATSG). Die örtliche Zuständigkeit ist gegeben (Art. 69 Abs. 1 lit. a des Bundesgesetzes über die Unfallversicherung vom 20. Dezember 1946 [UVG; SR 832.20]). Da auch die Bestimmungen über Frist (Art. 60 ATSG) sowie Form (Art. 61 lit. b ATSG; Art. 81 Abs. 1 i.V.m. Art. 32 des kantonalen Gesetzes über die Verwaltungsrechtspflege vom 23. Mai 1989 [VRPG; BSG 155.21]) eingehalten sind, ist auf die Beschwerde einzutreten. 1.2 Das Gericht überprüft den angefochtenen Entscheid frei und ist an die Begehren der Parteien nicht gebunden (Art. 61 lit. c und d ATSG; Art. 80 lit. c Ziff. 1 und Art. 84 Abs. 3 VRPG). 1.3 Der Beschwerdeführer hat am vorinstanzlichen Verfahren teilgenommen, ist durch den angefochtenen Entscheid besonders berührt und hat ein schutzwürdiges Interesse an dessen Aufhebung oder Änderung. Auf die Beschwerde ist einzutreten. 2. 2.1 Die Beschwerdegegnerin hat die Bestimmungen auf den Anspruch auf Heilbehandlung und Taggeld in Form von Heilbehandlung zufolge ihrer im August 2021 (act. II 42 und 46) eingereichten medizinischen Berichte zutreffend wiedergegeben. Darauf wird verwiesen. 2.2 Nach der Rechtsprechung ist der Beschwerdeführer durch den angefochtenen Entscheid berührt und haben ein schutzwürdiges
Scores	BERT: 91.69, BLEU: 84.17, METEOR: 62.43, ROUGE-1/2/L: 75.49 / 65.78 / 75.49
Target	considerations: EN DROIT 1. Conformément à l'art. 134 al. 1 let. a ch. 2 de la loi sur l'organisation judiciaire, du 26 septembre 2010 (LOJ - E 2 05) en vigueur dès le 1 er janvier 2011, la chambre des assurances sociales de la Cour de justice connaît, en instance unique, des contestations prévues à l'art. 56 de la loi fédérale sur la partie générale du droit des assurances sociales, du 6 octobre 2000 (LPGA - RS 830.1) relatives à la loi fédérale sur l'assurance-invalidité du 19 juin 1959 (LAI - RS 831.20).![endif]>![if> Sa compétence pour juger du cas d'espèce est ainsi établie. 2. À teneur de l'art. 1 al. 1 LAI, les dispositions de la LPGA s'appliquent à l'assurance-invalidité (ci-après : AI), à moins que la loi n'y déroge expressément.![endif]>![if> Toutefois, les modifications légales contenues dans la LPGA constituent, en règle générale, une version formalisée dans la loi de la jurisprudence relative aux notions correspondantes avant l'entrée en vigueur de la LPGA ; il n'en découle aucune modification du point de vue de leur contenu, de sorte que la jurisprudence développée à leur propos peut être reprise et appliquée (ATF 130 V 343 consid. 3). 3. Les modifications de la loi fédérale sur l'assurance-invalidité du 21 mars 2003 (4 ème révision), du 6 octobre 2006 (5 ème révision) et du 18 mars 2011 (révision 6a), entrées en vigueur le 1 er janvier 2004, respectivement, le 1 er janvier 2008 et le 1 er janvier 2012, entraînent la modification de nombreuses dispositions légales dans le domaine de l'assurance-invalidité.
Predicted	considerations: EN DROIT 1. Conformément à l'art. 134 al. 1 let. a ch. 4 de la loi sur l'organisation judiciaire, du 26 septembre 2010 (LOJ; RS E 2 05) en vigueur dès le 1 er janvier 2011, la Chambre des assurances sociales de la Cour de justice connaît, en instance unique, des contestations prévues à l'art. 56 de la loi fédérale sur la partie générale du droit des assurances sociales, du 6 octobre 2000 (LPGA; RS 830.1) relatives à la loi fédérale sur l'assurance-invalidité du 19 juin 1959 (LAI; RS 831.20).![endif]>![if> Sa compétence pour juger du cas d'espèce est ainsi établie.![endif]>![if> 2. Les dispositions de la LPGA s'appliquent à l'assurance-invalidité, à moins que la loi n'y déroge expressément.![endif]>![if> Toutefois, les modifications légales contenues dans la LPGA constituent, en règle générale, une version formalisée dans la loi de la jurisprudence relative aux notions correspondantes avant l'entrée en vigueur de la LPGA; il n'en découle aucune modification du point de vue de leur contenu, de sorte que la jurisprudence développée à leur propos peut être reprise et appliquée (ATF 130 V 230 consid. 1.1; 335 consid. 1.2; ATF 129 V 4 consid. 1.2; ATF 129 V 4 consid. 1.2; ATF 127 V 467 consid. 1, 126 V 136 consid. 4b et les références). 3. A teneur de l'art. 17 al. 1 de la loi fédérale sur la partie générale du droit des assurances sociales, du 6 octobre 2000 (LPGA; RS 830.1), les modifications légales contenues dans la LPGA constituent, en règle générale, une version formalisée dans la LAI.
Scores	BERT: 91.27, BLEU: 85.8, METEOR: 68.12, ROUGE-1/2/L: 78.29 / 66.82 / 78.29
Target	considerations: in diritto In ordine 2.1. La presente vertenza non pone questioni giuridiche di principio e non è di rilevante importanza (ad esempio per la difficoltà dell'istruttoria o della valutazione delle prove). Il TCA può dunque decidere nella composizione di un Giudice unico ai sensi dell'articolo 49 cpv. 2 LOG (STF 9C_699/2014 del 31 agosto 2015; 8C_855/2010 dell'11 luglio 2011; 9C_211/2010 del 18 febbraio 2011). Nel merito 2.2. Oggetto del contendere è sapere se a ragione l'Ufficio AI ha soppresso all'assicurato il diritto ad una rendita intera con effetto dal 1. settembre 2015 (cfr. consid. 1.3). L'insorgente postula l'annullamento della decisione impugnata del 10 maggio 2017 e quindi il mantenimento del diritto alla rendita intera. 2.3. Secondo l'art. 4 cpv. 1 LAI in relazione con gli art. 7 e 8 della LPGA, con invalidità s'intende l'incapacità al guadagno presunta permanente o di rilevante durata, cagionata da un danno alla salute fisica o psichica, conseguente ad infermità congenita, malattia o infortunio. Gli elementi fondamentali dell'invalidità, secondo la surriferita definizione, sono quindi un danno alla salute fisica o psichica conseguente a infermità congenita, malattia o infortunio, e la conseguente incapacità di guadagno. Occorre quindi che il danno alla salute abbia cagionato una diminuzione della capacità di guadagno, perché il caso possa essere sottoposto all'assicurazione per l'invalidità (Duc, L'assurance invalidité, in: Meyer (Hrsg.), Schweizerisches Bundesverwaltungsrecht, Band XIV, Soziale Sicherheit, 2a ed., Basilea/Ginevra/Monaco di Baviera 2007, pag. 1411, n. 46). Giusta l'art. 28 cpv.
Predicted	considerations: in diritto In ordine 2.1. La presente vertenza non pone questioni giuridiche di principio e non è di rilevante importanza (ad esempio per la difficoltà dell'istruttoria o della valutazione delle prove). Il TCA può dunque decidere nella composizione di un Giudice unico ai sensi degli articoli 26 c cpv. 2 della Legge organica giudiziaria civile e penale e 2 cpv. 1 della Legge di procedura per le cause davanti al Tribunale delle assicurazioni (cfr. STF 9C_211/2010 del 18 febbraio 2011; STF 9C_792/2007 del 7 novembre 2008; STF H 180/06 e H 183/06 del 21 dicembre 2007; STFA I 707/00 del 21 luglio 2003; STFA H 335/00 del 18 febbraio 2002; STFA U 347/98 del 10 ottobre 2001, pubblicata in RDAT I-2002 pag. 190 seg.; STFA H 304/99 del 22 dicembre 2000; STFA I 623/98 del 26 ottobre 1999). Nel merito 2.2. Oggetto della lite è sapere se l'Ufficio AI ha diritto alla rendita intera intera dal 1. settembre 2015 (doc. AI 122/445-450), a causa dell'invalidità della persona assicurata. Secondo l'art. 4 cpv. 1 LAI in relazione con gli art. 7 e 8 della LPGA, con invalidità s'intende l'incapacità al guadagno presunta permanente o di rilevante durata, cagionata da un danno alla salute fisica o psichica, conseguente a infermità congenita, malattia o infortunio. Gli elementi fondamentali dell'invalidità, secondo la surriferita definizione, sono quindi un danno alla salute fisica o psichica conseguente a infermità congenita, malattia o infortunio, e la conseguente incapacità di guadagno. Occorre quindi che il danno alla salute abbia cagionato una diminuzione della capacità di guadagno, perché il caso possa
Scores	BERT: 86.57, BLEU: 81.39, METEOR: 65.76, ROUGE-1/2/L: 68.54 / 58.87 / 66.85

Target	considerations: Das Bundesverwaltungsgericht zieht in Erwägung: 1. Nach Art. 53 Abs. 1 KVG kann gegen Beschlüsse der Kantonsregierungen nach Art. 39 KVG beim Bundesverwaltungsgericht Beschwerde geführt werden. Der angefochtene Beschluss Nr. 617/2019 des Regierungsrats des Kantons Zürich vom 26. Juni 2019 wurde gestützt auf Art. 39 KVG er- lassen. Das Bundesverwaltungsgericht ist deshalb zur Beurteilung der Be- schwerde zuständig (vgl. auch Art. 90a Abs. 2 KVG). Die Beschwerdefüh- rerin hat am vorinstanzlichen Verfahren teilgenommen, ist als Adressatin durch den angefochtenen Regierungsratsbeschluss besonders berührt und hat insoweit an dessen Aufhebung beziehungsweise Abänderung ein schutzwürdiges Interesse (Art. 48 Abs. 1 VwVG). Sie ist daher zur Be- schwerde legitimiert. Auf die frist- und formgerecht erhobene Beschwerde ist, nachdem auch der Kostenvorschuss rechtzeitig geleistet wurde, einzu- treten (Art. 50 Abs. 1, Art. 52 Abs. 1 und Art. 63 Abs. 4 VwVG). Das Haupt- rechtsbegehren der Beschwerdeführerin richtet sich auf Aufhebung der Zif- fer I des Dispositivs des angefochtenen Beschlusses. Soweit sie dabei Be- zug auf «C. Neue Regelung der Verlegungen ab Geburtshaus» nimmt, ist das im Lichte der Begründung der Beschwerde nicht so zu verstehen, dass sie auch die Aufhebung eines Teils der Begründung des angefochtenen Beschlusses beantragt. Der Antrag der Vorinstanz, auf die Beschwerde sei teilweise nicht einzutreten, weil sie sich auch gegen die Begründung des angefochtenen RRB richte, ist daher unbegründet. 2. 2.1 Mit Beschwerde gegen einen Spitallistenentscheid im Sinne von Art. 39 Abs. 1 KVG kann die Verletzung von Bundesrecht unter Einschluss des Missbrauchs oder der Überschreitung des Ermessens und die unrichtige oder unvollständige Feststellung des rechtserheblichen Sachverhalts ge- rügt werden. Nicht zulässig ist hingegen die Rüge der
Predicted	considerations: Das Bundesverwaltungsgericht zieht in Erwägung: 1. 1.1 Gemäss Art. 31 des Verwaltungsgerichtsgesetzes vom 17. Juni 2005 (VGG, SR 173.32) beurteilt das Bundesverwaltungsgericht Beschwerden gegen Verfügungen nach Art. 5 des Verwaltungsverfahrensgesetzes vom 20. Dezember 1968 (VwVG, SR 172.021), sofern keine Ausnahme nach Art. 32 VGG vorliegt. Als Vorinstanzen gelten die in Art. 33 und 34 VGG genannten Behörden. Dazu gehören u.a. Verfügungen im Sinne von Art. 5 VwVG. Das Bundesverwaltungsgericht ist somit zuständig für die Beurteilung der vorliegenden Beschwerde und entscheidet auf dem Gebiet des Gesundheits- schutzes endgültig, ausser bei Eintreten auf einen Entscheid, der eine Verfügung im Sinne von Art. 5 VwVG betrifft. Eine Ausnahme nach Art. 32 VGG liegt nicht vor. 1.2 Das Bundesverwaltungsgericht ist zur Beurteilung der vorliegenden Beschwerde zuständig. 1.3. Gegen Entscheide des Regierungsrats ist ein Beschwerdeverfahren zulässig, wenn dieses eine Rechtsmittelinstanz ist und grundsätzlich der Rechts- schutzinteresse der Beschwerde-berechtigten gegeben ist (vgl. art. 32 VGG; BGE 138 II 274 E. 5.1 S. 280). 1.4 Das Bundesverwaltungsgericht wendet im Beschwerde- verfahren das Bundesrecht von Amtes wegen an (Art. 106 Abs. 1 des Bundesgesetzes vom 20. Dezember 1968 über das Verwaltungsverfahren [VwVG, SR 172.021]). Es wird gemäss Art. 62 Abs. 4 VwVG zur Beurteilung von Beschwerden in Zivilsachen gemäss Art. 82 Abs. 1 und Art. 86 Abs. 1 Bst. d bis Bst. e VwVG grundsätzlich nur geprüft, wenn diese vor Bundesgericht nicht mehr vorgetragen wurden und keine Verletzung des Op-fers gegeben wurde. Das Bundesverwal- tungsgericht wendet im Beschwerdeverfahren das Bundesrecht von Amtes wegen an. Es ist gemäss Art. 62 Abs. 4 VwVG an die Begründung der Begehren nicht gebunden und kann die Beschwerde auch aus anderen als den geltend gemachten Gründen gutheissen
Scores	BERT: 75.72, BLEU: 67.17, METEOR: 31.26, ROUGE-1/2/L: 26.43 / 8.84 / 25.23

Target	considerations: Considérant en droit: 1. Selon l'art. 5 al. 1 des dispositions du RPAC relatives à la Commission de recours individuel, la décision rendue par cette autorité peut faire l'objet d'un recours de droit administratif auprès du Tribunal cantonal dans les trente jours suivant la communication de la décision motivée, conformément à la loi cantonale du 28 octobre 2008 sur la procédure administrative (LPA-VD; BLV 173.36). D'après l'art. 92 al. 1 LPA-VD, le Tribunal cantonal connaît des recours contre les décisions et décisions sur recours rendues par les autorités administratives, lorsque la loi ne prévoit aucune autre autorité pour en connaître. En l'espèce, le recours a été déposé en temps utile (art. 5 al. 1 RPAC) et il satisfait aux autres conditions formelles de recevabilité (art. 75 al. 1 let. a, 79 al. 1 et 99 LPA-VD, applicables par renvoi de l'art. 5 al. 1 RPAC), si bien qu'il convient d'entrer en matière. 2. a) La classification salariale du poste qu'occupe le recourant est litigieuse en l'occurrence. b) L'organisation de l'administration fait partie des tâches propres des autorités communales (art. 2 de la loi vaudoise du 28 février 1956 sur les communes [LC; BLV 175.11]). Selon cette loi, il incombe au Conseil général ou communal de définir le statut des collaborateurs communaux et la base de leur rémunération (art. 4 al. 1 ch. 9 LC), la municipalité ayant la compétence de nommer les collaborateurs et employés de la commune, de fixer leur traitement et d'exercer le pouvoir disciplinaire (art. 42 al. 1 ch. 3 LC). Conformément à l'art. 33 al. 1 RPAC, le traitement du fonctionnaire comprend le traitement de base (let. a), les allocations
Predicted	considerations: Considérant en droit: 1. 1.1 Le personnel doit maintenir, par contrat de droit privé, une commune, un partenaire public, pour assurer la bonne exécution de leurs missions spéciales, ainsi qu'un minimum d'administration (art. 10 al. 1 du RPAC). L'art. 10 al. 2 RPAC prévoit que les fonctionnaires dans toute la ville de Lausanne sont soumis à la qualification communale et leur remplacement a lieu par des fonctions communales cantonales, comme applicable en la matière. Elle constitue en revanche une décision sujette à recours auprès de la juridiction compétente (art. 20 al. 2 et art. 30 al. 1 de la loi sur le personnel des communes du 15 novembre 1976 (LPAC). 1.2 Selon la jurisprudence relative aux dispositions transitoires applicables au moment de la notification de la décision pour connaître du jugement du service, la décision de rémunération fait apparaître un changement de système de rémunération selon le RPAC. Cette disposition, qui comprend ainsi des critères objectifs et suffisants, permet au fonctionnaire de connaître de tous les aspects de la classification (Tappy, CPC, n. 1 ad art. 10 LPAC), ainsi que par le service de réduire son salaire. En l'espèce, l'autorité intimée a constaté que l'établissement des faits importants du dossier et de la satisfaction exigée sur le plan d'administration devait être complété dans la mesure où cela n'était pas possible, ce qui constituait effectivement un changement de système (Tappy, op. cit., n. 13 ad art. 10 LPAC). Il a constaté que le Service du personnel avait fixé le niveau de fonction attribué au fonctionnaire pour son titre de travail. Il a en outre estimé que cette classification était une décision de mise en œuvre déléguée à celui qui a entrepris une modification légale ou réglementaire du service. Parcon
Scores	BERT: 75.74, BLEU: 66.19, METEOR: 31.69, ROUGE-1/2/L: 28.32 / 4.21 / 26.01

Target	considerations: Diritto: 1. Il ricorso in materia di diritto pubblico può essere presentato per violazione del diritto, così come stabilito dagli art. 95 e 96 LTF. Il Tribunale federale fonda la sua sentenza sui fatti accertati dall'autorità inferiore (art. 105 cpv. 1 LTF). Può scostarsi da questo accertamento solo qualora esso sia avvenuto in modo manifestamente inesatto, ovvero arbitrario (DTF 134 V 53 consid. 4.3 pag. 62), oppure in violazione del diritto ai sensi dell'art. 95 LTF (art. 105 cpv. 2 LTF). Salvo i casi in cui tale inesattezza sia lampante, la parte ricorrente che intende contestare i fatti accertati dall'autorità inferiore deve spiegare, in maniera circostanziata, per quale motivo ritiene che le condizioni di una delle eccezioni previste dall'art. 105 cpv. 2 LTF sarebbero realizzate (cfr. DTF 133 II 249 consid. 1.4.3 pag. 254 con riferimento). 2. Oggetto del contendere è sapere se sia lesivo del diritto federale il giudizio cantonale, che conferma la decisione su opposizione, la quale a sua volta nega all'assicurato le indennità per insolvenza per non avere rivendicato tempestivamente le proprie pretese salariali. 3. 3.1. Il Tribunale cantonale delle assicurazioni, esposto lo svolgimento del processo, le norme e la prassi ritenute applicabili, ha accertato che il ricorrente è stato legato da un contratto di lavoro con la ditta B._ SA, dichiarata fallita il 15 maggio 2018, dal 15 aprile 2017 al 18 settembre 2017 come informatico e con un salario di fr. 3'360.- mensili. L'ultimo giorno di lavoro effettuato è stato il 1° settembre 2017. Il salario è stato saldato dal 15 al 30 aprile 2017. Per contro, non è stato pagato per i mesi di maggio, giugno, luglio e agosto, nonché per il 1° settembre 2017. Il
Predicted	considerations: Diritto: 1. 1.1. La procedura ricorsuale in materia di diritto pubblico è disciplinata dal diritto federale (art. 82 lett. a LTF) e dalla legge federale, in vigore dal 1° gennaio 2003, sull'assicurazione contro la disoccupazione: si tratta infatti di una causa di diritto pubblico contro una decisione finale adottata dalla Corte cantonale delle assicurazioni, entrata in vigore il 1° gennaio 2003. Il ricorso presentato dall'assoggettato al diritto di ricorrere (art. 89 cpv. 1 LTF) è di massima ammissibile dinanzi al Tribunale federale (art. 89 cpv. 1 LTF). 1.2. Il ricorso in materia di diritto pubblico può essere presentato per violazione del diritto, così come stabilito dagli art. 95 e 96 LTF, oltre che per violazione del diritto ai sensi dell'art. 95 LTF. Il Tribunale federale applica il diritto federale (art. 106 cpv. 1 LTF), senza essere vincolato né dai motivi addotti nel ricorso (art. 106 cpv. 2 LTF). Per contro, nel ricorso in materia di diritto pubblico il Tribunale federale esamina d'ufficio e con piena cognizione l'ammissibilità dei gravami che gli vengono sottoposti (DTF 133 III 439 consid. 1.3). 1.3. La critica del giudizio impugnato esplica degli effetti (art. 105 cpv. 1 LTF). Non è ammissibile che i ricorsi in materia di diritto pubblico possano essere decisi in base al diritto federale, ai sensi dell'art. 95 LTF, senza istruttoria (art. 97 cpv. 1 LTF, Art. 105 cpv. 2 LTF). 1.3. Il Tribunale federale esamina d'ufficio e con piena cognizione l'ammissibilità dei ricorsi che gli vengono sottoposti (DTF 133 III 439 consid. 1.3). 1.4. Con il ricorso in materia di pubblico contro la decisione di primo grado, il Tribunale cantonale ha emesso
Scores	BERT: 75.79, BLEU: 66.28, METEOR: 30.29, ROUGE-1/2/L: 37.74 / 20.43 / 36.48

Target	considerations: Das Versicherungsgericht zieht in Erwägung: 1. Streitig und zu prüfen ist der Rentenanspruch der Beschwerdeführerin. - 3 - 2. Am 1. Januar 2022 sind die Änderungen betreffend Weiterentwicklung der IV (WEIV) in Kraft getreten. Weder dem IVG noch der IVV sind besondere Übergangsbestimmungen betreffend die Anwendbarkeit dieser Änderun- gen im Hinblick auf nach dem 1. Januar 2022 beurteilte mögliche Ansprü- che des Zeitraums bis zum 31. Dezember 2021 zu entnehmen. Es sind daher nach den allgemeinen übergangsrechtlichen Grundsätzen jene Be- stimmungen anzuwenden, die bei der Erfüllung des rechtlich zu ordnenden oder zu Rechtsfolgen führenden Tatbestands Geltung haben beziehungs- weise hatten (vgl. Urteil des Bundesgerichts 8C_136/2021 vom 7. April 2022 E. 3.2.1 mit Hinweis unter anderem auf BGE 144 V 210 E. 4.3.1 S. 213). Da vorliegend Leistungen mit Anspruchsbeginn vor dem 1. Januar 2022 streitig sind, ist für deren Beurteilung die bis zum 31. Dezember 2021 geltende Rechtslage massgebend. 3. 3.1. Die Beschwerdegegnerin stützte sich in der Verfügung vom 14. Juni 2022 (Vernehmlassungsbeilage [VB] 162) in medizinischer Hinsicht im Wesent- lichen auf das bidisziplinäre Gutachten der SMAB vom 3. August 2021 (VB 139; Fachdisziplinen: Neurologie und Psychiatrie). 3.2. Die SMAB-Gutachterinnen stellten in der interdisziplinären Gesamtbeurtei- lung die folgenden Diagnosen mit Auswirkungen auf die Arbeitsfähigkeit (VB 139.1 S. 6): "1. Neuropathische Schmerzen an der Innenseite des Mittelfingers links und der Handinnenfläche links (Schnittverletzung am Mittelfinger links mit initialer Nervenverletzung 19.11.2014, epineurale und weitere mehrfache operative Eingriffe 2014-2018), 2. Chronische Schmerzstörung mit somatischen und psychischen (ICD-10: F45.41)". Ohne Auswirkung auf die Arbeitsfähigkeit sei eine rezidivierende depres- sive Störung, gegenwärtig remittiert. Die Gutachterinnen hielten fest, ins- gesamt bestehe aufgrund der psychiatrischen und neurologi
Predicted	considerations: Das Gericht zieht in Erwägung: 1. Gemäss Art. 87 Abs. 2 des Bundesgesetzes über die Unfallversicherung (UVG) besteht Anspruch auf eine ganze Rente, wenn die versicherte Person mindestens zu 66 2/3 Prozent, auf eine halbe Rente, wenn sie mindestens zu 50 Prozent oder auf eine Viertelsrente, wenn sie mindestens zu 40 Prozent invalid ist. In Härtefällen besteht gemäss Art. 87 Abs. 2 UVG auch ein Anspruch auf eine halbe Rente, wenn sie mindestens zu 50 Prozent, wenn sie mindestens zu 40 Prozent invalid ist (Art. 8 Abs. 1 UVG). Anspruch auf eine Rente haben gemäss Art. 28 Abs. 2 des Bundes- gesetzes über die Invalidenversicherung (IVG) Versicherte, die im Sinne von Art. 16 Abs. 1 IVG in Verbindung mit Art. 28 bis Abs. 1 IVG in Verbindung mit Abs. 1 und 2 IVG auch dann Anspruch auf eine Viertelsrente, wenn sie mindestens zu 40 Prozent invalid sind (Art. 28 Abs. 2 IVG i.V.m. Art. 16 ATSG). 2. 2.1 Anspruch auf eine Rente haben gemäss Art. 28 Abs. 1 IVG Versicherte, die: a. ihre Erwerbsfähigkeit oder die Fähigkeit, sich im Aufgabenbereich zu betä- tigen, nicht durch zumutbare Eingliederungsmassnahmen wieder herstellen, erhalten oder verbessern können; b. während eines Jahres ohne wesentlichen Unterbruch durchschnittlich min- destens 40 Prozent arbeitsunfähig (Art. 6 ATSG) gewesen sind; und c. nach Ablauf dieses Jahres zu mindestens 40 Prozent invalid (Art. 8 ATSG) sind. 2.2 Die Beschwerdeführerin bringt vor, die ärztlich eingeholten ärztlichen Berichte seien als diagnostisch zu qualifizieren. Das trifft vorliegend nicht zu. Ihr Gesundheitszustand sei gemäss Abklärungen vom RAD mit einer Invalidität von mindestens 40 Prozent zu vereinbaren. Die Leistungsfähigkeit sei in Art. 16 ATSG eingetreten. 2.2 Die Leistungsfähigkeit sei in Art. 16 ATSG i.V.m. Art. 28 Abs. 1 IVG
Scores	BERT: 62.42, BLEU: 51.67, METEOR: 19.69, ROUGE-1/2/L: 17.58 / 2.56 / 17.58
Target	considerations: le conseil de B_ a rappelé que celle-ci contestait que son frère aurait été en droit de loger gratuitement au 1_ dès le mois d'avril 2015. Afin d'éviter la notification d'une poursuite interruptive de la prescription, A_ était invité à signer une déclaration de renonciation à la prescription concernant les loyers qui, selon B_, étaient dus à la succession de leurs défunts parents. Etait annexée à ce courrier une déclaration de renonciation à invoquer la prescription libellée en ces termes : " Par la présente, A_ déclare renoncer à invoquer toute exception de prescription vis-à-vis des successions de [feu D_ et feu C_], eu égard à toutes prétentions éventuelles de loyers et charges alors qu'il logeait chemin 1_ à _ Fribourg. Cette déclaration de renonciation est valable jusqu'au 31 décembre 2020. La présente déclaration ne vaut en aucun cas reconnaissance de dette ou de responsabilité quelconque. La présente déclaration n'a d'effet que si la prescription n'est pas déjà acquise à ce jour ". A_ n'a pas donné suite à ce courrier. f. Par réquisition de poursuite du 7 avril 2020, reçue le lendemain par l'Office cantonal des poursuites de Genève (ci-après : l'Office), Me H_, agissant en qualité de représentant du créancier, à savoir " Pour l'hoirie de feu C_, décédée le _ 2019, Madame B_, Rte 2_, _ Fribourg ", a requis la poursuite de A_, domicilié avenue 3_, _ Genève, pour la somme de 112'500 fr. avec intérêts à 5% dès le 1 er mai 2017, réclamée à titre de " Loyers et charges dus pour le logement du 1_ à _ Fribourg d'avril 2015 à mai 2019 (50 x
Predicted	considerations: EN DROIT 1. 1.1 Le litige porte sur la désignation d'un représentant de la communauté héréditaire et la désignation d'un administrateur de la régie G_ SA (art. 602 al. 3 CC). Le motif de la désignation du représentant est incontestable. Selon l'art. 602 al. 5 CC, la désignation d'un représentant de la communauté héréditaire est donc une décision de dernière instance cantonale au sens des art. 80 al. 1 et 82 al. 1 CPC, qui doit être notifiée à l'unique partie plaignante qui a pris part à la procédure (ATF 110 IV 92 consid. 1a; arrêt du Tribunal fédéral X. du 27 juillet 2004 ; ATF 99 III 58 consid. 1; ATF 118 Ia 397 consid. 1b; Tappy, CPC-VD, n. 3 ad art. 602 CC ; Replin, Le représentant de la communauté héréditaire, 5 ème éd., Lausanne 2013, p. 569; ATF 118 IV 286 consid. 2a; TF 6B_211/2007 du 29 janvier 2008, consid. 5.3; ATF 117 IV 29 consid. 3b; TF 8B_44/2007 du 15 août 2007, consid. 3.2; Tappy, Procédure civile, tome II, ad art. 602; Piquerez, in : Kuhn/Jeanneret [éd.], Basler Kommentar, n. 6 ad art. 602; TF 7B_51/2007 du 1 er janvier 2008, consid. 3.2; ATF 130 III 136 consid. 1.2.1; TF 9C_438/2007 du 30 septembre 2007, consid. 3.1; ATF 134 III 102 consid. 3.1; TF 6B_71/2007 du 24 août 2007, consid. 5b; TF 9C_792/2007 du 11 août 2007, consid. 4.2). 1.2 La désignation d'un représentant et de l'administration régulière de l'ensemble de la succession ont été rejetées. 2. 2.1. L'art. 602 al. 3 CC ouvre un recours au Tribunal cantonal, sans être lié par l
Scores	BERT: 63.90, BLEU: 46.65, METEOR: 19.12, ROUGE-1/2/L: 11.52 / 1.64 / 10.91
Target	considerations: A.c. Con progetto di decisione dell'8 gennaio 2014 l'UAI ha concluso, in via di riconsiderazione, per la soppressione della mezza rendita d'invalidità con effetto retroattivo al 30 aprile 2011. In sede di audizione, l'UAI ha esperito nuovi accertamenti, in particolare la perizia pluridisciplinare dell'8 aprile 2016. Con decisione del 14 luglio 2016 l'UAI ha riconosciuto che al momento dell'assegnazione della mezza rendita d'invalidità è incorso in un errore manifesto avendo applicato il "Prozentvergleich" in luogo del metodo ordinario di confronto dei redditi, da cui risulta un grado d'invalidità del 26 %. Ritenuta l'importanza della sua correzione, l'UAI, riconsiderando le decisioni 14 dicembre 2006 e 14 maggio 2007 nonché la comunicazione del 20 marzo 2009, ha predisposto la soppressione del diritto alla mezza rendita invalidità con effetto retroattivo al 30 aprile 2011. B. A._ si è aggravato il 13 settembre 2016 al Tribunale delle assicurazioni del Cantone Ticino, che con giudizio del 17 maggio 2017 ha respinto il gravame. C. Il 19 giugno 2017 (timbro postale) A._ inoltra un ricorso in materia di diritto pubblico al Tribunale federale, cui chiede di ammettere il ricorso e riformare il giudizio cantonale nel senso di confermare il diritto ad almeno una mezza rendita d'invalidità e ordinare all'UAI di versare tutte le prestazioni trattenute dal momento della revisione ad oggi. L'UAI e l'Ufficio federale delle assicurazioni sociali rinunciano a determinarsi. Diritto: 1. Il ricorso in materia di diritto pubblico può essere presentato per violazione del diritto così come stabilito dagli art. 95 e 96 LTF. Il Tribunale federale esamina d'ufficio l'applicazione de l diritto federale (art. 106 cpv. 1 L
Predicted	considerations: Diritto: 1. Il Tribunale delle assicurazioni esamina d'ufficio e con piena cognizione l'ammissibilità dei gravami che gli vengono sottoposti (DTF 134 V 131 consid. 1.1 pag. 131; 133 V 477 consid. 1.1 pag. 481). In particolare, i gravami inoltrati già in tempo utile contro le decisioni di prima istanza rese dall'UAI in materia di assicurazioni sociali possono essere decisi sulla scorta degli atti, senza istruttoria (DTF 133 V 477 consid. 1.2 pag. 481). Il giudizio può essere reso sulla base degli atti, senza istruttoria (DTF 133 V 477 consid. 1.2 pag. 481; 133 V 477 consid. 1.2 pag. 481). 2. Il Tribunale delle assicurazioni esamina d'ufficio e con piena cognizione l'ammissibilità dei gravami che gli vengono sottoposti (DTF 133 V 477 consid. 1.2 pag. 481). 3. 3.1. Giusta l'art. 4 cpv. 1 LAI in relazione con gli art. 7 e 8 della LPGA, con invalidità s'intende l'incapacità al guadagno presunta permanente o di rilevante durata, cagionata da un danno alla salute fisica o psichica, conseguente a infermità congenita, malattia o infortunio. Gli elementi fondamentali dell'invalidità, secondo la surriferita definizione, sono quindi un danno alla salute fisica o psichica conseguente a infermità congenita, malattia o infortunio, e la conseguente incapacità di guadagno. Occorre quindi che il danno alla salute abbia cagionato una diminuzione della capacità di guadagno, perché il caso possa essere sottoposto all'assicurazione per l'invalidità (DTF 134 V 329 consid. 3.2.3 pag. 331; 133 V 477 consid. 3.3.1 pag. 481; 129 V 4 consid. 1.2 pag. 5 con rinvii). 3.2. Giusta
Scores	BERT: 67.86, BLEU: 55.76, METEOR: 25.41, ROUGE-1/2/L: 24.32 / 5.59 / 23.65

Table 31: Three examples of generated summaries by $mT5_{Base}$ with input length 4096, showcasing high, average, and low scored outputs in LDS Task

Target	regeste: Regeste Teilklage und negative Feststellungswiderklage (Art. 86 und 224 ZPO). Die Ausnahme vom Erfordernis der gleichen Verfahrensart gemäss Art. 224 Abs. 1 ZPO für negative Feststellungswiderklagen ist nicht auf den Fall beschränkt, dass es sich bei der Hauptklage um eine sogenannte echte Teilklage handelt, sondern gilt allgemein dann, wenn die Teilklage eine Ungewissheit zur Folge hat, die es rechtfertigt, die Feststellung des Nichtbestands einer Forderung oder eines Rechtsverhältnisses zu verlangen (E. 2).

Predicted	regeste: Regeste Voraussetzungen, unter denen eine Widerklage erhoben werden darf, wenn eine echte Teilklage beurteilt wird (E. 2.1). Zur Anwendung von Art. 224 Abs. 1 ZPO ist die Ausnahme vom Erfordernis der gleichen Verfahrensart gemäss Art. 224 Abs. 1 ZPO nicht auf diesen Fall beschränkt, sondern gilt allgemein dann, wenn die Teilklage eine Ungewissheit zur Folge hat, die es rechtfertigt, im Sinne von Art. 88 ZPO die Feststellung des Nichtbestands einer Forderung oder eines Rechtsverhältnisses zu verlangen (E. 2.2).
Scores	BERT: 88.03, BLEU: 30.45, METEOR: 43.78, ROUGE-1/2/L: 72.44 / 54.05 / 70.87
Target	regeste: Regeste Uferschutzplanung Wohlensee; Gesamtinteressenabwägung. Die Behörden nehmen bei der Genehmigung einer Sondernutzungsplanung eine umfassende Interessenabwägung vor (vgl. Art. 3 RPV ; E. 3.2). Würdigung des Gutachtens der Vogelwarte Sempach und der Stellungnahme des BAFU als Fachbehörde (E. 5.5, 6.3 und 6.4). Im zu beurteilenden Fall wird das öffentliche Interesse an einer ufernahen Wegführung (vgl. Art. 3 Abs. 2 lit. c RPG) relativiert, weil ein durchgehender Sichtschutz zum Schutz der im Gebiet vorkommenden Vogelarten erforderlich wäre, was Spaziergängern das Erleben der Uferlandschaft nur sehr beschränkt ermöglichen würde (E. 6.5 sowie 3.3). Das Gebiet des Wohlensees stellt eines von insgesamt 25 Wasser- und Zugvogelreservaten von nationaler Bedeutung dar (vgl. Art. 11 Abs. 2 JSG, Art. 2 Abs. 1 i.V.m. Anhang 1 WZVV).
Predicted	regeste: Regeste Art. 3 Abs. 2 lit. c RPG ; schweizerische Vogelwarte Sempach; Umweltschutz. Gutachten im Bereich des Umweltrechts, insbesondere von dem UVG/USG ergangenen Streitwert (E. 3.1). Ein öffentlicher Weg, der über Privatgrundstücke führt und mit dem Enteignungsrecht zugunsten des Gemeinwesens verbunden ist, stellt eine öffentlich-rechtliche Eigentumsbeschränkung dar (E. 3.3). Interessenabwägungen sind grundsätzlich frei (E. 3.4 und 3.5). Umfang des geplanten Uferweges; ein ausweisenes öffentliches Interesse an einem direkt dem Ufer entlangführenden Weg sowie an einer zumindest ufernahen Wegführung (E. 3.5).
Scores	BERT: 73.91, BLEU: 47.21, METEOR: 30.91, ROUGE-1/2/L: 32.75 / 10.63 / 32.75
Target	regeste: Regeste a Schwere Bedrängnis, grosse seelische Belastung und heftige Gemütsbewegung; Art. 48 lit. a Ziff. 2 und Art. 48 lit. c StGB. Abgrenzung dieser Strafmilderungsgründe untereinander (E. 2.1-2.3). Regeste b Mord, heftige Gemütsbewegung und grosse seelische Belastung; Art. 48 lit. c und Art. 112 StGB. Die für den Mord typische Skrupellosigkeit schliesst die Bejahung der Strafmilderungsgründe der heftigen Gemütsbewegung und der grossen seelischen Belastung aus (E. 2.5).
Predicted	regeste: Regeste Art. 112 und 49 Abs. 2 StGB ; tiefe Gewalt, Schwere Gewalt, Mord. Die Beeinträchtigung oder der schwere Gewalt ist einen psychischen Klumpen im Sinne des Art. 48 StGB, wenn beide Personen einen Beziehungskonflikt bilden und sich ein bestimmtes Verhältnis zwischen einseitigen Umständen des Tatvergehens und der Begründung der Tatbegehren haben (E. 2 und 3).
Scores	BERT: 69.3, BLEU: 12.14, METEOR: 14.72, ROUGE-1/2/L: 30.43 / 10.26 / 30.43

Resolving Legalese: A Multilingual Exploration of Negation Scope Resolution in Legal Documents

Ramona Christen[1] * **Anastassia Shaitarova**[2] **Matthias Stürmer**[1,3]

Joel Niklaus[1,3,4] *

[1]University of Bern [2]University of Zurich
[3]Bern University of Applied Sciences [4]Stanford University

Abstract

Resolving the scope of a negation within a sentence is a challenging NLP task. The complexity of legal texts and the lack of annotated in-domain negation corpora pose challenges for state-of-the-art (SotA) models when performing negation scope resolution on multilingual legal data. Our experiments demonstrate that models pre-trained without legal data underperform in the task of negation scope resolution. Our experiments, using language models exclusively fine-tuned on domains like literary texts and medical data, yield inferior results compared to the outcomes documented in prior cross-domain experiments. We release a new set of annotated court decisions in German, French, and Italian and use it to improve negation scope resolution in both zero-shot and multilingual settings. We achieve token-level F1-scores of up to 86.7% in our zero-shot cross-lingual experiments, where the models are trained on two languages of our legal datasets and evaluated on the third. Our multilingual experiments, where the models were trained on all available negation data and evaluated on our legal datasets, resulted in F1-scores of up to 91.1%.

1 Introduction

Negation scope resolution is an important research problem in the field of Natural Language Processing (NLP). It describes the detection of words that are affected by a negation cue (e.g. no or not) in a sentence, which is important for understanding its true meaning. Although this task is far from trivial, deep learning approaches have shown promising results (Khandelwal and Sawant, 2020; Shaitarova et al., 2020; Shaitarova and Rinaldi, 2021).

As with many NLP tasks, the largest amount of annotated data is available in English.[1] Multilingual datasets are less common and often not easily accessible. For example, on the huggingface hub, hosting most important open-source datasets, 4559 datasets are tagged as English. The next most common language is Chinese with 10 times fewer datasets for a total of 469.[2] In addition, much of the work conducted in the area of negation scope resolution has been done in the medical domain in order to automatically process clinical reports and discharge summaries (Szarvas et al., 2008). Other datasets consist of literary texts (Morante and Blanco, 2012) or more informal data such as online reviews (Konstantinova et al., 2012). The legal domain differs from all of the above in that it is often very complex (i.e., legalese) and uses highly specific vocabulary and knowledge that is not common outside the legal domain (Friedrich, 2021; Ruhl et al., 2017). This poses a challenge to any model tackling tasks in the legal domain. While a large amount of legal data is publicly available and has been annotated for various tasks (Chalkidis et al., 2021; Rasiah et al., 2023; Niklaus et al., 2021, 2023a; Brugger et al., 2023; Niklaus et al., 2023b; Chalkidis et al., 2022), *inter alia*, to the best of our

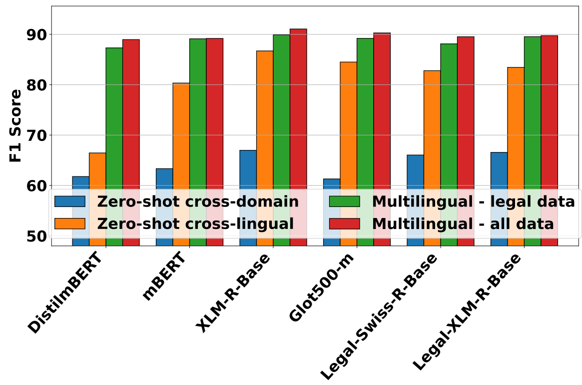


Figure 1: Results over main experiments from select models. For all results see Appendix B.

* Equal contribution.

[1](Mielke, 2016) analyzed all ACL conference proceedings from 2004, 2008, 2012, and 2016 and found that between 58% and 69% of papers only evaluated in English.

[2]Numbers extracted from https://huggingface.co/ datasets on 13.08.2023.

knowledge there exists no legal negation corpus.

We annotate four new datasets containing legal judgments from Swiss and German courts in German, French and Italian for negation cues and scopes. We find that these legal documents contain on average longer sentences as well as longer annotated negation scopes, compared to existing datasets. Our experiments show that the legal domain poses a significant challenge to models attempting negation scope resolution. The results achieved by models pre-trained in different domains and evaluated on legal data are lower than those seen in other cross-corpus experiments (Khandelwal and Sawant, 2020; Shaitarova and Rinaldi, 2021). Using our newly annotated datasets, we can improve these results. We conduct experiments where the models are fine-tuned on two languages of the legal data and evaluated on the third. In these zero-shot cross-lingual experiments, our models achieve higher F1-scores than the models pre-trained only on different domains. By training on all available data, we are able to further improve these results, achieving F1-scores around 90% for our multilingual experiments. Our results provide an interesting insight into how even smaller datasets can make a valuable contribution to improving the performance of language models (LMs) on a specific downstream task such as negation scope resolution.

Contributions

The contributions of this paper are three-fold:

- We annotate new datasets of legal documents for negation in German, French, and Italian each containing around 1000 sentences.
- We train and evaluate models on the task of negation scope resolution on the newly annotated datasets to provide a reference point and achieve token-level F1-scores in the mid eighties for cross-lingual zero-shot experiments and up to 91% in multilingual experiments.
- We publicly release the annotation guidelines, the data, the models and the experimentation code as resources and for reproducibility.[3]

[3]The annotation guidelines as well as the code to fine-tune our models can be found on GitHub: https://github.com/RamonaChristen/Multilingual_Negation_Scope_Resolution_on_Legal_Data. Our best model (https://huggingface.co/rcds/neg-xlm-roberta-base) and dataset (https://huggingface.co/rcds/MulitLegalNeg) are published on huggingface.

2 Related Work

Different approaches have been used to address the issue of negation detection and negation scope resolution. Early research focused mainly on rule-based approaches. NegEx, a simple regular expression algorithm developed by Chapman et al. (2001), was successfully able to identify negations in the medical domain. Morante et al. (2008) first took a machine learning approach to negation scope resolution. They used two memory-based classifiers, one to identify the negation cue in a sentence, and one to identify the scope of the negation. On the negation scope resolution task, they achieved an F1-score of 81% on the BioScope corpus (Szarvas et al., 2008). These results were later surpassed by Fancellu et al. (2017), achieving an F1-score of 92% by using neural networks for scope detection. Khandelwal and Sawant (2020) achieved the best results on the BioScope corpus, as well as on two other publicly available negation corpora, the SFU Review Corpus (Konstantinova et al., 2012) and the ConanDoyle-neg corpus (Morante and Blanco, 2012). Their NegBERT model uses Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2019) and applies a transfer learning approach for negation detection and scope resolution.

Only a limited amount of work has been conducted on negation scope resolution across different languages. Fancellu et al. (2018) developed a cross-lingual system, trained on English data and tested on a Chinese corpus. By employing cross-lingual universal dependencies in English they were able to achieve an F1-score of 72% on the Chinese data. Shaitarova et al. (2020) investigated cross-lingual zero-shot negation scope resolution between English, Spanish, and French. They built on NegBERT but used the multilingual BERT (mBERT) model. Shaitarova and Rinaldi (2021) built on this using NegBERT with mBERT and XLM-R_{Large} (Conneau et al., 2020), and were able to achieve a token-level F1-score of 87% on zero-shot transfer from Spanish to Russian.

The sparse amount of cross-lingual research can be explained by the lack of annotated data in languages other than English. There are few corpora annotated with negations in German and Italian (Jiménez-Zafra et al., 2020). The only German corpus annotated for negation and speculation contains medical data and clinical notes (Cotik et al., 2016). However, the corpus is not publicly available and no annotation guidelines have been published. For Italian, Altuna et al. (2017) presented a framework for the annotation of negations and applied it to a corpus of news articles and tweets, parts of which are publicly available. In French, Dalloux et al. (2020) annotated a medical corpus, available on request. To our knowledge, no legal corpus annotated with negations currently exists.

3 Data

3.1 Legal Data

We use court decisions in our legal datasets, also often referred to as judgments. The judgments form German courts were collected from *Bayern.Recht*[4] and include a variety of legal domains and structures (Glaser. et al., 2021). The Swiss court decisions in French, Italian, and German (CH) were collected from the Federal Supreme Court of Switzerland (FSCS). The FSCS is the highest legal authority in Switzerland and oversees federal criminal, administrative, patent, and cantonal courts.

Judgments published by the FSCS usually consist of four sections: 1) The introduction gives information about the date, chamber, involved judge(s) and parties, and the topic of the court decision. 2) The facts outline the important case information. 3) The considerations form the basis for the final ruling by providing relevant case law and other cited rulings. 4) The rulings gives the final decision made by the court.

3.2 Datasets

We annotated four new datasets in three languages for negation cues and scopes, and standardized the existing French and English datasets to make them more accessible. Our datasets consist of publicly available legal judgments from Swiss and German courts. Since negation scope resolution is a sentence-level task, we first split the data into sentences using sentence boundary annotations. The French (fr) and Italian (it) datasets consist of a subset of Swiss court decisions from the Swiss-Judgment-Prediction (SJP) dataset (Niklaus et al., 2022) and the Multi-Legal-Pile (Niklaus et al., 2023b) which were annotated for sentence spans by Brugger et al. (2023). The main German data (de (DE)) is a subset of judgments from German courts collected by Glaser. et al. (2021). Only judgments were included in our dataset because they include a variety of sources and legal areas, they also have a higher density of negation cues compared to other legal texts. To validate that the negation scope prediction also works on German court data from Switzerland, we curated a small dataset of German-Swiss court decisions (de (CH)) that is also a subset of the SJP corpus. We separated each dataset into a train (70%), test (20%), and validation (10%) split.

To ensure that sufficient negation data is available in each dataset, a negation score was assigned to each document based on a simple word search for the most common negation words in each language. The documents with the highest negation scores were then selected to be annotated. Table 1 shows the amount of data and the distribution of negations for the newly created datasets in comparison to the existing datasets in English and French. Our datasets contain a slightly higher ratio of negated sentences compared to the other datasets. This can be attributed to the nature of legal data and our pre-selection procedure. Because we annotated only a subset of an existing corpus we were able to exclude documents without or only few negations while other corpora like ConanDoyle-neg and SFU annotated complete existing datasets or stories.

	Dataset	Total	Negated	%neg
legal	fr	1059	382	36.07
	it	1001	418	41.76
	de (DE)	1098	454	41.35
	de (CH)	208	112	53.85
external	SFU	17672	3528	19.96
	BioScope	14700	2095	14.25
	ConanDoyle-neg	5714	1421	24.87
	Dalloux	11032	1817	16.47

Table 1: Total number of sentences, and number and percentage of sentences containing at least one negation.

Annotations were done by native-language human annotators using the tool Prodigy. All annotators are university students but not part of a legal study program. The annotations were cross-checked by one annotator, who has a linguistic background, with the help of an online translator to ensure that they adhere to the annotation guidelines and are consistent across all three languages. The annotation guidelines are based on existing guidelines for the English datasets, and have been extended to cover all three languages included in our data, as well as the characteristics of the legal domain. Key guidelines are summarized below.

[4] https://www.gesetze-bayern.de/

Negation Cues Cues were not annotated as part of the negation scope following the annotation guidelines for the ConanDoyle-neg corpus (Morante et al., 2011). We excluded affixal cues[5] in our annotations and kept all annotations to the word as the level of the minimal syntactic unit.

Multiple negations Annotators were instructed to annotate one negation per sentence. Sentences with multiple negations were duplicated before annotation based on the most common negation cues. To ensure that the same cue was not annotated twice, duplicates were displayed next to each other in the annotation tool to allow annotators to see which clues had yet to be annotated.

Maxiumum scope strategy As with BioScope, we used a maximum scope strategy. This means that the scope extends to the largest possible unit. If a negated clause has subordinate clauses providing additional information to the clause, the scope extends over the negated clause and all of its subordinate clauses, as illustrated in example 1. This sentence structure is very common in our set of legal data. In all following examples we mark the cue in **bold** and underline the scope. We provide an English translation for clarity.

1 Vorliegend ginge es **nicht** darum, dass ein Arbeitgeber über Fristen oder Pflichten nicht aufgeklärt habe, somit eine blosse Untätigkeit des Arbeitgebers [...]

EN: *In the present case, it was* ***not*** *a matter of an employer not having provided information about deadlines or obligations, thus a mere inactivity on the part of the employer [...].*

Case citations Our dataset contains two main types of citations: inline citations and parenthesized citations. Inline citations, as in example 2 were annotated as part of the scope, while parenthesized citations, as in example 3, were excluded from the negation scope.

2 Da der Kläger **kein** ähnlicher leitender Angestellter i.S.d 14 Abs. 2Satz 2 KSchG ist [...]

EN: *Since the plaintiff is* ***not*** *a similar executive employee in the sense of 14 Abs. 2Satz 2 KSchG* [...]

[5] Affixal cues are cues within a word such as impossible

3 Seit dem 06.02.2017 ist der Kläger im Handelsregister **nicht mehr** als Geschäftsführer eingetragen (vgl. Auszug aus dem Handelsregister in Anlage K9, Bl 75 ff. d.A).

EN: *Since 06.02.2017 the plaintiff is* ***no longer*** *registered in the commercial register as managing director (see extract from the commercial register in annex K9, Bl 75 ff. d.A)*

Punctuation Punctuation marks, such as periods or exclamation points, were excluded from the scope, unless the scope spans multiple clauses separated by commas.

Table 2 shows the average number of tokens in a sentence for all datasets, as well as the average length of the annotated scopes as a ratio between annotated and not annotated tokens. On average, the sentences in our legal datasets are slightly longer than in other datasets. Furthermore, the mean length of the annotated scopes in our data is higher than in all other datasets. For de (DE), more than 50% of tokens were annotated as scope, which is around twice as much as with the biomedical, literary, and review corpora. This is due to the legal domain's sentence structure and our annotation guidelines, which include the subject in the scope. Additionally, nested sentences with multiple subordinate clauses are common in our dataset. This, combined with our maximum scope strategy, leads to longer scopes compared to other datasets.

	Dataset	Sentence	Scope
legal	fr	48.52	37.96%
	it	40.84	30.17%
	de (DE)	31.14	50.18%
	de (CH)	27.65	36.03%
external	SFU	24.46	21.87%
	BioScope	28.49	25.91%
	ConanDoyle-neg	22.11	32.37%
	Dalloux	25.96	19.82%

Table 2: The average number of tokens per sentence. Scopes are shown as a percentage of negated tokens.

4 Experimental setup

We performed experiments to assess negation scope resolution model performance on our multilingual legal data. We integrated the NegBERT architecture (Khandelwal and Sawant, 2020), successful in this task on prior datasets, with various pre-trained multilingual LMs outlined in Table 3. We ran each experiment five times with different random seeds and report the mean token-level F1-score averaged over random seeds, together with the standard deviation. All experiments were conducted with the same hyperparameters for all models, optimized with a search over learning rate (5e-7, 1e-6, 3e-6, 1e-5, 3e-5, 5e-5) and batch size (4, 8, 16, 32, 64, and 128). We optimized the hyperparameters for mBERT and XLM-R and concluded that the best results can be achieved with an initial learning rate of 1e-5 and a batch size of 16. To avoid overfitting, we used early stopping with patience set to 8 as a compromise between the patience of 6 used in the original NegBERT experiments (Khandelwal and Sawant, 2020) and 9 used in the multilingual experiments of Shaitarova and Rinaldi (2021). We extended the maximum input length to 252 tokens for our data. Experiments ran on an NVIDIA A100 GPU via Google Colab, totaling around 105 hours of training time.

Firstly, we evaluated ChatGPT in zero- and few-shot experiments to interpret the results of a non-fine-tuned model in the negation scope resolution task. For all subsequent experiments, we used the NegBERT architecture. In the first NegBERT experiment, models were fine-tuned on all existing French and English datasets and evaluated on our new legal datasets, representing a Zero-shot cross-domain transfer. For a second series of zero-shot experiments, we attempted a Zero-shot cross-lingual transfer within our legal datasets. In each cross-lingual experiment, models were trained on two dataset languages and evaluated on the third. We also executed Multilingual experiments using our datasets and all available data.

5 Results

ChatGPT We evaluated the performance of ChatGPT-3.5 (Brown et al., 2020), one of the leading LMs, in the task of negation scope resolution on our legal datasets. Other researchers have found that ChatGPT performs well on simple annotation tasks such as text classification (Gilardi et al., 2023). To analyze ChatGPT's understanding of negation scopes, we conducted a small test over the chat interface (See Appendix A) which showed that it was able to correctly identify the negation scope of a simple German sentence. For the same request with an example sentence from our legal dataset, ChatGPT was not able to accurately identify the negation scope. To evaluate the performance on the whole dataset, we used the ChatGPT API with 'gpt-3.5-turbo-16k' to accommodate longer inputs. We set the temperature to 0 to reduce randomness and receive a coherent output in json format. Similar to the experiments with the NegBERT architecture, we gave the sentence as well as the negation cues as input and prompted ChatGPT to return the sentence annotated for negation scopes. In a zero-shot experiment, we did not give any annotated examples and only provided a short definition of negation scopes. The results show that ChatGPT's performance on our datasets is subpar (Table 4). In an effort to increase the performance, we conducted some few-shot experiments where 1, 5 or 10 examples of annotated sentences were provided with the prompt, but it did not lead to improvement. The results of the 1-shot experiments averaged lower than the 0-shot experiments. Overall the standard deviation is very high which can be explained by the fact that a random set of annotated examples was selected for each of the five experiment runs. Overall we can conclude, that ChatGPT is currently not suited to solve negation scope resolution in the legal domain without fine-tuning.

Zero-shot cross-domain transfer The results for our zero-shot cross-domain transfer experiments are presented in Table 5. The best results over all datasets were achieved by the Legal-XLM-R_{Large} model, scoring an F1-score of 71.6%. Overall, the LMs pre-trained on legal data demonstrated a 4-percentage point advantage, with a mean F1 of 68.3% averaged over all four legal models, compared to the other models pre-trained on different domains. Furthermore, we notice that the standard deviation for the experiments conducted with the LMs pre-trained on legal data is higher compared to the other models. A possible explanation is that pre-training on legal data improved negation predictions in some areas but adversely affected others, likely due to bias in the legal models, thereby increasing standard deviations across experiments. Generally, cross-domain transfer to the legal domain is less successful than other zero-shot experiments across languages and domains (i.e., Shaitarova et al. (2020); Khandelwal and Sawant (2020)). This suggests that transferring from non-legal to legal domains is challenging.

Zero-shot cross-lingual transfer Table 6 presents the results of our zero-shot cross-lingual

Model	Source	InLen	Params	Vocab	NumTokens	Corpus	Langs
DistilmBERT	Sanh et al. (2020)	512	134M	120K	n/a	Wikipedia	104
mBERT	Devlin et al. (2019)	512	177K	120K	n/a	Wikipedia	104
XLM-$R_{Base/Large}$	Conneau et al. (2020)	512	278M/560M	250K	6'291B	2.5TB CC100	100
Glot500-m	ImaniGooghari et al. (2023)	512	395M	401K	94B	glot500-c	511
Legal-Swiss-$R_{Base/Large}$	Rasiah et al. (2023)	512	184M/435M	128K	262B/131B	CH Rulings/Legislation	3
Legal-XLM-$R_{Base/Large}$	Niklaus et al. (2023b)	512	184M/435M	128K	262B/131B	CH Rulings/Legislation	3

Table 3: Model stats. InLen: max input length during pre-training. Params: total parameter count. NumTokens: Batch size × Steps × InLen

Test Dataset	0-shot	1-shot	5-shot	10-shot	Mean F1 by Dataset
fr	$13.00_{\pm 2.1}$	$16.63_{\pm 10.3}$	$14.90_{\pm 7.5}$	$22.53_{\pm 10.7}$	$16.77_{\pm 8.5}$
it	$25.11_{\pm 1.5}$	$18.22_{\pm 6.5}$	$31.07_{\pm 7.1}$	$26.10_{\pm 3.8}$	$25.12_{\pm 6.7}$
de (DE)	$16.47_{\pm 2.6}$	$22.45_{\pm 9.1}$	$17.34_{\pm 2.7}$	$24.48_{\pm 10.7}$	$20.18_{\pm 7.5}$
de (CH)	$32.91_{\pm 7.9}$	$21.20_{\pm 5.8}$	$36.89_{\pm 18.6}$	$19.83_{\pm 10.3}$	$27.71_{\pm 13.1}$
Mean F1 by experiment	$21.87_{\pm 8.9}$	$19.62_{\pm 7.8}$	$25.05_{\pm 13.6}$	$23.23_{\pm 8.9}$	

Table 4: Results for zero- and few-shot experiments conducted over the ChatGPT API.

experiments conducted with only our legal data. Although these datasets are considerably smaller than the existing English and French datasets, we were able to increase the F1-score by an average of 15.6% across all models and datasets. The legal models still performed well in these experiments, but they no longer showed an advantage over the other LMs. XLM-R_{Base} achieved the best results. All models, except for DistilmBERT, performed significantly better than in the previous experiment across all datasets. DistilmBERT performed worse on the German datasets than in the previous experiment. One explanation for this might be that DistilmBERT is the only cased model used in our experiments. While cased models usually outperform uncased models, this does not seem to apply to cross-lingual experiments. Similar results were found by Macková and Straka (2020), who conducted cross-lingual reading comprehension experiments from English to Czech and found that the uncased models outperformed the cased models in these experiments. They theorized that the overlap of sub-words is larger between English and Czech for uncased models because they disregard diacritical marks, which are common in Czech. A similar argument could be made for the cross-lingual transfer between Italian, French, and German because German includes a lot of casing information while Italian and French do not.

Multilingual experiments The best results for negation scope resolution on our legal datasets were achieved by training our models on the entirety of the available data (Table 7). This multilingual approach achieved an average F1-score of 90% across all models and datasets and outperformed all of the previous setups. This indicates that a relatively small amount of training data in the domain and language of the test dataset can significantly improve the performance of a LM. It is also notable that there seems to be no substantial difference in the performance of the different LMs in this experiment, with a standard deviation of only ± 3.6 over all models and datasets. Although DistilmBERT obtained the lowest scores in this experiment, its performance is not significantly inferior to that of the mBERT model. This could be attributed to the fact that the training data also included German examples which might have mitigated the advantage of the uncased models with regard to shared vocabulary. We also conducted multilingual experiments only using our new datasets which achieved very similar results with an overall F1-score of $89.1_{\pm 4}$ (see Appendix C).

5.1 Error analysis

We investigated the length of the predicted negation scopes as well as random samples of the predictions on the French and German test data to identify some common error cases.

Model \ Test Dataset	fr	it	de (DE)	de (CH)	Mean F1 by Model
DistilmBERT	$61.43_{\pm1.9}$	$63.40_{\pm2.6}$	$63.50_{\pm4.3}$	$58.78_{\pm4.5}$	$61.78_{\pm3.8}$
mBERT	$66.39_{\pm2.1}$	$68.49_{\pm0.8}$	$64.17_{\pm3.1}$	$54.31_{\pm4.8}$	$63.34_{\pm6.2}$
XLM-R$_{\text{Base}}$	$66.80_{\pm1.9}$	$71.40_{\pm0.8}$	$67.29_{\pm3.7}$	$62.44_{\pm2.9}$	$66.98_{\pm4.0}$
XLM-R$_{\text{Large}}$	$72.30_{\pm2.0}$	$70.30_{\pm0.9}$	$73.81_{\pm4.2}$	$\mathbf{63.72}_{\pm4.6}$	$70.03_{\pm5.0}$
Glot500-m	$63.78_{\pm0.8}$	$65.54_{\pm1.1}$	$61.38_{\pm4.0}$	$54.51_{\pm2.5}$	$61.30_{\pm4.9}$
Legal-Swiss-R$_{\text{Base}}$	$69.48_{\pm2.3}$	$68.64_{\pm1.0}$	$71.81_{\pm3.8}$	$54.26_{\pm4.9}$	$66.05_{\pm7.7}$
Legal-Swiss-R$_{\text{Large}}$	$\mathbf{74.66}_{\pm2.4}$	$72.68_{\pm1.5}$	$\mathbf{76.5}_{\pm1.6}$	$51.75_{\pm6.6}$	$68.89_{\pm10.8}$
Legal-XLM-R$_{\text{Base}}$	$71.50_{\pm3.1}$	$71.48_{\pm2.2}$	$71.35_{\pm5.4}$	$51.93_{\pm3.5}$	$66.57_{\pm9.3}$
Legal-XLM-R$_{\text{Large}}$	$74.52_{\pm2.1}$	$\mathbf{74.48}_{\pm3.3}$	$76.06_{\pm3.3}$	$61.30_{\pm8.9}$	$\mathbf{71.59}_{\pm7.7}$
ChatGPT	$13.00_{\pm2.1}$	$25.11_{\pm1.5}$	$16.47_{\pm2.6}$	$32.91_{\pm7.9}$	$21.87_{\pm8.9}$
Mean F1 by Dataset	$68.99_{\pm4.9}$	$\mathbf{69.60}_{\pm3.7}$	$69.54_{\pm6.4}$	$57.00_{\pm6.4}$	$66.28_{\pm7.6}$

Table 5: Cross-domain zero-shot results from existing datasets to our new legal datasets. All models except for ChatGPT were pre-trained on all external datasets, ChatGPT did not receive any training data. The bottom right entry shows the average across all datasets and models except ChatGPT.

Predicted scope length As expected, our cross-domain zero-shot experiments without legal training data achieved the lowest F1-scores overall. This can mostly be attributed to the differences in annotation for each dataset, as well as the different domains. Although the external corpora included French data, this did not improve the performance on the French dataset compared to the other legal datasets. A possible reason is that the subject was not annotated as part of the scope in the Dalloux dataset opposed to the French legal dataset.

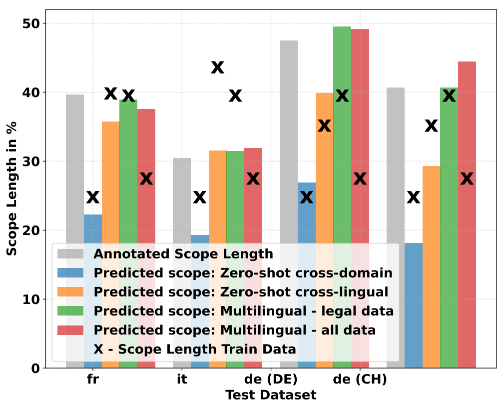


Figure 2: Actual scope length and scope length predicted by Legal-XLM-R$_{\text{Large}}$ for each experiment. X marks the scope length of the train data.

Analyzing the predicted scope length compared to the actual scope length reveals one main issue with the zero-shot transfer from the external datasets of different domains to our legal datasets. Figure 2 shows the analysis of the predicted scopes by the Legal-XLM-R$_{\text{Large}}$ model. In our cross-domain zero-shot experiment, the predicted scope length is significantly shorter than the actual annotated scope length. This is clarified by Table 2, revealing the external datasets have a shorter annotated scope length (24%) compared to our legal datasets (38.6%). Sample predictions confirm that the model often omits the subject from the annotated scope.

Annotation : Es sei festzustellen, dass der Rückerstattungsanspruch **nicht** verjährt sei.

*EN: It should be noted that the claim for restitution is **not** forfeited.*

Prediction : Es sei festzustellen , dass der Rückerstattungsanspruch **nicht** verjährt sei.

*EN: It should be noted that the claim for restitution is **not** forfeited.*

As soon as some legal data is added to our training sets, the predicted scope length as well as the F1-score increases. An inspection of the predictions made by the legal and multilingual models shows that the additional training data helps to predict the subject as part of the scope. One exception where the subject was not annotated in the prediction is for subjects represented by an initial instead of a pronoun or a full name, which is common in

Model \ Test Dataset	fr	it	de (DE)	de (CH)	Mean F1 by Model
DistilmBERT	$79.56_{\pm 1.0}$	$74.94_{\pm 1.7}$	$58.74_{\pm 9.6}$	$52.59_{\pm 11.3}$	$66.46_{\pm 13.3}$
mBERT	$87.22_{\pm 1.6}$	$81.94_{\pm 1.3}$	$81.39_{\pm 3.6}$	$70.78_{\pm 6.7}$	$80.33_{\pm 7.1}$
XLM-R$_{\text{Base}}$	$88.70_{\pm 0.8}$	$\mathbf{86.43}_{\pm 2.2}$	$88.00_{\pm 1.9}$	$\mathbf{83.71}_{\pm 4.8}$	$\mathbf{86.71}_{\pm 3.3}$
XLM-R$_{\text{Large}}$	$\mathbf{90.55}_{\pm 0.9}$	$84.93_{\pm 1.7}$	$\mathbf{91.36}_{\pm 0.8}$	$76.65_{\pm 4.5}$	$85.87_{\pm 6.4}$
Glot500-m	$86.77_{\pm 2.3}$	$83.41_{\pm 1.3}$	$90.10_{\pm 2.0}$	$77.73_{\pm 4.6}$	$84.50_{\pm 5.4}$
Legal-Swiss-R$_{\text{Base}}$	$87.42_{\pm 1.2}$	$84.54_{\pm 1.6}$	$88.24_{\pm 1.0}$	$70.95_{\pm 3.6}$	$82.79_{\pm 7.4}$
Legal-Swiss-R$_{\text{Large}}$	$84.63_{\pm 1.0}$	$83.88_{\pm 1.9}$	$88.47_{\pm 3.9}$	$70.33_{\pm 6.0}$	$81.83_{\pm 7.8}$
Legal-XLM-R$_{\text{Base}}$	$86.40_{\pm 2.1}$	$83.28_{\pm 1.4}$	$89.56_{\pm 2.5}$	$74.52_{\pm 8.0}$	$83.44_{\pm 7.0}$
Legal-XLM-R$_{\text{Large}}$	$85.51_{\pm 1.7}$	$85.76_{\pm 0.3}$	$89.58_{\pm 1.8}$	$80.16_{\pm 4.0}$	$85.25_{\pm 4.1}$
Mean F1 by dataset	$\mathbf{86.31}_{\pm 3.2}$	$83.23_{\pm 3.5}$	$85.05_{\pm 10.4}$	$73.05_{\pm 10.3}$	$81.91_{\pm 9.3}$

Table 6: Multilingual zero-shot experiments within our legal datasets. Each column represents a different set of test and train data where the test data includes all legal datasets in languages that are not the language of the test dataset i.e. models evaluated on fr were trained with it and de (DE,CH).

legal documents for anonymization reasons. We suspect that in these cases the models were not able to identify the initial as the subject because these kinds of subjects might be more uncommon outside of the legal domain.

Annotation: E._ **ne** disposait d'**aucune** autonomie budgétaire;

EN: E._ had ***no*** *budgetary autonomy*

Prediction: E. _ **ne** disposait d'**aucune** autonomie budgétaire;

EN: E._ had ***no*** *budgetary autonomy*

Non-continuous scopes Another error case is sentences where the scope is not continuous because it is interrupted by an interjection or contrasting statement. These kinds of sentences are more complex than the average sentence and not very common in the training data. A larger amount of training data containing similar sentence structures could improve accuracy.

Annotation: Eine ordentliche Kündigung ist während der vereinbarten Laufzeit beiderseits nur zum Vertragsende und **nicht** zu einem früheren Zeitpunkt zulässig.

EN An ordinary termination during the agreed term is only permissible on both sides at the end of the contract and ***not*** *at an earlier time*

Prediction: Eine ordentliche Kündigung ist während der vereinbarten Laufzeit beiderseits nur zum Vertragsende und **nicht** zu einem früheren Zeitpunkt zulässig.

EN An ordinary termination during the agreed term is only permissible on both sides at the end of the contract and ***not*** *at an earlier time*

6 Conclusions and Future Work

6.1 Conclusion

We released new legal datasets in German, French and Italian, annotated for negation cues and scopes and showed that the legal domain does pose a challenge for models in negation scope resolution. Cross-domain zero-shot experiments showed that models without legal training data do not perform as well on multilingual legal datasets as they do on other domains. The task is also too complex for ChatGPT, which was not able to reach F1-scores above 37%. Using our new datasets we fine-tuned different models on the legal domain, significantly improving the results and showing that even relatively small amounts of training data in a specific domain and language can improve the performance of multilingual LMs for negation scope resolution.

6.2 Future Work

Negation scope resolution models in the legal domain could benefit from more training data to increase the accuracy of predictions of more complex sentence structures such as non-continuous scopes. More diverse data from different legal fields could further improve the performance of negation scope models in the legal domain.

With our new datasets we were able to show that existing systems performing well on datasets

Model \ Test Dataset	fr	it	de (DE)	de (CH)	Mean F1 by Model
DistilmBERT	$87.54_{\pm 0.6}$	$82.90_{\pm 1.3}$	$94.63_{\pm 0.5}$	$90.77_{\pm 1.2}$	$88.96_{\pm 4.5}$
mBERT	$89.98_{\pm 2.1}$	$83.72_{\pm 1.0}$	$95.21_{\pm 0.5}$	$87.83_{\pm 1.0}$	$89.10_{\pm 4.4}$
XLM-R$_{\text{Base}}$	$\mathbf{91.31}_{\pm 1.2}$	$88.81_{\pm 1.1}$	$94.74_{\pm 0.7}$	$89.39_{\pm 1.8}$	$\mathbf{91.06}_{\pm 2.6}$
XLM-R$_{\text{Large}}$	$90.77_{\pm 1.8}$	$87.44_{\pm 0.5}$	$93.40_{\pm 1.1}$	$90.20_{\pm 3.9}$	$90.45_{\pm 3.0}$
Glot500-m	$89.65_{\pm 1.0}$	$85.54_{\pm 2.3}$	$94.94_{\pm 0.7}$	$\mathbf{91.00}_{\pm 2.7}$	$90.28_{\pm 3.8}$
Legal-Swiss-R$_{\text{Base}}$	$89.08_{\pm 1.6}$	$87.40_{\pm 1.9}$	$94.60_{\pm 1.0}$	$87.02_{\pm 1.5}$	$89.52_{\pm 3.4}$
Legal-Swiss-R$_{\text{Large}}$	$89.07_{\pm 1.4}$	$86.72_{\pm 1.5}$	$\mathbf{95.94}_{\pm 0.2}$	$89.39_{\pm 0.9}$	$90.28_{\pm 3.7}$
Legal-XLM-R$_{\text{Base}}$	$90.71_{\pm 0.5}$	$86.67_{\pm 0.5}$	$95.41_{\pm 0.7}$	$86.17_{\pm 2.4}$	$89.74_{\pm 4.0}$
Legal-XLM-R$_{\text{Large}}$	$90.75_{\pm 1.4}$	$\mathbf{89.46}_{\pm 0.8}$	$93.87_{\pm 0.8}$	$89.18_{\pm 1.0}$	$90.82_{\pm 2.1}$
Mean F1 by Dataset	$89.87_{\pm 1.2}$	$86.52_{\pm 2.4}$	$\mathbf{94.74}_{\pm 1.0}$	$88.99_{\pm 2.4}$	$90.03_{\pm 3.6}$

Table 7: Results from multilingual experiments over all available data.

across different domains are not necessarily able to perform as well on legal data. This should motivate future work to focus on this complex domain and evaluate the performance of existing systems in diverse NLP tasks.

Limitations

Due to resource constraints, our datasets are relatively small compared to other publicly available corpora. A larger set of legal data accross a diverse set of sources, annotated with negations could further improve the performance of LMs for negation scope resolution in this field. We also did not investigate the potential of cross-lingual cue detection since this is the more trivial part of negation research and can easily be replaced by a list of negation cues for each language.

Ethics Statement

The goal of our work was to improve the performance of negation scope resolution systems in the legal domain. These improved systems could be used to support legal professionals in processing and analysing legal texts. These systems should only be used as an assistance to human experts with considerations to their limitations and possible biases. To the best of our knowledge there is currently no real world application of a negation scope resolution system in the legal domain.

The legal data that we annotated and used to train our models is all publicly available and has all been anonymized. It should therefore not include any sensitive information.

References

Begoña Altuna, Anne-Lyse Minard, and Manuela Speranza. 2017. The scope and focus of negation: A complete annotation framework for Italian. In *Proceedings of the Workshop Computational Semantics Beyond Events and Roles*, pages 34–42, Valencia, Spain. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tobias Brugger, Matthias Stürmer, and Joel Niklaus. 2023. MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset. ArXiv:2305.01211 [cs].

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6974–6996, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Wendy W. Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv:1911.02116 [cs]*. ArXiv: 1911.02116.

Viviana Cotik, Roland Roller, Feiyu Xu, Hans Uszkoreit, Klemens Budde, and Danilo Schmidt. 2016. Negation detection in clinical reports written in German. In *Proceedings of the Fifth Workshop on Building and Evaluating Resources for Biomedical Text Mining BioTxtM2016)*, pages 115–124, Osaka, Japan. The COLING 2016 Organizing Committee.

Clément Dalloux, Vincent Claveau, Natalia Grabar, Lucas Oliveira, Claudia Moro, Yohan Gumiel, and Deborah Carvalho. 2020. Supervised learning for the detection of negation and of its scope in french and brazilian portuguese biomedical corpora. *Natural Language Engineering*, 27:1–21.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv:1810.04805.

Federico Fancellu, Adam Lopez, and Bonnie Webber. 2018. Neural networks for cross-lingual negation scope detection.

Federico Fancellu, Adam Lopez, Bonnie Webber, and Hangfeng He. 2017. Detecting negation scope is easy, except when it isn't. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 58–63, Valencia, Spain. Association for Computational Linguistics.

Roland Friedrich. 2021. Complexity and entropy in legal language. *Frontiers in Physics*, 9:671882.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Ingo Glaser., Sebastian Moser., and Florian Matthes. 2021. Sentence boundary detection in german legal documents. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART*, pages 812–821. INSTICC, SciTePress.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Salud María Jiménez-Zafra, Roser Morante, María Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2020. Corpora annotated with negation: An overview. *Computational Linguistics*, 46(1):1–52.

Aditya Khandelwal and Suraj Sawant. 2020. NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution. *arXiv:1911.04211 [cs]*. ArXiv: 1911.04211.

Natalia Konstantinova, Sheila CM De Sousa, Noa P Cruz Díaz, Manuel J Mana López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3190–3195.

Kateřina Macková and Milan Straka. 2020. Reading comprehension in czech via machine translation and cross-lingual transfer. In *Text, Speech, and Dialogue*, pages 171–179, Cham. Springer International Publishing.

Sabrina J. Mielke. 2016. Language diversity in ACL 2004 - 2016.

Roser Morante and Eduardo Blanco. 2012. * sem 2012 shared task: Resolving the scope and focus of negation. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Roser Morante, Anthony Liekens, and Walter Daelemans. 2008. Learning the scope of negation in biomedical texts. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 715–724, Honolulu, Hawaii. Association for Computational Linguistics.

Roser Morante, Sarah Schrauwen, and Walter Daelemans. 2011. Annotation of negation cues and their scope: Guidelines v1. *Computational linguistics and psycholinguistics technical report series, CTRS-003*, pages 1–42.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023a. LEXTREME: A Multi-Lingual and Multi-Task Benchmark for the Legal Domain. ArXiv:2301.13126 [cs].

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E. Ho. 2023b. MultiLegalPile: A 689GB Multilingual Legal Corpus. ArXiv:2306.02069 [cs].

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An empirical study on cross-x transfer for legal judgment prediction. ArXiv:2209.12325.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E. Ho, and Joel Niklaus. 2023. SCALE: Scaling up the Complexity for Advanced Language Model Evaluation. ArXiv:2306.09237 [cs].

JB Ruhl, Daniel Martin Katz, and Michael J Bommarito. 2017. Harnessing legal complexity. *Science*, 355(6332):1377–1378.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. ArXiv: 1910.01108.

Anastassia Shaitarova, Lenz Furrer, and Fabio Rinaldi. 2020. Cross-lingual transfer-learning approach to negation scope resolution. In *CEUR Workshop proceedings*. CEUR-WS.

Anastassia Shaitarova and Fabio Rinaldi. 2021. Negation typology and general representation models for cross-lingual zero-shot negation scope resolution in Russian, French, and Spanish. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 15–23, Online. Association for Computational Linguistics.

György Szarvas, Veronika Vincze, Richárd Farkas, and János Csirik. 2008. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the workshop on current trends in biomedical natural language processing*, pages 38–45. Association for Computational Linguistics.

A ChatGPT negation scope example

User: In the following German sentence give me a list of all words that are part of a negation scope. Negation scope describes all words affected by a negation in a sentence. Format the output as json. Sentence: Ich weiss nicht was eine Kartoffel ist.

ChatGPT: Certainly, here's the JSON output for the words that are part of the negation scope in the provided German sentence:

```
{
"Neg_scope": [
    "weiss",
    "was",
    "eine",
    "Kartoffel",
    "ist"
}
```

Figure 3: ChatGPT annotation of a simple sentence.

User: In the following German sentence give me a list of all words that are part of a negation scope. Negation scope describes all words affected by a negation in a sentence. Format the output as json. Sentence: Die Gewährung der tariflichen Leistungen, somit auch die Versorgungszusage, die laut Tarifvertrag für Auszubildende nicht gelte, hätte überhaupt nur dann Sinn gemacht, wenn die Versorgungszusage entgegen des Geltungsbereichs des Tarifvertrages auch dem Auszubildenden erteilt hätte werden sollen.

ChatGPT: Certainly, here's the JSON output for the words that are part of the negation scope in the provided German sentence:

```
{
  "Neg_scope": [
    "gelte"
  ]
}
```

Figure 4: ChatGPT annotation of a more complex legal sentence.

B Mean results over all models, datasets and experiments

Model \ Test Dataset	Zero-shot cross-domain	Zero-shot cross-lingual	Multilingual - legal data	Multilingual - all data	Mean F1 by model
DistilmBERT	$61.78_{\pm 3.77}$	$66.46_{\pm 13.33}$	$0.87_{\pm 0.05}$	$88.96_{\pm 4.51}$	$72.40_{\pm 14.54}$
mBERT	$63.34_{\pm 6.24}$	$80.33_{\pm 7.10}$	$0.89_{\pm 0.04}$	$89.19_{\pm 4.41}$	$77.62_{\pm 12.33}$
XLM-R_{Base}	$66.98_{\pm 4.02}$	$86.71_{\pm 3.25}$	$0.90_{\pm 0.03}$	$91.06_{\pm 2.63}$	$81.59_{\pm 11.07}$
XLM-R_{Large}	$70.03_{\pm 4.98}$	$85.87_{\pm 6.44}$	$0.90_{\pm 0.04}$	$90.45_{\pm 3.00}$	$82.12_{\pm 10.10}$
Glot500-m	$61.30_{\pm 4.85}$	$84.50_{\pm 5.36}$	$0.89_{\pm 0.04}$	$90.28_{\pm 3.84}$	$78.70_{\pm 13.46}$
Legal-Swiss-R_{Base}	$66.05_{\pm 7.72}$	$82.79_{\pm 7.41}$	$0.88_{\pm 0.05}$	$89.52_{\pm 3.41}$	$79.45_{\pm 11.82}$
Legal-Swiss-R_{Large}	$68.89_{\pm 10.80}$	$81.83_{\pm 7.81}$	$0.90_{\pm 0.03}$	$90.28_{\pm 3.66}$	$80.33_{\pm 11.84}$
Legal-XLM-R_{Base}	$66.57_{\pm 9.33}$	$83.44_{\pm 7.01}$	$0.90_{\pm 0.04}$	$89.74_{\pm 3.99}$	$79.92_{\pm 12.10}$
Legal-XLM-R_{Large}	$71.59_{\pm 7.73}$	$85.25_{\pm 4.07}$	$0.89_{\pm 0.04}$	$90.82_{\pm 2.12}$	$82.55_{\pm 9.61}$
Mean F1 by experiment	$66.28_{\pm 7.64}$	$81.91_{\pm 9.26}$	$0.89_{\pm 0.04}$	$90.03_{\pm 3.57}$	

Table 8: Mean Results over all models and experiments

C Multilingual results legal data

Model \ Test Dataset	fr	it	de (DE)	de (CH)	Mean F1 by Model
DistilmBERT	$86.06_{\pm 0.76}$	$81.82_{\pm 0.79}$	$93.98_{\pm 0.82}$	$87.40_{\pm 2.36}$	$87.32_{\pm 4.65}$
mBERT	$90.16_{\pm 1.33}$	$84.56_{\pm 1.63}$	$94.95_{\pm 0.80}$	$86.81_{\pm 2.06}$	$89.12_{\pm 4.25}$
XLM-R-Base	$90.26_{\pm 0.96}$	$88.05_{\pm 1.81}$	$94.12_{\pm 0.59}$	$87.21_{\pm 2.66}$	$89.91_{\pm 3.16}$
XLM-R-Large	$90.23_{\pm 1.40}$	$86.93_{\pm 0.73}$	$94.56_{\pm 0.85}$	$86.44_{\pm 3.56}$	$89.54_{\pm 3.80}$
Glot500-m	$88.81_{\pm 1.47}$	$85.62_{\pm 1.12}$	$94.23_{\pm 1.40}$	$88.13_{\pm 2.60}$	$89.20_{\pm 3.59}$
Legal-Swiss-R-Base	$87.98_{\pm 1.46}$	$89.53_{\pm 0.54}$	$93.15_{\pm 0.44}$	$81.82_{\pm 3.88}$	$88.12_{\pm 4.62}$
Legal-Swiss-R-Large	$88.35_{\pm 0.88}$	$88.20_{\pm 1.13}$	$95.30_{\pm 0.37}$	$89.39_{\pm 1.37}$	$90.31_{\pm 3.13}$
Legal-XLM-R-Base	$88.89_{\pm 1.58}$	$88.41_{\pm 1.84}$	$95.56_{\pm 0.88}$	$85.27_{\pm 3.83}$	$89.53_{\pm 4.39}$
Legal-XLM-R-Large	$88.86_{\pm 0.95}$	$87.98_{\pm 0.64}$	$94.46_{\pm 0.69}$	$85.30_{\pm 3.12}$	$89.15_{\pm 3.76}$
Mean F1 by dataset	$88.84_{\pm 1.70}$	$86.79_{\pm 2.55}$	$94.48_{\pm 1.01}$	$86.42_{\pm 3.36}$	$89.13_{\pm 3.97}$

Table 9: Results of multilingual experiments using only our legal datasets.

Anonymity at Risk?
Assessing Re-Identification Capabilities of Large Language Models

Alex Nyffenegger[1,2*] **Matthias Stürmer** [2,3] **Joel Niklaus** [2,3,4*]
[1]University of Fribourg
[2]University of Bern
[3]Bern University of Applied Sciences
[4]Stanford University

Abstract

Anonymity of both natural and legal persons in court rulings is a critical aspect of privacy protection in the European Union and Switzerland. With the advent of LLMs, concerns about large-scale re-identification of anonymized persons are growing. In accordance with the Federal Supreme Court of Switzerland, we explore the potential of LLMs to re-identify individuals in court rulings by constructing a proof-of-concept using actual legal data from the Swiss federal supreme court. Following the initial experiment, we constructed an anonymized Wikipedia dataset as a more rigorous testing ground to further investigate the findings. With the introduction and application of the new task of re-identifying people in texts, we also introduce new metrics to measure performance. We systematically analyze the factors that influence successful re-identifications, identifying model size, input length, and instruction tuning among the most critical determinants. Despite high re-identification rates on Wikipedia, even the best LLMs struggled with court decisions. The complexity is attributed to the lack of test datasets, the necessity for substantial training resources, and data sparsity in the information used for re-identification. In conclusion, this study demonstrates that re-identification using LLMs may not be feasible for now, but as the proof-of-concept on Wikipedia showed, it might become possible in the future. We hope that our system can help enhance the confidence in the security of anonymized decisions, thus leading to the courts being more confident to publish decisions.

1 Introduction

The swift advancements in Natural Language Processing (NLP) (Vaswani et al. 2017; Brown et al. 2020; Ouyang et al. 2022; Khurana et al. 2023) have introduced new challenges to the security of traditional legal processes (Tsarapatsanis and Aletras 2021). As public access to data increases in tandem with digital advancements (Katz et al. 2023; EUGH 2018; Lorenz 2017), the potential risks associated with data disclosure have become increasingly significant. Increasingly larger and more capable Large Language Models (LLMs), more powerful vector stores and potent embeddings together have the capacity to extract unintended information from public data (Borgeaud et al. 2022; Carlini et al. 2021). This poses a security risk, as the identification of individuals involved in legal proceedings can lead to privacy breaches, providing undue advantage to certain legal actors, and risking public defamation. Over the past decade, at least 18 requests for name changes following the re-identification of convicts have been registered in Switzerland, indicating that this issue already exists due to imprudent media coverage (Stückelberger, Evin, and Damian 2021). The number of cases where the accused become victims of unlawful personal information disclosure is likely to rise as further re-identifications occur. The prevention of re-identification is critical not only for the protection of the accused, but also for the courts. Munz (2022) even suggests that the state could be held accountable for non-monetary damages to judged persons, underscoring the urgent need for courts to address the re-identification issue proactively. Vokinger and Mühlematter (2019) have shown that some re-identifications are possible by applying regular expressions.

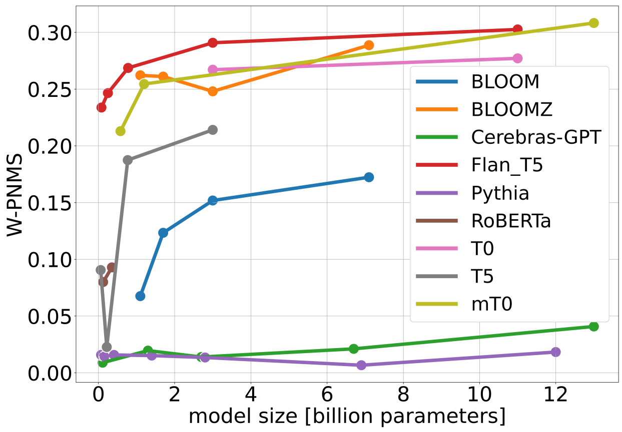


Figure 1: Re-identification rate by parameter count

We use state-of-the-art transformer models (Vaswani et al. 2017) like LLaMA-2, GPT-4 or BLOOM (Touvron et al. 2023; OpenAI 2023; Scao et al. 2023) to re-identify individuals in publicly released Swiss court decisions. Such models have the ability to store information within their parameters and extract this information when prompted (Roberts, Raffel, and Shazeer 2020; AlKhamissi et al. 2022; Ippolito et al. 2023; Carlini et al. 2023). We find that while the best models are capable of identifying persons from masked Wikipedia

*These authors contributed equally.

articles, in the much more difficult case of re-identification from court decisions, they mostly fail. Only using a highly curated set of manually identified relevant news articles, they are capable of identifying the anonymized defendants from cases. Additionally, we identify three main factors influencing the re-identification risk: input length, model size, and instruction tuning.

To both ensure responsible research and maximize downstream usability, we closely collaborated with the Federal Supreme Court of Switzerland (FSCS). The FSCS currently uses regular expressions and a BERT-based (Devlin et al. 2018) token classifier to provide suggestions to human anonymizers for what entities should be masked. Together with the FSCS we improved its recall on anonymization tokens from 83% to 93% by pre-training a legal specific model. In accordance with their anonymization team, in this work we apply what could be called penetration testing to their method of anonymization by developing a tool that could ensure that the applied anonymization is sufficient for safe publication even with stronger LLMs emerging.

A tool with these capabilities can help identify whether affected parties in rulings could still be identified despite anonymization efforts, thus the results from our research can guide legal entities, data privacy advocates, and NLP practitioners in devising strategies to mitigate potential re-identification risks. This is relevant beyond Switzerland, as anonymization of court rulings became mandatory across the EU with the introduction of the DSGVO (See Appendix E.3). The German Supreme Court even ruled that all rulings should be anonymized and published. However, in 2021 barely one percent of rulings were being published (Hamann 2021) (See Appendix E.3). This may be partially caused by fears that publications are insufficiently anonymized and courts could be held accountable. A tool to ensure privacy for anonymized documents could lead to more publications in Germany as well as in the EU.

Main Research Questions

This study is guided by the following key research questions:
RQ1: Performance of LLMs on re-identifications: How effectively can various LLMs re-identify masked persons within Wikipedia pages and in Swiss court rulings?
RQ2: Influential Factors: What are the key factors that influence the performance of LLMs in re-identification tasks?
RQ3: Privacy Implications: How will evolving LLM capabilities and their use in re-identifications affect the preservation of privacy in anonymized court rulings in Switzerland?

By addressing these questions, we aim to highlight LLMs' capabilities and limitations in re-identification tasks and enhance understanding of required privacy considerations in the ongoing digital transformation of legal practice.

Contributions

The contributions of this paper are threefold:

1. We curate and publish a unique, large-scale Wikipedia dataset with masked entities.
2. We introduce new metrics to evaluate performance of re-identifications of entities within texts. Using those metrics, we provide a thorough evaluation and benchmark of various state-of-the-art LLMs in the context of re-identifying masked entities within Wikipedia entries and Swiss court rulings. This includes an exploration of the most critical factors that can influence a model's performance. The results demonstrate that some models are more effective than others for re-identification tasks.
3. We underscore and investigate the potential privacy implications of using LLMs for re-identification tasks.

2 Related Work

Chen et al. (2017) used Language Models (LMs) for machine reading to answer open domain questions by giving models the required context within Wikipedia articles so they would be able to extract the required knowledge. With the advent of the transformer (Vaswani et al. 2017), more powerful models became able to store information within their parameters (Petroni et al. 2019; AlKhamissi et al. 2022) and the idea of using models directly without additional context became viable. Petroni et al. (2019) found that language models can be used as knowledge bases, drawing information from their training set to answer open domain questions. Roberts, Raffel, and Shazeer (2020) went a step further and evaluated LLMs in different sizes, namely T5 (Raffel et al. 2020) showing that larger models can store more information, but unlike other Question Answering (QA) systems are not able to show where facts come from. This is especially a problem when models hallucinate an answer when they are unsure, as correctness of a answer is hard to factually check without any source (Petroni et al. 2019). With Lewis, Stenetorp, and Riedel (2020) finding that good results on open domain question answering heavily depends on the overlap of questions and training data, Wang, Liu, and Zhang (2021) showed that even without overlapping data, knowledge retrieval is possible, although with much lower performance. Finding that knowledge might be present in the models parameters but not retrieved correctly, Wang, Liu, and Zhang (2021) applied a new method, named QA-bridge-tune, to allow the model to more reliably retrieve the relevant information from its parameters. To improve reliability of results even further (Lewis et al. 2021) introduced the combination of pretrained models and a dense vector index of Wikipedia, finding that QA tasks are answered with more specific and factual knowledge than parametric models alone, while hallucinations are reduced when using Retrieval Augmented Generation (RAG) (Shuster et al. 2021). While previous works concentrated on the English language, more recent research (Kassner, Dufter, and Schütze 2021) found that multilingual models might perform better on knowledge retrieval tasks, while the retrieval works much better when the question is asked in the same language as the training information was ingested. Inter-language information retrieval does not perform well, meaning the performance for questions in a language other than the language of the data source is worse than when the question is posed in the data source language (Jiang et al. 2020). Poerner, Waltinger, and Schütze (2020) showed that while pretrained models without specific knowledge retention targets might be able to answer some questions, training on data specifically prepared for a certain knowledge retrieval task can produce much better results without altering the models architecture. In the domain of re-identifications in court rulings, Vokinger and Mühlematter (2019) used linkage methods to connect medical keywords from public information to medical keywords in court rulings, through which they were able to identify persons via their connection to medical terms, specifically drugs and medicine. Their successful attempt to partially re-identify entities within rulings implies the possibility for language models to do the same.

3 Datasets

3.1 Court Decisions Dataset

We used the Swiss caselaw corpus by Rasiah et al. (2023) to benchmark re-identification on court rulings. The FSCS likely rules the most publicised cases as the final body of appeal in Switzerland and offered to validate re-identifications in a limited fashion, leading us to discard cases from other courts. This decision aligned well with the fact that federal court cases occur more often in the news, elevating the likelihood of potential re-identifications. To make sure that all evaluated models have been trained on relevant data, we only used cases from 2019, resulting in approx. 8K rulings.

3.2 Hand Picked Rulings Dataset

Constructing a representative dataset linking news articles with corresponding court rulings would demand extensive data and computational resources. To address this, we crafted a smaller dataset by manually connecting court rulings with pertinent news articles. By probing our complete news dataset using keywords (for file numbers, "judgment", etc.), we pinpointed articles that referenced the file number of a related ruling. While these often safeguarded individuals' identities, other cues or associated stories sometimes hinted at articles naming the individuals. Leveraging the expertise of law students, we received insights on notable court case individuals spotlighted in the news and became familiar with court-specific terminology. This collaboration helped us detect more rulings, resulting in a set of seven cases distinctly cited in news articles, albeit references were fragmented across various articles. To gather information on each entity, we filtered news articles using keywords, like the entity's name or ruling's file number, amassing about 700 relevant articles. These articles varied in content, with some mentioning the file number and others naming unrelated individuals with similar names. To diversify the dataset and ensure models would discern accurate information, we blended these 700 articles with 1K random news articles spanning the same date range. To maintain privacy, the connected news articles and rulings are not disclosed. The news articles are proprietary and were sourced from swissdox.ch.

3.3 Wikipedia Dataset

Data Acquisition We extracted a random subset of 0.6M entries from the Hugging Face Wikipedia dataset (20220301.en) based on individuals identified through the Wikipedia query interface, without specific sorting. Given the large size of the Wikipedia corpus, we favored entries with more extended text — arguably featuring more notable individuals. Prioritizing entries over 4K characters for higher entity prevalence within texts, we excluded bibliography and references, leaving around 71K entries.

Paraphrasing Wikipedia Pages To evaluate how much the models rely on the exact phrasing of text in the training data (Carlini et al. 2021), the Wikipedia pages were paraphrased and stored alongside the original contents. We paraphrased the pages on a sentence-by-sentence basis using PEGASUS fine-tuned for paraphrasing (Zhang et al. 2019)[1]. This approach ensured that the text varied slightly, yet retained the overall structure and essential details.

Masking To prepare the dataset for model prediction, we replaced all occurrences of the individual associated with a entry by a mask token using BERT, fine-tuned for Named Entity Recognition (NER) (Devlin et al. 2018; Lim 2021). The identified entities were concatenated into a single string and matched against the title of the Wikipedia entry using a regular expression. Matches were replaced with the mask token. This process occasionally led to erroneous matches, usually involving family members with similar names. For instance, 'Gertrude Scharff Goldhaber' might mask 'Maurice Goldhaber' (husband) as well. This issue is, as discussed in Section 4, unlikely to have a significant impact on performance due to its rarity relative to the vast number of examples. Unmatched entries, from NER limitations, misaligned names, or mask removal during paraphrasing, were discarded, leaving about 69K entries. A random 10K subset was chosen to better mirror the diverse court rulings dataset. This choice, motivated by performance, likely wouldn't impact results even with a larger corpus.

4 Metrics

Re-identification of persons is a known problem for imaging (Karanam et al. 2018), but comparable metrics for re-identifications within texts are, to the best of our knowledge, not established. To allow the quantification of produced results, we introduce the following four novel metrics to measure re-identification performance of a person in a text:

Partial Name Match Score (PNMS) evaluates predictions against a regular expression requiring any part of an entity's name to be a match for the prediction to be considered as correct. For example, "Max Orwell" would match "George Orwell". This allows for matches with predictions that only contain a part of the name. Manual experimentation suggested that persons can be re-identified by using just a part of their name. The predicted name might be near exact, hence the allowance for partial matches. The metric accepts n predictions and deems any collection of predictions correct if at least one of the n predictions is correct.

Normalized Levenshtein Distance (NLD) is introduced to assess the precision of predictions deemed correct by PNMS. Given that there is no clear-cut distinction between

[1]When the dataset was created, GPT-3.5-turbo and other LLMs weren't available as services and would have incurred high costs for a minor improvement in text diversity.

correct and incorrect, using the Levenshtein distance provides a more nuanced perspective on how close the predictions are to the target. For the top five predictions, the smallest distance of all five was used. Using the best distance of n given predictions, the distance was normalized against the length of the target name to avoid distortions in results. As example, the distance between "Alice Cooper" and "Alina Cooper" would be two, and with the normalization by $len("AlinaCooper")$ applied result in 0.16.

Last Name Match Score (LNMS) works the same way as PNMS, but only the last name is considered. The last name is defined as the last whitespace-separated part of a full name string. Partial matches are accounted as correct as well meaning that the name "Mill" would also be counted as correct if the target was "Miller". This overlap might cause a very slight imprecision but does not lead to problems in evaluations as all models have the same advantage.

Weighted Partial Name Match Score (W-PNMS) blends PNMS and the LNMS using a weighted sum, emphasizing the significance of last names for re-identification. Let $\alpha = 0.35$ be the weight for PNMS. Thus, W-PNMS is calculated as W-PNMS $= \alpha \times$ PNMS $+ (1 - \alpha) \times$ LNMS.

5 Experimental Setup

Models were run using the HuggingFace Transformers library on two 80GB NVIDIA A100 GPUs, using default model configurations in 8-bit precision. For efficiency, only the first 1k characters of each Wikipedia page were used to compute five predictions per example. For the court rulings we employed the same procedure but extended the input length to 10K characters, fully utilizing the available sequence lengths of models evaluated, automatically truncating sequences exceeding the maximum input length.

5.1 Prompt Engineering

The effectiveness of model responses is significantly influenced by the design of input prompts (Liu et al. 2022; Wei et al. 2023). Various models require distinct prompting strategies to perform optimally. In this study, we tailored prompts for each model, but without extensive optimization, ensuring a consistent effort across all models. Experimental results indicated that once a prompt successfully communicated the re-identification task to a model, further refinement of the prompt did not substantially improve any metrics.

5.2 Retrieval Augmented Generation

To estimate how well an LLM could use information from news articles without training one we used RAG (Lewis et al. 2021): From the 1.7K news articles gathered for the hand-picked decision dataset, we split texts into 1K-character chunks, embedded them with OpenAI's text-embedding-ada-002, and stored the embeddings in a Chroma vector database (https://www.trychroma.com/). To re-identify a ruling, we fed it to GPT-3.5-turbo-16k, prompting it to summarize the decision, emphasizing facts in news articles and retaining key details, including masked entities.

We then embedded this shorter version the same way as the articles and matched against the stored article chunks using the similarity search function provided by the Chroma database. The top five retrieved documents together with the shortened version of the ruling were given to GPT-4 with the prompt to use the information given in the documents to re-identify the entity referred to as <mask> in the given decision. This method skips the large training effort required to store knowledge in LLMs while still demonstrating the capability of LLMs to comprehend multi-hop information from news articles and apply it to a re-identification task.

5.3 Evaluated Models

For the rulings dataset, we utilized models that were specifically trained on news articles and court rulings, alongside the two multilingual models, GPT-4 and mT0. The selection of these models, as detailed in Table 3, was informed by their pre-training on relevant news content. For the Wikipedia dataset a plethora of different models with different pre-training datasets and architectures were used. By using a large and diverse selection of models, prominent factors for good performance can be found more easily and results are more reliable. A full list is available in Table 3. All models except the commercial models ChatGPT and GPT-4 are publicly available on the HuggingFace Hub.

5.4 Baselines

We introduce two baselines for easier interpretation:

Random Name Guessing Baseline predicts for every example five first and last names paired up to full names at random. This gives a good impression on predictive performance when models understand the task or at least guess while not actually knowing the entities name. Names were chosen from a GPT-3.5-generated list of 50 names.

Majority Name Guessing Baseline predicts the top five common first and last names for the English language, with the names being paired up to full names in their order of commonness. First names were sourced from the US Social Security Administration[2] and last names from Wiktionary[3].

6 Results

6.1 Performance on Court Rulings

Re-identifications on Rulings Test Set Among all evaluated models, only legal_xlm_roberta (561M) and legal_swiss_roberta (561M) re-identified a single entity from 7673 rulings. As discussed later in Section 6.2, this aligns with expectations since evaluated models, excluding GPT-4 and mT0, do not meet key factors for effective re-identification: input length, model size, and instruction tuning. Despite their smaller size and lack of instruction tuning, these models made some reasonable guesses. Conversely, larger multilingual models like GPT-4 and mT0 failed to give credible guesses. Notably, GPT-4 was tested on just the top 50 most reasonably predicted examples from other models due to resource constraints. Potentially reflecting OpenAI's commitment to privacy alignment, GPT-4 consistently

[2]https://www.ssa.gov/oact/babynames/decades/century.html

[3]https://en.wiktionary.org/wiki/Appendix:English_surnames_(England_and_Wales)

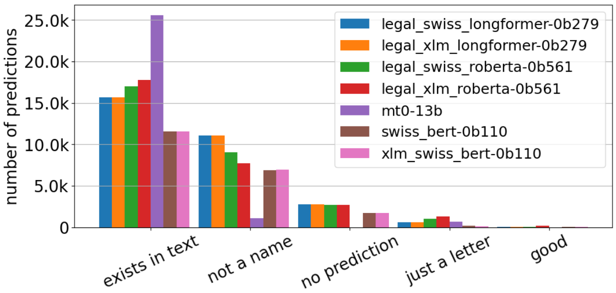


Figure 2: Categorized predictions for rulings

Model	Size [B]	**PNMS** ↑	NLD ↓	W-PNMS ↑
GPT-4	1800	0.71	0.17	0.65
GPT-3.5	175	0.52	0.23	0.46
mT0	13	0.37	0.42	0.31
Flan_T5	11	0.37	0.45	0.30
incite	3	0.37	0.53	0.30
Flan_T5	3	0.35	0.48	0.29
BLOOMZ	7.1	0.34	0.45	0.29
T0	11	0.34	0.45	0.28

Table 1: Models w/ W-PNMS > 0.28 on Wikipedia dataset

indicated that the person was not present in the text, refraining from leaking training data or making speculative guesses. mT0, trained on mC4 likely containing Swiss news articles, underperformed despite strong performance on the Wikipedia dataset, treating the text as cloze test instead of attempting to guess names. Due to resource constraints, only top two predictions from mT0 were possible. However, they yielded no reasonable output, suggesting the top three to five wouldn't have improved results. While mT0's predictions lacked meaningful output, the success of smaller models to predict some believable speculations suggests they might not have been relying solely on chance but made informed guesses. As shown in Figure 2, most predictions corresponded to words already present in the ruling or were not a name. Excluding the few good predictions, the rest consisted of empty predictions or single letters.

Re-identifications on Hand-picked Rulings Applying the same models on the hand-picked dataset, the results were not better even though for this small dataset we had the confirmation that all rulings were re-identifiable with the information in the training data. None of the models were able to predict any entity correctly. However, using the RAG approach worked much better. When passing the relevant news articles and the corresponding court ruling to the context, GPT-3.5-turbo-16k was able to identify 4 out of 7 entities, with the full name for one example. GPT-4 performed even better, correctly identifying 5 out of 7, with the full name for one example. Interestingly, the two cases which were easiest for us humans to identify were not identified by either model. This result not only suggests that re-identification by training on enough news articles could be possible, but that models powerful enough to understand the task and the given information are capable of using not only their training data information, but simultaneously ingest relevant additional information. It could even be possible to re-identify decisions without any pre-training by ingesting the full news dataset and embed information on a large scale, leading to large scale re-identifications in the worst case.

6.2 Factors for Re-identification on Wikipedia

Performance in re-identification tasks varied significantly. Some larger models like Flan_T5 or mT0 achieved high scores, with GPT-4 even surpassing 0.6 in W-PNMS and low NLD. Conversely, models like Pythia or Cerebras-GPT underperformed, often falling below the guessing baseline. Table 1 lists the top performers on the Wikipedia dataset. Due to resource constraints, ablations focus on these models, offering clearer insights into methodological differences. Comprehensive model performance is detailed in Table 4. Analyzing factors for good performance in re-identification tasks, we found that performance varied strongly, with some larger models such as Flan_T5 or mT0 reaching scores above 0.3 or for GPT-4 even above 0.6 for W-PNMS with very low NLD while models like Pythia or cerebras-GPT performed very poorly, below the guessing baseline even. Table 1 shows the best performing models on the Wikipedia dataset. Ablations prioritize top-performing models because of resource constraints and the need for interpretability. Not every model is assessed on all datasets, as comparing high-performing models across different benchmarks provides clearer insights into methodological differences than their lower-performing counterparts. The full list for all models and their performance is shown in Table 4.

Input length Testing a selection of models (Figure 3) revealed that performance improves with increasing input size, though the degree of improvement varies among models. While models which performed better at 1k input characters gained performance logistically with increasing input length, the initially poorly performing models were likely to increase their performance gain more steeply. The initially better performing models are all much larger and are all instruction tuned. The model roberta_squad which is only 355M parameters but fine-tuned on a QA dataset was able to gain a strong increase in performance nearly matching the top performers. The small models which were not instruction tuned remained at poor performance or with a slow increase in performance. It can be stated that longer input is most likely a critical factor for good performance as long as the maximum sequence length for a model is not exceeded.

Instruction tuning As stated in Section 5.1, the prompt given to models heavily influences the accuracy of the predictions (Li and Liang 2021; Zamfirescu-Pereira et al. 2023). As shown in Figure 4, instruction tuned models perform much better at re-identification. Even though both versions of each model were pretrained on the same datasets and contain the same knowledge, the instruction tuned models were far more likely to understand the task and retrieve the correct name, which is consistent with previous research (Longpre et al. 2023; Ouyang et al. 2022; Muennighoff et al. 2023).

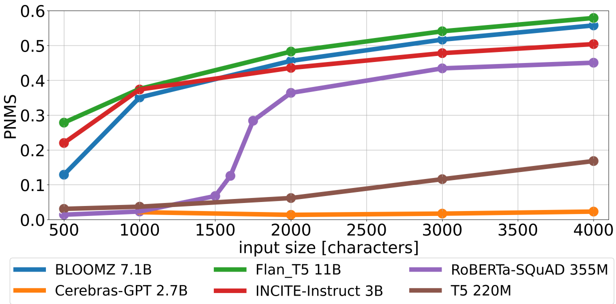


Figure 3: Comparing models across input lengths

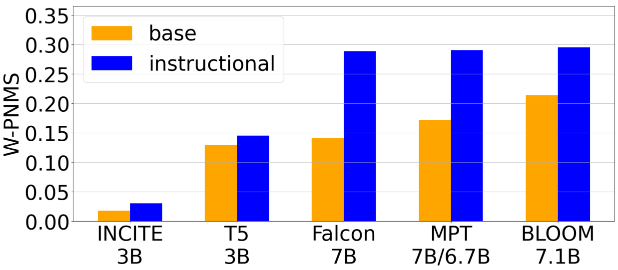


Figure 4: Base vs. instruction tuned performance

Sampling methods We see in Figure 5 that overall the variation in performance is small. Only the greedy algorithm performed much worse; however, it only predicts a single entity while the others may give five different predictions. Performance varies most for beam search: Incite_instruct performed worst, while BLOOMZ achieved its best results. However, this does not mean that top-k is the best sampling method for re-identifications. Looking at the precision of decisions, the NLD is better for predictions produced with beam search, meaning beam search can deliver more precise re-identifications, while top-k might find generally more likely names, but not necessarily the exact full name. With two out of three evaluated models performing best with beam search and NLD being best with this sampling strategy we used beam search for all other experiments.

Re-Identification methods In Figure 6 we compare fill

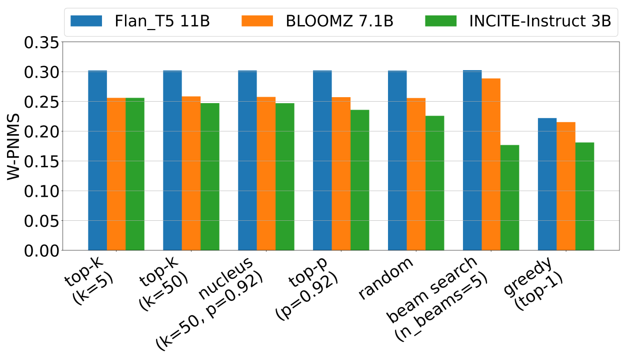


Figure 5: Generation methods of top performing models

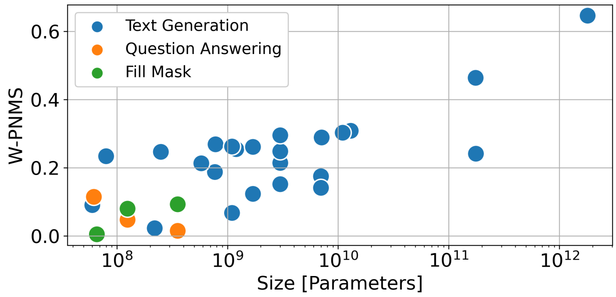


Figure 6: Parameter efficiency across model types

Data Config	PNMS ↑	NLD ↓	LNMS ↑	W-PNMS ↑
	input constrained to 1000 characters			
original	$0.35_{\pm 0.04}$	$0.52_{\pm 0.05}$	$0.25_{\pm 0.03}$	$0.29_{\pm 0.03}$
paraphrased	$0.33_{\pm 0.03}$	$0.48_{\pm 0.03}$	$0.24_{\pm 0.02}$	$0.27_{\pm 0.02}$
	input constrained to eight sentences			
original	$0.33_{\pm 0.05}$	$0.57_{\pm 0.11}$	$0.22_{\pm 0.04}$	$0.26_{\pm 0.05}$
paraphrased	$0.28_{\pm 0.03}$	$0.51_{\pm 0.04}$	$0.19_{\pm 0.03}$	$0.22_{\pm 0.03}$

Table 2: Average and std over top performers (incite_instruct, Flan_T5, T0, BLOOMZ, mT0)

mask, QA and text generation models across model sizes. Note that we excluded text generation models below the random name guessing baseline because they failed to follow the instructions (i.e., Pythia, Cerebras-GPT, Falcon, Falcon-Instruct, GPT-J). We find models performing the fill mask and question answering tasks to underperform the text generation models across the board, and even at the same model size. While performance increases for models performing the fill mask task, the opposite happens for models doing QA when scaling up model size. Given that most large-scale models are text generation models, they tend to outperform fill mask and QA counterparts. The improved performance of these models can be attributed to their ability to retain more information, a characteristic inherent to larger models (Roberts, Raffel, and Shazeer 2020).

Original vs paraphrased In Table 2 we compare the effect of paraphrases on re-identification performance. We find models to perform slightly better on the original text, both when we constrain the input by the number of characters and by a number of sentences (to ensure that the same amount of information is given). Remember that the average paraphrased sentence is significantly shorter than the average original sentence (95 vs 141 characters, see Appendix E.1).

This comes with the danger that very specific details which would have otherwise given the clue for a re-identification could be lost.

Model Size Comparing differently sized versions of a model as shown in Figure 1, a clear performance boost is observed as model size increases, consistent with prior research suggesting better knowledge retrieval with larger models (Roberts, Raffel, and Shazeer 2020). Performance typically improves significantly when transitioning from smaller to medium-sized models, though the gains diminish for larger models. While not all models performed the same for the larger model sizes, the general performance progression indicates that performance gains stagnate when models are scaled beyond their sweet spot between size and performance. On average this turning point appears to be at around 3B parameters but varies for different models with some models still reaching better performances for much larger sizes. Models with overall low performance do not see as large of a performance increase with increasing model size. The small increase might be due to the model understanding the task better but still not being able to retrieve the requested name, but by chance giving more diverse answers and coincidentally matching some predictions.

7 Conclusions and Future Work

7.1 Answering the Main Research Questions

RQ1: Performance of LLMs on re-identifications: How effectively can various LLMs re-identify masked persons within Wikipedia pages and in Swiss court rulings?
We find that vanilla LLMs can not re-identify individuals in court rulings. Additionally, relatively small models trained on news articles and court rulings respectively can barely guess credible names. Finally, by augmenting strong LLMs with retrieval on a manually curated dataset, a small subset of individuals can be re-identified.
RQ2: Influential factors: What are the key factors that influence the performance of LLMs in re-identification tasks?
We identified three influential factors affecting the performance of LLMs in re-identification tasks: model size, input length, and instruction tuning.
RQ3: Privacy Implications: How will evolving LLM capabilities and their use in re-identifications affect the preservation of privacy in anonymized court rulings in Switzerland?
We demonstrate that, for now, significant privacy breaches using LLMs on a large scale are unattainable without considerable resources. Yet, the Wikipedia benchmark revealed that larger models, when exposed to adequate pre-training information, can proficiently identify entities.

7.2 Conclusions

Currently, the risk of vanilla LLMs re-identifying individuals in Swiss court rulings is limited. However, if a malicious actor were to invest significant resources by pre-training on relevant data and augmenting the LLM with retrieval, we fear increased re-identification risk. We identified three major factors influencing re-identification performance: the model's size, the length of the input, and instruction tuning. As technology progresses, the implications for privacy become more pronounced. It is imperative to tread cautiously to ensure the sanctity of privacy in legal documentation remains uncompromised.

7.3 Future Work

Liu et al. (2023) showed that models extract information better if it is located at the start or end of large contexts. For the large models which can ingest full court rulings, this could mean that ordering parts of the rulings by their relevancy for re-identifications could improve chances for successful re-identifications. Further research is required to analyze which parts of rulings are the most relevant for re-identification.

Specific pre-training of large models on relevant data and sophisticated prompting techniques such as chain of thought (Wei et al. 2023) may increase re-identification risk.

In this work, we only considered information in textual form, either embedded in the weights by pretraining or put into the context with retrieval. Future work may additionally investigate the use of more structured information, such as structured databases or knowledge graphs.

Limitations

Ambiguity in Re-identification Metrics: The metrics employed to gauge the re-identification risk present inherent ambiguities. By comparing exact name matches and assessing the general similarity to the target name, we can infer the likelihood of manual re-identification. Yet, for lesser-known individuals or those with widespread names, a generic first name paired with a surname might be insufficient for precise identification. Thus, manual scrutiny remains necessary to distill the correct person from the model's suggested candidates. Essentially, while models scoring highly on our metrics can suggest potential identities, they might not always identify a person with certainty, especially when common names or lesser-known individuals are involved.

Scope of the Study: Our research focused on Swiss court decisions, and we did not extend our study to public court decisions from other jurisdictions. Differences in legal cultures, language nuances, and documentation standards across jurisdictions could introduce variables that could affect the generalizability of our findings.

Ethics and Broader Impact

Abundant open publication of court rulings is crucial for holding the judicial system accountable and thus for a functioning democratic state. Additionally, it greatly facilitates legal research by eliminating barriers to accessing case documents. However, courts are reluctant to publish rulings, fearing repercussions due to possible privacy breaches. Solid automated anonymization is key for courts publishing decisions more plentiful, faster, and regularly. Strong re-identification methods can be a valuable tool to stress-test anonymization systems in the absence of formal guarantees of security. However, re-identification techniques, akin to penetration testing in security, are dual-use technologies by nature and thus pose a certain risk if misused. Fortunately, our findings indicate that without a significant investment of resources and expertise, large scale re-identification using LLMs is currently not feasible.

Acknowledgements

We thank Daniel Brunner from the Swiss federal Supreme court for evaluating our predictions on court rulings. A kind thanks goes to Dominique Schläfli, the law student who helped us navigate the complicated texts of court rulings to curate a dataset of re-identified court rulings allowing us to benchmark different re-identification strategies.

References

AI, T. 2023. Releasing 3B and 7B RedPajama-INCITE family of models including base, instruction-tuned & chat models.

AlKhamissi, B.; Li, M.; Celikyilmaz, A.; Diab, M.; and Ghazvininejad, M. 2022. A Review on Language Models as Knowledge Bases. *arXiv:2204.06031 [cs]*. ArXiv: 2204.06031.

Almazrouei, E.; Alobeidli, H.; Alshamsi, A.; Cappelli, A.; Cojocaru, R.; Debbah, M.; Goffinet, E.; Heslow, D.; Launay, J.; Malartic, Q.; Noune, B.; Pannier, B.; and Penedo, G. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Biderman, S.; Schoelkopf, H.; Anthony, Q.; Bradley, H.; O'Brien, K.; Hallahan, E.; Khan, M. A.; Purohit, S.; Prashanth, U. S.; Raff, E.; Skowron, A.; Sutawika, L.; and van der Wal, O. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. ArXiv:2304.01373 [cs].

Black, S.; Biderman, S.; Hallahan, E.; Anthony, Q.; Gao, L.; Golding, L.; He, H.; Leahy, C.; McDonell, K.; Phang, J.; Pieler, M.; Prashanth, U. S.; Purohit, S.; Reynolds, L.; Tow, J.; Wang, B.; and Weinbach, S. 2022. GPT-NeoX-20B: An Open-Source Autoregressive Language Model. ArXiv:2204.06745 [cs].

Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Driessche, G. v. d.; Lespiau, J.-B.; Damoc, B.; Clark, A.; Casas, D. d. L.; Guy, A.; Menick, J.; Ring, R.; Hennigan, T.; Huang, S.; Maggiore, L.; Jones, C.; Cassirer, A.; Brock, A.; Paganini, M.; Irving, G.; Vinyals, O.; Osindero, S.; Simonyan, K.; Rae, J. W.; Elsen, E.; and Sifre, L. 2022. Improving language models by retrieving from trillions of tokens. ArXiv:2112.04426 [cs].

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Carlini, N.; Ippolito, D.; Jagielski, M.; Lee, K.; Tramer, F.; and Zhang, C. 2023. Quantifying Memorization Across Neural Language Models. ArXiv:2202.07646 [cs].

Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; Oprea, A.; and Raffel, C. 2021. Extracting Training Data from Large Language Models. ArXiv:2012.07805 [cs].

Chan, B.; Möller, T.; Pietsch, M.; and Soni, T. 2020. roberta-base for QA.

Chen, D.; Fisch, A.; Weston, J.; and Bordes, A. 2017. Reading Wikipedia to Answer Open-Domain Questions. *arXiv:1704.00051 [cs]*. ArXiv: 1704.00051.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. ArXiv:2210.11416 [cs].

Dettmers, T.; Lewis, M.; Belkada, Y.; and Zettlemoyer, L. 2022. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. Number: arXiv:2208.07339 arXiv:2208.07339 [cs].

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805. _eprint: 1810.04805.

Dey, N.; Gosal, G.; Zhiming; Chen; Khachane, H.; Marshall, W.; Pathria, R.; Tom, M.; and Hestness, J. 2023. Cerebras-GPT: Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster. ArXiv:2304.03208 [cs].

EUGH. 2018. Ab 1. Juli 2018 werden Vorabentscheidungssachen, an denen natürliche Personen beteiligt sind, anonymisiert. *Pressemitteilung*.

Hamann, H. 2021. Der blinde Fleck der deutschen Rechtswissenschaft – Zur digitalen Verfügbarkeit instanzgerichtlicher Rechtsprechung. *JuristenZeitung (JZ)*, 76(13): 656–665. Place: Tübingen Publisher: Mohr Siebeck.

Ippolito, D.; Tramèr, F.; Nasr, M.; Zhang, C.; Jagielski, M.; Lee, K.; Choquette-Choo, C. A.; and Carlini, N. 2023. Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy. ArXiv:2210.17546 [cs].

Jiang, Z.; Anastasopoulos, A.; Araki, J.; Ding, H.; and Neubig, G. 2020. X-FACTR: Multilingual Factual Knowledge Retrieval from Pretrained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5943–5959. Online: Association for Computational Linguistics.

Karanam, S.; Gou, M.; Wu, Z.; Rates-Borras, A.; Camps, O.; and Radke, R. J. 2018. A Systematic Evaluation and Benchmark for Person Re-Identification: Features, Metrics, and Datasets. ArXiv:1605.09653 [cs].

Kassner, N.; Dufter, P.; and Schütze, H. 2021. Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models. *arXiv:2102.00894 [cs]*. ArXiv: 2102.00894.

Katz, D. M.; Hartung, D.; Gerlach, L.; Jana, A.; and Bommarito II, M. J. 2023. Natural Language Processing in the Legal Domain. ArXiv:2302.12039 [cs].

Khurana, D.; Koli, A.; Khatter, K.; and Singh, S. 2023. Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3): 3713–3744.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv:2005.11401 [cs].

Lewis, P.; Stenetorp, P.; and Riedel, S. 2020. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. ArXiv:2008.02637 [cs].

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. ArXiv:2101.00190 [cs].

Lim, D. S. 2021. dslim/bert-base-NER · Hugging Face.

Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2023. Lost in the Middle: How Language Models Use Long Contexts. ArXiv:2307.03172 [cs].

Liu, X.; Ji, K.; Fu, Y.; Tam, W. L.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-Tuning v2: Prompt Tuning Can Be Comparable to Fine-tuning Universally Across Scales and Tasks. ArXiv:2110.07602 [cs].

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Longpre, S.; Hou, L.; Vu, T.; Webson, A.; Chung, H. W.; Tay, Y.; Zhou, D.; Le, Q. V.; Zoph, B.; Wei, J.; and Roberts, A. 2023. The Flan Collection: Designing Data and Methods for Effective Instruction Tuning. ArXiv:2301.13688 [cs].

Lorenz, P. 2017. Machtwort vom BGH: Urteile sind für alle da.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T. L.; Bari, M. S.; Shen, S.; Yong, Z.-X.; Schoelkopf, H.; and others. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Muennighoff, N.; Wang, T.; Sutawika, L.; Roberts, A.; Biderman, S.; Scao, T. L.; Bari, M. S.; Shen, S.; Yong, Z.-X.; Schoelkopf, H.; Tang, X.; Radev, D.; Aji, A. F.; Almubarak, K.; Albanie, S.; Alyafeai, Z.; Webson, A.; Raff, E.; and Raffel, C. 2023. Crosslingual Generalization through Multitask Finetuning. ArXiv:2211.01786 [cs].

Munz, T. 2022. Staatshaftung für mangelhafte Anonymisierung von publizierten Gerichtsurteilen. *Richterzeitung*, (1).

Niklaus, J.; Matoshi, V.; Sturmer, M.; Chalkidis, I.; and Ho, D. E. 2023. MultiLegalPile: A 689GB Multilingual Legal Corpus. *ArXiv*, abs/2306.02069.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. ArXiv:2203.02155 [cs].

Petroni, F.; Rocktäschel, T.; Riedel, S.; Lewis, P.; Bakhtin, A.; Wu, Y.; and Miller, A. 2019. Language Models as Knowledge Bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2463–2473. Hong Kong, China: Association for Computational Linguistics.

Poerner, N.; Waltinger, U.; and Schütze, H. 2020. E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 803–818. Online: Association for Computational Linguistics.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Technical Report arXiv:1910.10683, arXiv. ArXiv:1910.10683 [cs, stat] type: article.

Rasiah, V.; Stern, R.; Matoshi, V.; Stürmer, M.; Chalkidis, I.; Ho, D. E.; and Niklaus, J. 2023. SCALE: Scaling up the Complexity for Advanced Language Model Evaluation. ArXiv:2306.09237 [cs].

Remy, P. 2021. Name Dataset. Publication Title: GitHub repository.

Roberts, A.; Raffel, C.; and Shazeer, N. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? *arXiv:2002.08910 [cs, stat]*. ArXiv: 2002.08910.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Sanh, V.; Webson, A.; Raffel, C.; Bach, S.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Raja, A.; Dey, M.; Bari, M. S.; Xu, C.; Thakker, U.; Sharma, S. S.; Szczechla, E.; Kim, T.; Chhablani, G.; Nayak, N.; Datta, D.; Chang, J.; Jiang, M. T.-J.; Wang, H.; Manica, M.; Shen, S.; Yong, Z. X.; Pandey, H.; Bawden, R.; Wang, T.; Neeraj, T.; Rozen, J.; Sharma, A.; Santilli, A.; Fevry, T.; Fries, J. A.; Teehan, R.; Scao, T. L.; Biderman, S.; Gao, L.; Wolf, T.; and Rush, A. M. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; Ilić, S.; Hesslow, D.; Castagné, R.; Luccioni, A. S.; Yvon, F.; Gallé, M.; Tow, J.; Rush, A. M.; Biderman, S.; Webson, A.; Ammanamanchi, P. S.; Wang, T.; Sagot, B.; Muennighoff, N.; del Moral, A. V.; Ruwase, O.; Bawden, R.; Bekman, S.; McMillan-Major, A.; Beltagy, I.; Nguyen, H.; Saulnier, L.; Tan, S.; Suarez, P. O.; Sanh, V.; Laurençon, H.; Jernite, Y.; Launay, J.; Mitchell, M.; Raffel, C.; Gokaslan, A.; Simhi, A.; Soroa, A.; Aji, A. F.; Alfassy, A.; Rogers, A.; Nitzav, A. K.; Xu, C.; Mou, C.; Emezue, C.; Klamm, C.; Leong, C.; van Strien, D.; Adelani, D. I.; Radev, D.; Ponferrada, E. G.; Levkovizh, E.; Kim, E.; Natan, E. B.; De Toni, F.; Dupont, G.; Kruszewski, G.; Pistilli, G.; Elsahar, H.; Benyamina, H.; Tran, H.; Yu, I.; Abdulmumin, I.; Johnson, I.; Gonzalez-Dios, I.; de la Rosa, J.; Chim, J.; Dodge, J.; Zhu, J.; Chang, J.; Frohberg, J.; Tobing, J.; Bhattacharjee, J.; Almubarak, K.; Chen, K.; Lo, K.; Von Werra, L.; Weber, L.; Phan, L.; allal, L. B.; Tanguy, L.; Dey, M.; Muñoz, M. R.; Masoud, M.; Grandury, M.; Šaško, M.; Huang, M.; Coavoux, M.; Singh, M.; Jiang, M. T.-J.; Vu, M. C.; Jauhar, M. A.; Ghaleb, M.; Subramani, N.; Kassner, N.; Khamis, N.; Nguyen, O.; Espejel, O.; de Gibert, O.; Villegas, P.; Henderson, P.; Colombo, P.; Amuok, P.; Lhoest, Q.; Harliman, R.; Bommasani, R.; López, R. L.; Ribeiro, R.; Osei, S.; Pyysalo, S.; Nagel, S.; Bose, S.; Muhammad, S. H.; Sharma, S.; Longpre, S.; Nikpoor, S.; Silberberg, S.; Pai, S.; Zink, S.; Torrent, T. T.; Schick, T.; Thrush, T.; Danchev, V.; Nikoulina, V.; Laippala, V.; Lepercq, V.; Prabhu, V.; Alyafeai, Z.; Talat, Z.; Raja, A.; Heinzerling, B.; Si, C.; Taşar, D. E.; Salesky, E.; Mielke, S. J.; Lee, W. Y.; Sharma, A.; Santilli, A.; Chaffin, A.; Stiegler, A.; Datta, D.; Szczechla, E.; Chhablani, G.; Wang, H.; Pandey, H.; Strobelt, H.; Fries, J. A.; Rozen, J.; Gao, L.; Sutawika, L.; Bari, M. S.; Al-shaibani, M. S.; Manica, M.; Nayak, N.; Teehan, R.; Albanie, S.; Shen, S.; Ben-David, S.; Bach, S. H.; Kim, T.; Bers, T.; Fevry, T.; Neeraj, T.; Thakker, U.; Raunak, V.; Tang, X.; Yong, Z.-X.; Sun, Z.; Brody, S.; Uri, Y.; Tojarieh, H.; Roberts, A.; Chung, H. W.; Tae, J.; Phang, J.; Press, O.; Li, C.; Narayanan, D.; Bourfoune, H.; Casper, J.; Rasley, J.; Ryabinin, M.; Mishra, M.; Zhang, M.; Shoeybi, M.; Peyrounette, M.; Patry, N.; Tazi, N.; Sanseviero, O.; von Platen, P.; Cornette, P.; Lavallée, P. F.; Lacroix, R.; Rajbhandari, S.; Gandhi, S.; Smith, S.; Requena, S.; Patil, S.; Dettmers, T.; Baruwa, A.; Singh, A.; Cheveleva, A.; Ligozat, A.-L.; Subramonian, A.; Névéol, A.; Lovering, C.; Garrette, D.; Tunuguntla, D.; Reiter, E.; Taktasheva, E.; Voloshina, E.; Bogdanov, E.; Winata, G. I.; Schoelkopf, H.; Kalo, J.-C.; Novikova, J.; Forde, J. Z.; Clive, J.; Kasai, J.; Kawamura, K.; Hazan, L.; Carpuat, M.; Clinciu, M.; Kim, N.; Cheng, N.; Serikov, O.; Antverg, O.; van der Wal, O.; Zhang, R.; Zhang, R.; Gehrmann, S.; Mirkin, S.; Pais, S.; Shavrina, T.; Scialom, T.; Yun, T.; Limisiewicz, T.; Rieser, V.; Protasov, V.; Mikhailov, V.; Pruksachatkun, Y.; Belinkov, Y.; Bamberger, Z.; Kasner, Z.; Rueda, A.; Pestana, A.; Feizpour, A.; Khan, A.; Faranak, A.; Santos, A.; Hevia, A.; Unldreaj, A.; Aghagol, A.; Abdollahi, A.; Tammour, A.; HajiHosseini, A.; Behroozi, B.; Ajibade, B.; Saxena, B.; Ferrandis, C. M.; McDuff, D.; Contractor, D.; Lansky, D.; David, D.; Kiela, D.; Nguyen, D. A.; Tan, E.; Baylor, E.; Ozoani, E.; Mirza, F.; Ononiwu, F.; Rezanejad, H.; Jones, H.; Bhattacharya, I.; Solaiman, I.; Sedenko, I.; Nejadgholi, I.; Passmore, J.; Seltzer, J.; Sanz, J. B.; Dutra, L.; Samagaio, M.; Elbadri, M.; Mieskes, M.; Gerchick, M.; Akinlolu, M.; McKenna, M.; Qiu, M.; Ghauri, M.; Burynok, M.; Abrar, N.; Rajani, N.; Elkott, N.; Fahmy, N.; Samuel, O.; An, R.; Kromann, R.; Hao, R.; Alizadeh, S.; Shubber, S.; Wang, S.; Roy, S.; Viguier, S.; Le, T.; Oyebade, T.; Le, T.; Yang, Y.; Nguyen, Z.; Kashyap, A. R.; Palasciano, A.; Callahan, A.; Shukla, A.; Miranda-Escalada, A.; Singh, A.; Beilharz, B.; Wang, B.; Brito, C.; Zhou, C.; Jain, C.; Xu, C.; Fourrier, C.; Periñán, D. L.; Molano, D.; Yu, D.; Manjavacas, E.; Barth, F.; Fuhrimann, F.; Altay, G.; Bayrak, G.; Burns, G.; Vrabec, H. U.; Bello, I.; Dash, I.; Kang, J.; Giorgi, J.; Golde, J.; Posada, J. D.; Sivaraman, K. R.; Bulchandani, L.; Liu, L.; Shinzato, L.; de Bykhovetz, M. H.; Takeuchi, M.; Pàmies, M.; Castillo, M. A.; Nezhurina, M.; Sänger, M.; Samwald, M.; Cullan, M.; Weinberg, M.; De Wolf, M.; Mihaljcic, M.; Liu, M.; Freidank, M.; Kang, M.; Seelam, N.; Dahlberg, N.; Broad, N. M.; Muellner, N.; Fung, P.; Haller, P.; Chandrasekhar, R.; Eisenberg, R.; Martin, R.; Canalli, R.; Su, R.; Su, R.; Cahyawijaya, S.; Garda, S.; Deshmukh, S. S.; Mishra, S.; Kiblawi, S.; Ott, S.; Sang-aroonsiri, S.; Kumar, S.; Schweter, S.; Bharati, S.; Laud, T.; Gigant, T.; Kainuma, T.; Kusa, W.; Labrak, Y.; Bajaj, Y. S.; Venkatraman, Y.; Xu, Y.; Xu, Y.; Xu, Y.; Tan, Z.; Xie, Z.; Ye, Z.; Bras, M.; Belkada, Y.; and Wolf, T. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. ArXiv:2211.05100 [cs].

Shuster, K.; Poff, S.; Chen, M.; Kiela, D.; and Weston, J. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. ArXiv:2104.07567 [cs].

Stückelberger, B.; Evin, Y.; and Damian, C. 2021. Anzeige von Namensänderungen strafrechtlich Verurteilter nach identifizierender Medienberichterstattung | sui generis.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Touvron, H.; Martin, L.; and Stone, K. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Tsarapatsanis, D.; and Aletras, N. 2021. On the Ethical Limits of Natural Language Processing on Legal Text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3590–3599. Online: Association for Computational Linguistics.

Vamvas, J.; Graën, J.; and Sennrich, R. 2023. Swiss-BERT: The Multilingual Language Model for Switzerland. ArXiv:2303.13310 [cs].

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. *arXiv:1706.03762 [cs]*. ArXiv: 1706.03762.

Vokinger, K. N.; and Mühlematter, U. J. 2019. Re-Identifikation von Gerichtsurteilen durch "Linkage" von Daten(banken). 27.

Wang, B.; and Komatsuzaki, A. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

Wang, C.; Liu, P.; and Zhang, Y. 2021. Can Generative Pre-trained Language Models Serve as Knowledge Bases for Closed-book QA? Number: arXiv:2106.01561 arXiv:2106.01561 [cs].

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Zamfirescu-Pereira, J.; Wong, R. Y.; Hartmann, B.; and Yang, Q. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. Hamburg Germany: ACM. ISBN 978-1-4503-9421-5.

Zhang, J.; Zhao, Y.; Saleh, M.; and Liu, P. J. 2019. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. _eprint: 1912.08777.

A Technical Specifications

To run experiments with smaller models we used machines with 1024GB Memory and a NVIDIA GeForce 4090. For larger models we used the computing server of our research institute with 180GB Memory and two NVIDIA A100 80GB graphics card over NVMe. All models were run with bitsandbytes (Dettmers et al. 2022) 8bit quantization.

A.1 Hyperparameters

We did not tune any hyperparameters in this work and used default settings when not specifically stated otherwise. To optimize GPU usage we set batch sizes as large as possible, preferring multiples of 64 as suggested by NVIDIA. Exact batch sizes for all models are documented in the code base accompanying this work.

A.2 Repeatability and Variance

To verify the consistency of our results, given that each model was run only once per experiment, we conducted a brief test using mT0 with the same configuration across three separate runs without setting specific seeds. All results were identical, reinforcing our decision to conduct single runs for each model and configuration.

A.3 Code

All code for experiments, evaluation and plots is available at our official Github repository: https://github.com/Skatinger/Anonymity-at-Risk-Assessing-Re-Identification-Capabilities-of-Large-Language-Models.

B Error Analysis

For the court rulings, many predictions were single letters like *X.__*, common in rulings and often the correct content before the <mask> insertion. For mask-filling models, this is expected, hinting the name might be unknown or overshadowed by frequent fillers. Notably, GPT-4's dominant prediction was "I don't know," despite clear instructions to guess a name. We theorize that OpenAI's recent modifications, aimed at reducing GPT-4's tendency to make things up, might also deter it from making educated guesses when uncertain.

On Wikipedia, the majority of incorrect predictions were blank tokens such as newline characters or the mask token itself. Notably, smaller versions of T5 frequently predicted "True" or "False". In contrast, the largest Cerebras-GPT seemed to treat the text as a cloze test, often predicting "____," suggesting the text is a fill-in-the-blank.

Enhancements in performance could potentially be achieved by expanding prompt tuning to prompt models to make an educated guess if they do not know the correct answer, possibly reducing unusable tokens. It is likely that some models might have performed better if more time were invested in prompt engineering, but in fairness all models were tuned with a maximum of five tries.

B.1 Analyzing Model Predictions in Rulings

Analysis of predictions showed that a significant portion of predictions for rulings are names or terms already present in the ruling itself. On closer examination, many of these predictions turned out to be common legal terms or frequently mentioned law firm names. Tokens resembling anonymized entities, like *"A.___"*, fall into this category as well. While models occasionally guessed the anonymization token (<mask>) or single/double letters, the latter was less common. For terms not occurring in the text but representing full words, we used the name database by Remy (2021) to detect any possible names. With the largest part of words not categorized as names, only a small portion of predictions was classified as possible re-identifications. Our evaluation largely relied on fill mask models because no QA or text generation models were specifically designed for Swiss legal texts or news.

C In Depth Experimental Setup

Wikipedia pages that did not contain a mask within the first 1k characters in one of the configurations (original, paraphrased) were omitted. This led to 5% of examples being omitted in the worst case, leaving at least 9.5K examples for any model. For the court rulings the number of omitted pages was 915 of 7673, or 13,5%. Only GPT-3.5 and GPT-4 were able to ingest the full number of examples (see Table 3 for details). This is most likely due to the fact that some pages contain a lot of special characters from different languages, requiring many tokens for tokenizers with smaller vocabulary sizes, while tokenizers with large vocabularies can still tokenize very obscure terms into single tokens rather than requiring a token per character. Using an exact number of characters significantly simplified processing and facilitated more direct model comparisons, even when the models' maximum input token size varied from 512 to 4096 tokens. This is due to the fact that different tokenizers have different vocabulary sizes allowing models with larger tokenizers to ingest more text at once when a number of tokens rather than a number of characters or words is specified. All experiments were conducted as single runs since the test set is large enough to offset any minor variances between runs. Conducting multiple runs would have been too resource-intensive given the extensive amount of inference needed to benchmark all settings and configurations.

D Datasets

D.1 Court Rulings

The basis for our hand-picked rulings dataset and the rulings dataset with 6.7K entries from the year 2019 are both extracted from the publicly available swiss-courts rulings dataset published on HuggingFace. The dataset is available here: https://huggingface.co/datasets/rcds/swiss_rulings

D.2 Wikipedia Dataset

The created Wikipedia dataset with masked entities is publicly available on HuggingFace. Two versions exist, one version contains all data with each page as single example. The second version provides splits with examples already split into lengths which fit either 512 tokens or 4096 tokens. Consult the dataset cards for specific details.

Full dataset without splits (recommended for most tasks): https://huggingface.co/datasets/rcds/wikipedia-persons-masked

Dataset with precomputed splits (recommended for specific max sequence lengths): https://huggingface.co/datasets/rcds/wikipedia-for-mask-filling

E Additional Information

E.1 Wikipedia dataset paraphrasing

The generation used 10 beams and a temperature of 1.5, resulting in an average string edit distance of 76 per sentence between original and paraphrased versions, with original sentences averaging 141 characters and paraphrased sentences 95 characters.

E.2 Examples of Original and Paraphrased Wikipedia Text

Original sentence 1: Thomas Woodley ”Woody” Abernathy (October 16, 1908 – February 11, 1961) was a professional baseball player whose career spanned 13 seasons in minor league baseball.

Paraphrased sentence 1: There was a professional baseball player named Woody who played 13 seasons in minor league baseball.

Original sentence 2: Austin Sean Healey (born 26 October 1973 in Wallasey (now part of Merseyside, formerly Cheshire), is a former English rugby union player who played as a utility back for Leicester Tigers, and represented both England and the British & Irish Lions.

Paraphrased sentence 2: Austin Sean Healey is a former English rugby union player who played for both England and the British and Irish Lions.

E.3 Legal Concerns

The introduction of the **Datenschutz-Grundverordnung (DSGVO)** [4] on 27th of April 2018 has lead the court of justice of the European Union to enforce anonymization of court rulings. Press statement: https://curia.europa.eu/jcms/upload/docs/application/pdf/2018-06/cp180096de.pdf. The German Supreme court has ruled that all court rulings should be published anonymously [5]. A study[6] in 2021 found that less than a percent of German rulings are published.

E.4 Acronyms

PNMS Partial Name Match Score
LLM Large Language Model
LM Language Model
LNMS Last Name Match Score
NLD Normalized Levenshtein Distance
W-PNMS Weighted Partial Name Match Score
FSCS Federal Supreme Court of Switzerland
RAG Retrieval Augmented Generation
NER Named Entity Recognition
NLP Natural Language Processing
QA Question Answering

F Additional Graphs and Tables

[4] https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=celex%3A32016R0679

[5] https://juris.bundesgerichtshof.de/cgi-bin/rechtsprechung/document.py?Gericht=bgh&Art=en&nr=78212&pos=0&anz=1

[6] https://www.mohrsiebeck.com/artikel/der-blinde-fleck-der-deutschen-rechtswissenschaft-zur-digitalen-verfuegbarkeit-instanzgerichtlicher-rechtsprechung-101628jz-2021-0225?no_cache=1

Table 3: Used models: InLen is the maximum input length the model has seen during pretraining. # Parameters is the total parameter count (including the embedding layer). Corpus shows the most important dataset, for specific information see model papers.

Model	Source	InLen	# Parameters	Vocab	Corpus	# Langs
GPT-4	OpenAI (2023)	8K	1800B	n/a	n/a	n/a
GPT-3.5	Brown et al. (2020)	4K/16K	175B	256K	n/a	n/a
BLOOM	Scao et al. (2023)	2K	1.1B/1.7B/3B/7.1B	250K	ROOTS	59
BLOOMZ	Muennighoff et al. (2022)	2K	1.1B/1.7B/3B/7.1B	250K	mC4,xP3	109
T5	Raffel et al. (2020)	512	60M/220M/770M/3B/11B	32K	C4	1
Flan_T5	Chung et al. (2022)	512	80M/250M/780M/3B/11B	32K	collection (see paper)	60
T0	Sanh et al. (2022)	1K	3B/11B	32K	P3	1
mT0	Muennighoff et al. (2022)	512	580M/1.2B/13B	250K	mC4,xP3	101
Llama	Touvron et al. (2023)	2K	7B	32K	CommonCrawl,Github,Wikipedia,+others	20
Llama2	Touvron, Martin, and Stone (2023)	4K	7B/13B	32K	n/a	> 13
INCITE	AI (2023)	2K	3B	50K	RedPajama-Data-1T	1
INCITE-Instruct	AI (2023)	2k	3B	50K	RedPajama-Data-1T	1
Cerebras-GPT	Dey et al. (2023)	2K	111M/1.3/2.7/6.7/13B	50K	The Pile	1
GPT-NeoX	Black et al. (2022)	2K	20B	50K	The Pile	1
Pythia	Biderman et al. (2023)	512/768/1K/2K/2K/2.5K/4/5K	70/160/410M/1.4/2.8/6.9/12B	50K	The Pile	1
GPT-J	Wang and Komatsuzaki (2021)	4K	6B	50K	The Pile	1
Falcon	Almazrouei et al. (2023)	2K	7B	65K	RefinedWeb + custom corpora	11
Falcon-Instruct	Almazrouei et al. (2023)	2K	7B	65K	RefinedWeb,Baize + custom corpora	11
RoBERTa	Liu et al. (2019)	512	125M/355M	50K	BookCorpus,Wikipedia,+others	1
RoBERTa SQuAD	Chan et al. (2020)	386	125M/355M	50K	RoBERTa,SQuAD2.0	1
DistilBERT	Sanh et al. (2020)	768	66M	30K	Wikipedia	1
DistilBERT SQuAD	Sanh et al. (2020)	768	62M	28K	SQuAD	1
Models used only on court rulings						
SwissBERT	Vamvas, Graën, and Sennrich (2023)	514	110M	50K	Swissdox	4
Legal-Swiss-RobBERTa	Rasiah et al. (2023)	768	279M/561M	250K	Multi Legal Pile	3
Legal-Swiss-LongFormer-base	Rasiah et al. (2023)	4K	279M	250K	Multi Legal Pile	3
Legal-XLM-RobBERTa-base	Niklaus et al. (2023)	514	561M	250K	Multi Legal Pile	24
Legal-XLM-LongFormer-base	Niklaus et al. (2023)	4K	279M	250K	Multi Legal Pile	24

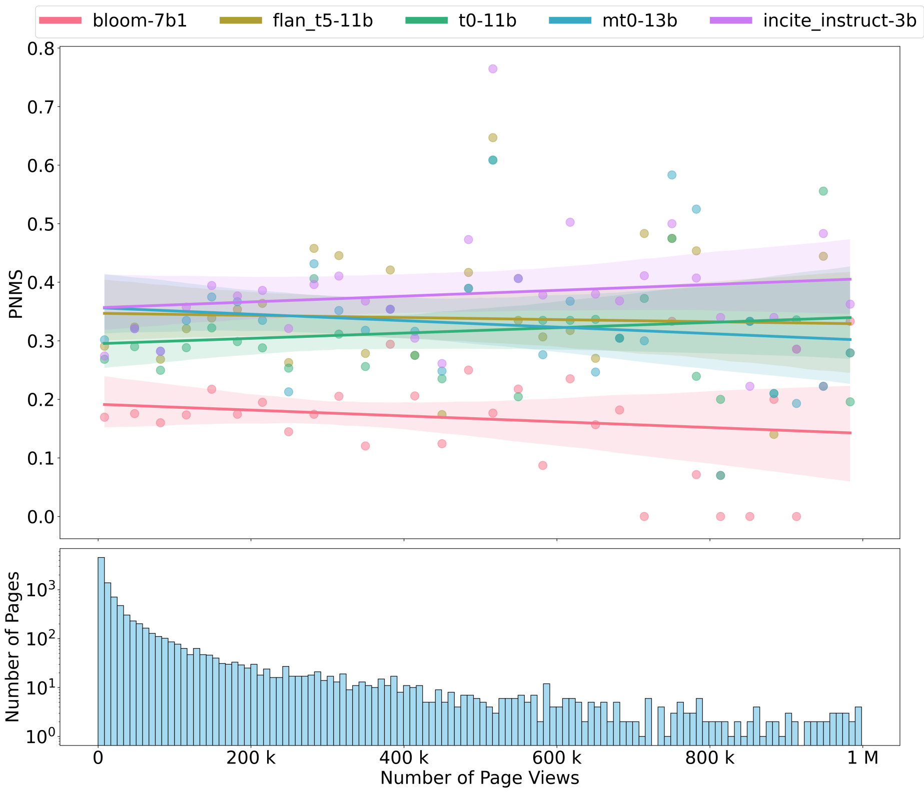


Figure 7: PNMS does not correlate with the number of views a Wikipedia page has.

Model	Size [B]	PNMS ↑	NLD ↓	W-PNMS ↑
GPT-4	1800.00	0.71	0.17	0.65
GPT-3.5	175.00	0.52	0.23	0.46
mT0	13.00	0.37	0.42	0.31
Flan_T5	11.00	0.37	0.45	0.30
INCITE-Instruct	3.00	0.37	0.53	0.30
Flan_T5	3.00	0.35	0.48	0.29
BLOOMZ	7.10	0.34	0.45	0.29
T0	11.00	0.34	0.45	0.28
Flan_T5	0.78	0.33	0.50	0.27
T0	3.00	0.32	0.46	0.27
BLOOMZ	1.10	0.31	0.48	0.26
BLOOMZ	1.70	0.31	0.47	0.26
mT0	1.20	0.31	0.47	0.25
BLOOMZ	3.00	0.29	0.48	0.25
Flan_T5	0.25	0.30	0.51	0.25
BLOOMZ	176.00	0.28	0.68	0.24
Flan_T5	0.08	0.28	0.51	0.23
T5	3.00	0.26	0.59	0.21
mT0	0.58	0.25	0.49	0.21
T5	0.77	0.23	0.56	0.19
Llama	7.00	0.26	0.54	0.17
BLOOM	7.10	0.21	0.57	0.17
BLOOM	3.00	0.18	0.58	0.15
MPT Instruct	6.70	0.19	0.61	0.15
MPT	7.00	0.20	0.53	0.14
Llama2	13.00	0.21	0.47	0.14
INCITE	3.00	0.16	0.58	0.13
Llama2	7.00	0.19	0.46	0.13
BLOOM	1.70	0.15	0.53	0.12
DistilBERT SQuAD	0.06	0.16	0.74	0.11
RoBERTa	0.35	0.18	1.03	0.09
T5	0.06	0.12	0.71	0.09
RoBERTa	0.12	0.17	1.04	0.08
BLOOM	1.10	0.09	0.60	0.07
RoBERTa SQuAD	0.12	0.07	1.40	0.05
Majority Name Baseline	-	0.11	0.64	0.04
Cerebras-GPT	13.00	0.05	1.56	0.04
Falcon-instruct	7.00	0.04	0.72	0.03
T5	0.22	0.04	0.63	0.02
Cerebras-GPT	6.70	0.03	0.78	0.02
Cerebras-GPT	1.30	0.03	0.75	0.02
GPT-NeoX	20.00	0.03	1.07	0.02
Pythia	12.00	0.04	0.82	0.02
Falcon	7.00	0.03	0.77	0.02
Pythia	0.07	0.02	0.82	0.02
Pythia	0.41	0.03	0.84	0.02
Pythia	1.40	0.03	0.84	0.02
RoBERTa SQuAD	0.35	0.02	1.61	0.02
Pythia	0.16	0.02	0.79	0.01
Cerebras-GPT	2.70	0.02	0.81	0.01
GPT-J	6.00	0.03	0.80	0.01
Pythia	2.80	0.02	0.81	0.01
Cerebras-GPT	0.11	0.02	0.92	0.01
Random Name Baseline	-	0.03	0.75	0.1
Pythia	6.90	0.01	0.97	0.01
DistilBERT	0.07	0.01	1.08	0.00

Table 4: All models on Wikipedia dataset using top five predictions and beam search with the first 1k characters as input, excluding prompt.

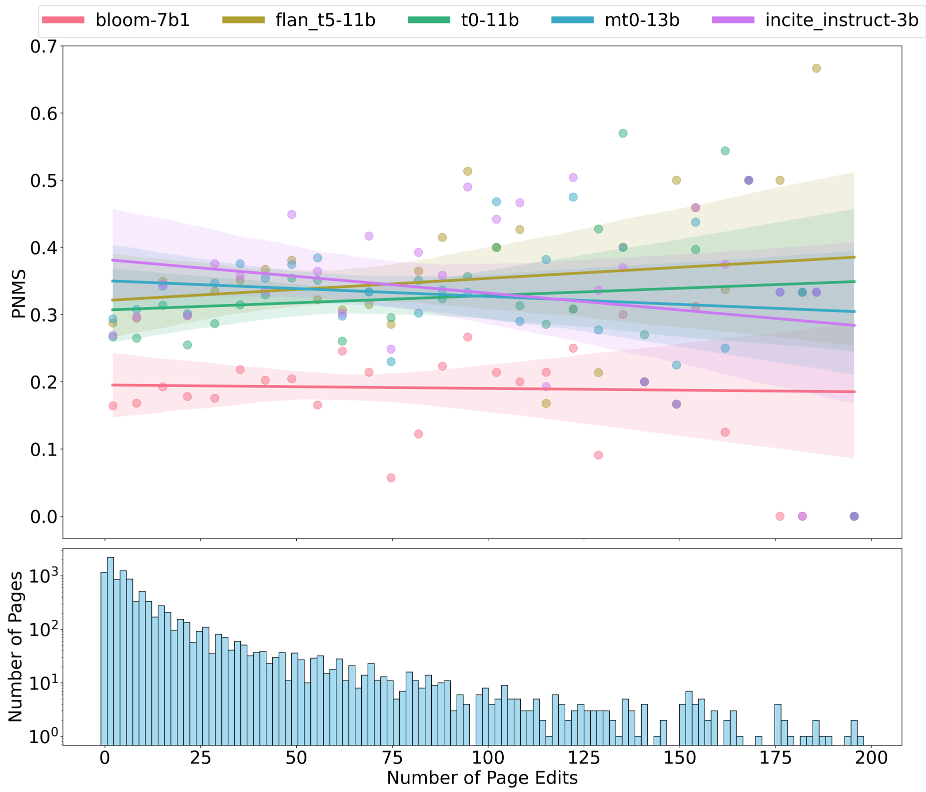


Figure 8: PNMS does not correlate with the number of edits a Wikipedia page has.

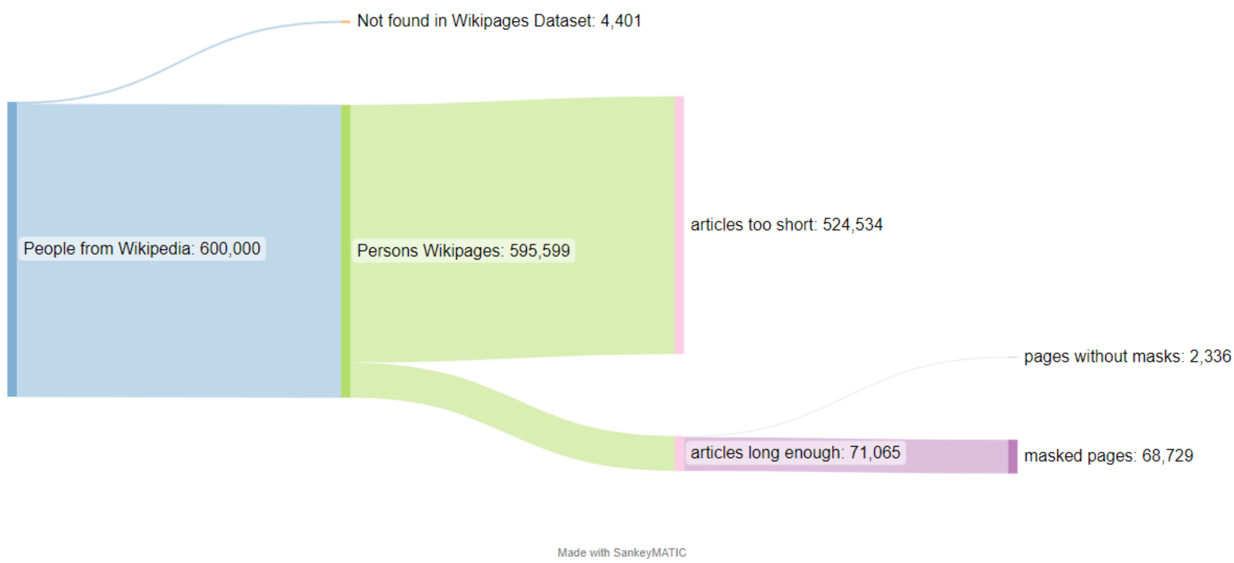


Figure 9: Selection Steps for Wikipedia Dataset

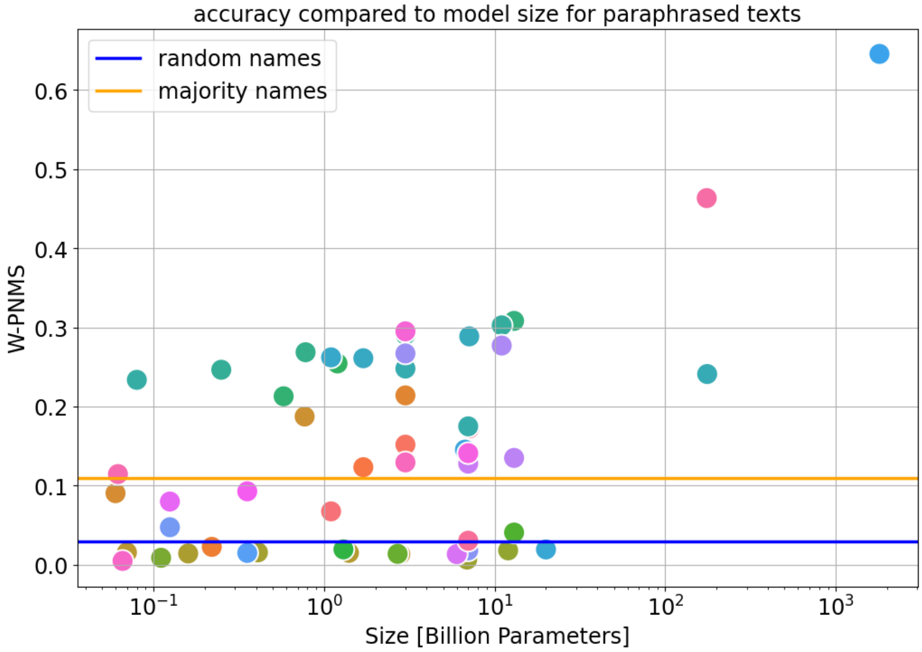


Figure 10: Overview over all evaluated models and their performance on the paraphrased config

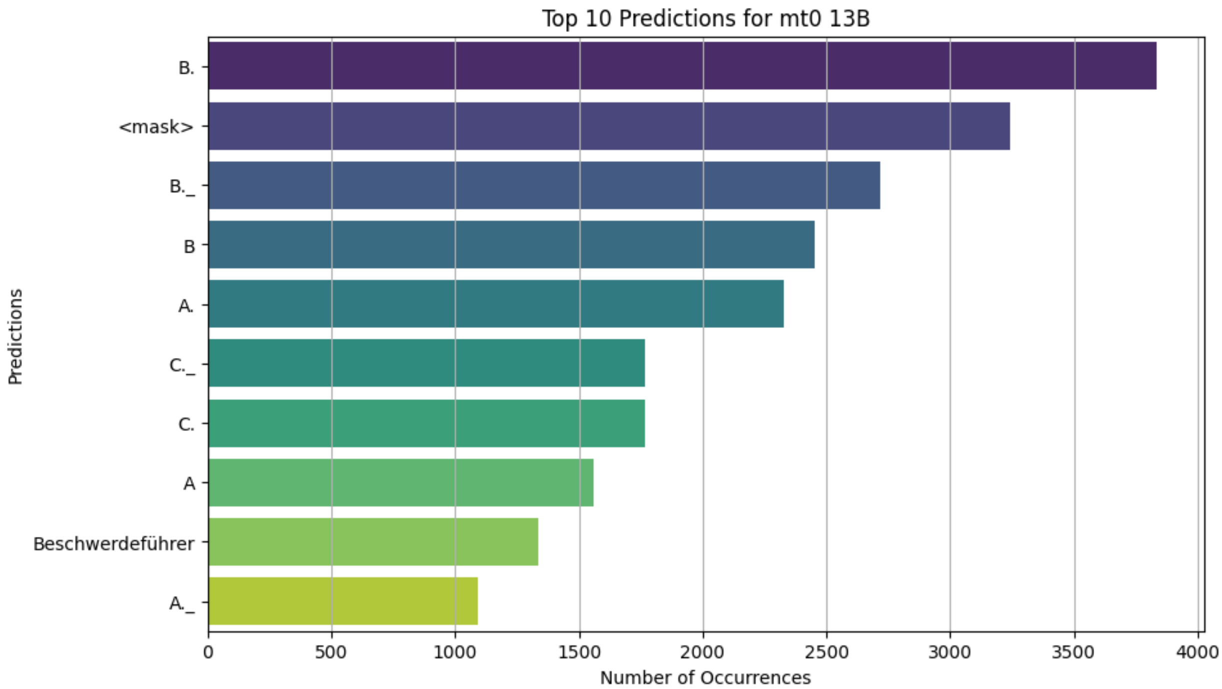


Figure 11: Most common predictions on court rulings for mT0 13B

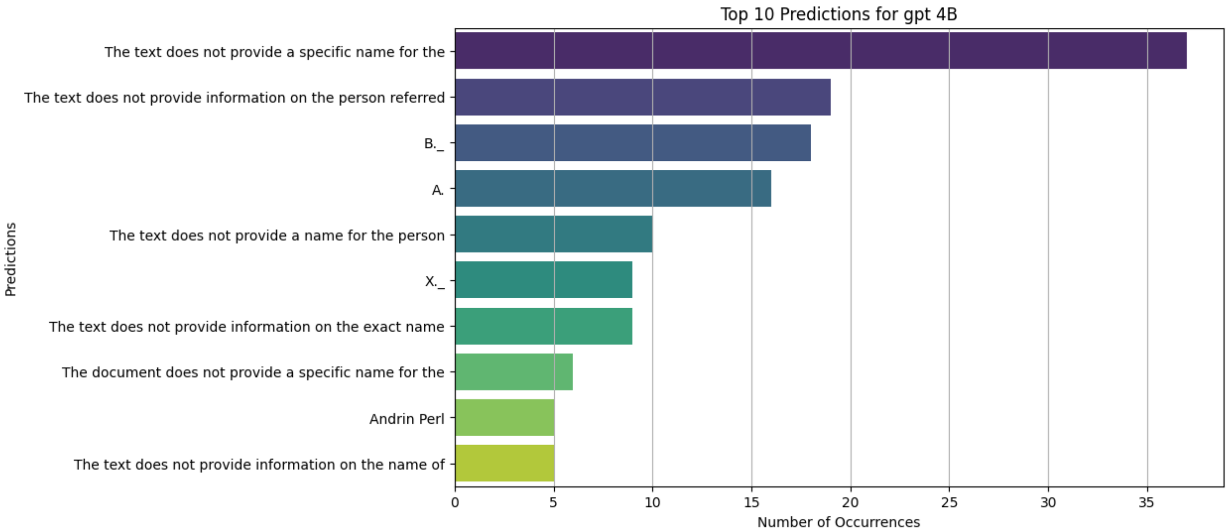


Figure 12: Most common predictions on court rulings for GPT-4

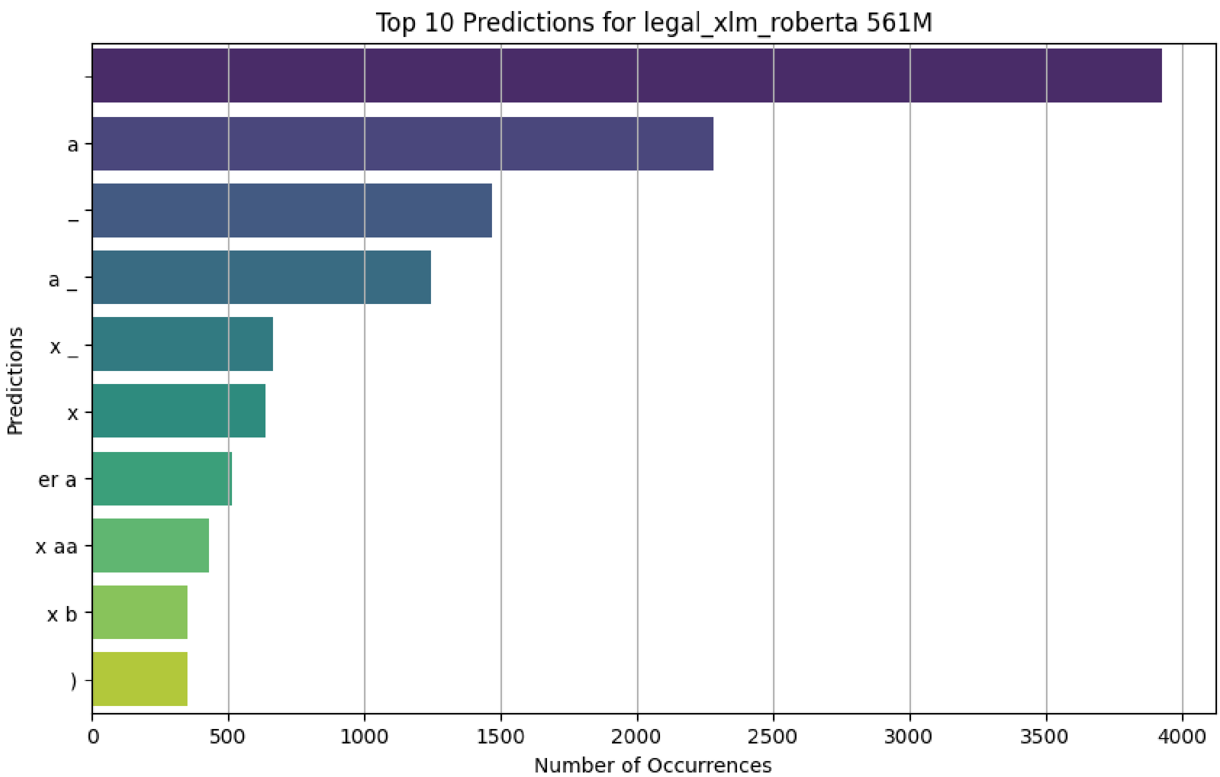


Figure 13: Most common predictions on court rulings for legal-xlm-roberta 561M

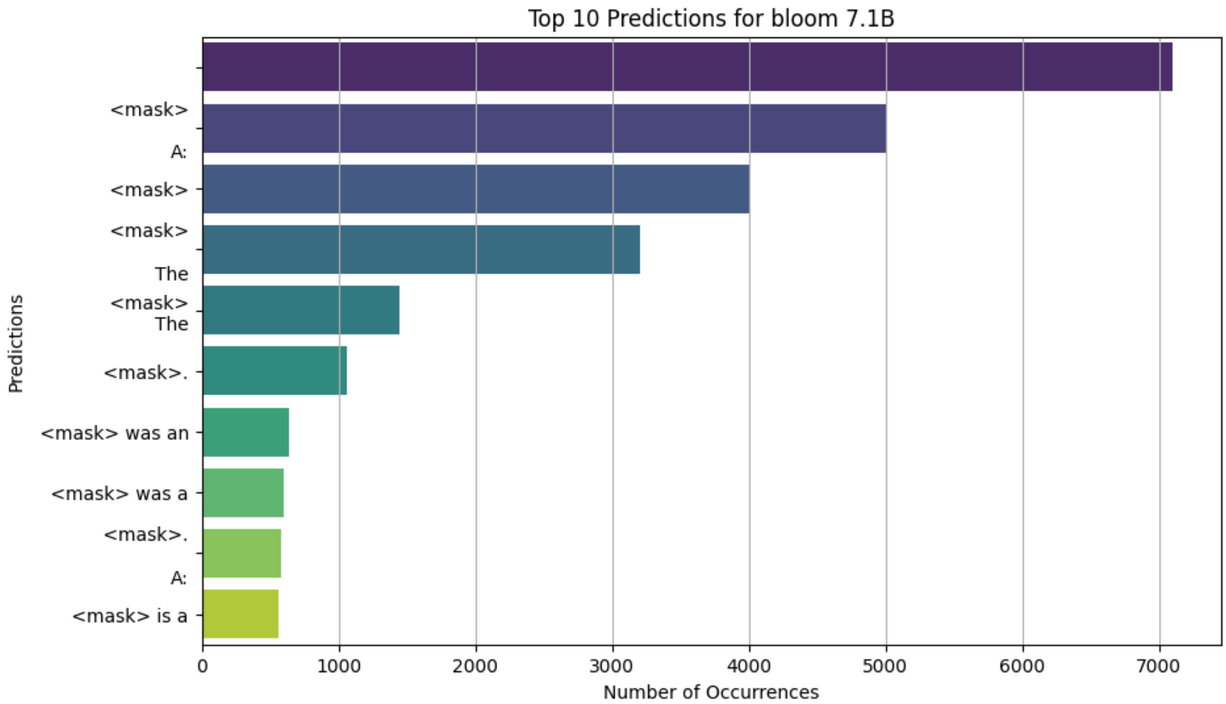


Figure 14: Most common predictions on Wikipedia for bloom 7.1B

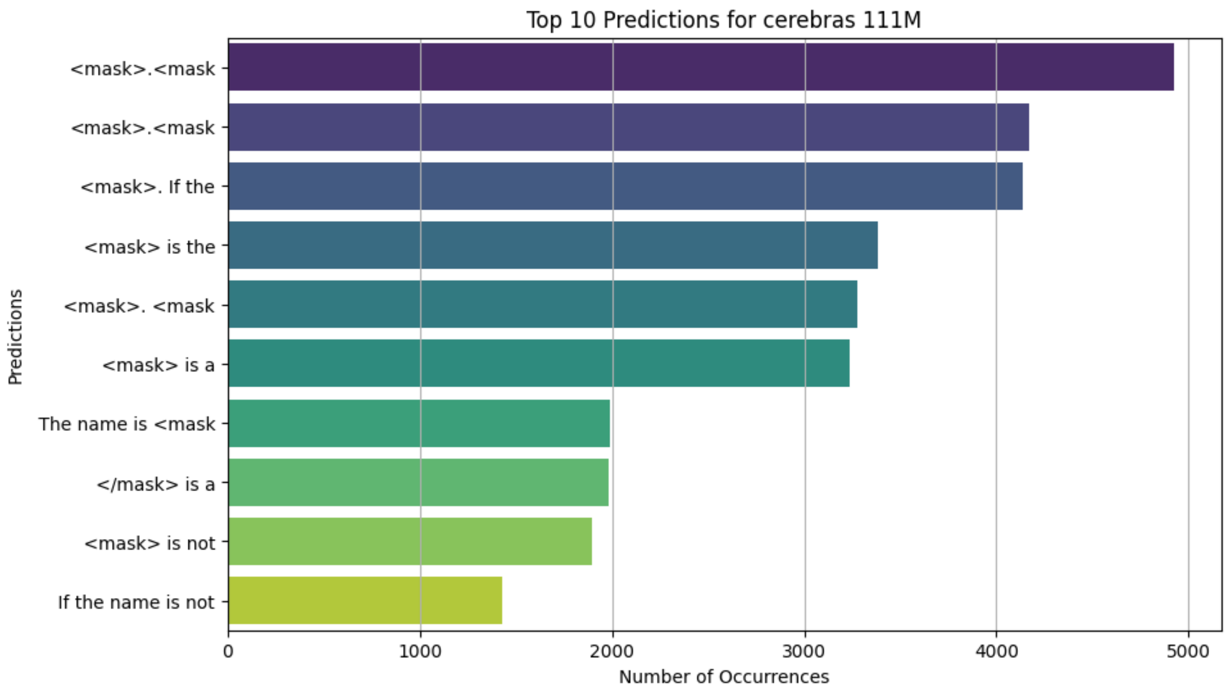


Figure 15: Most common predictions on Wikipedia for Cerebras-GPT 111M

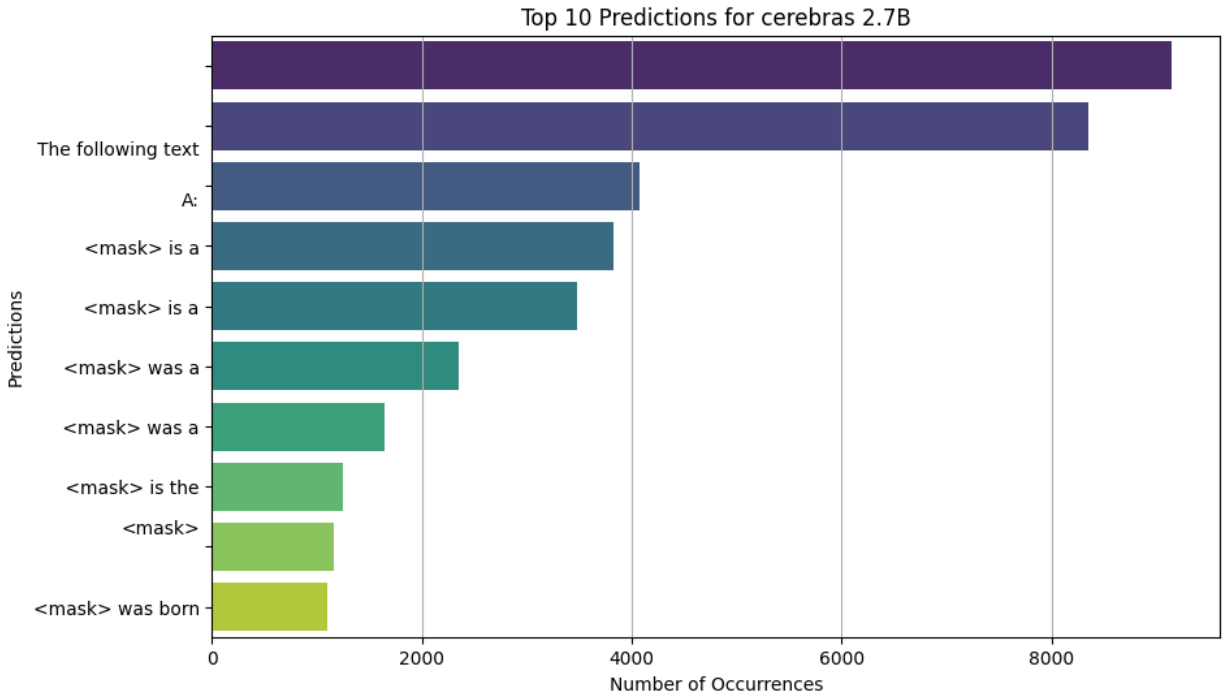


Figure 16: Most common predictions on Wikipedia for Cerebras-GPT 2.7B

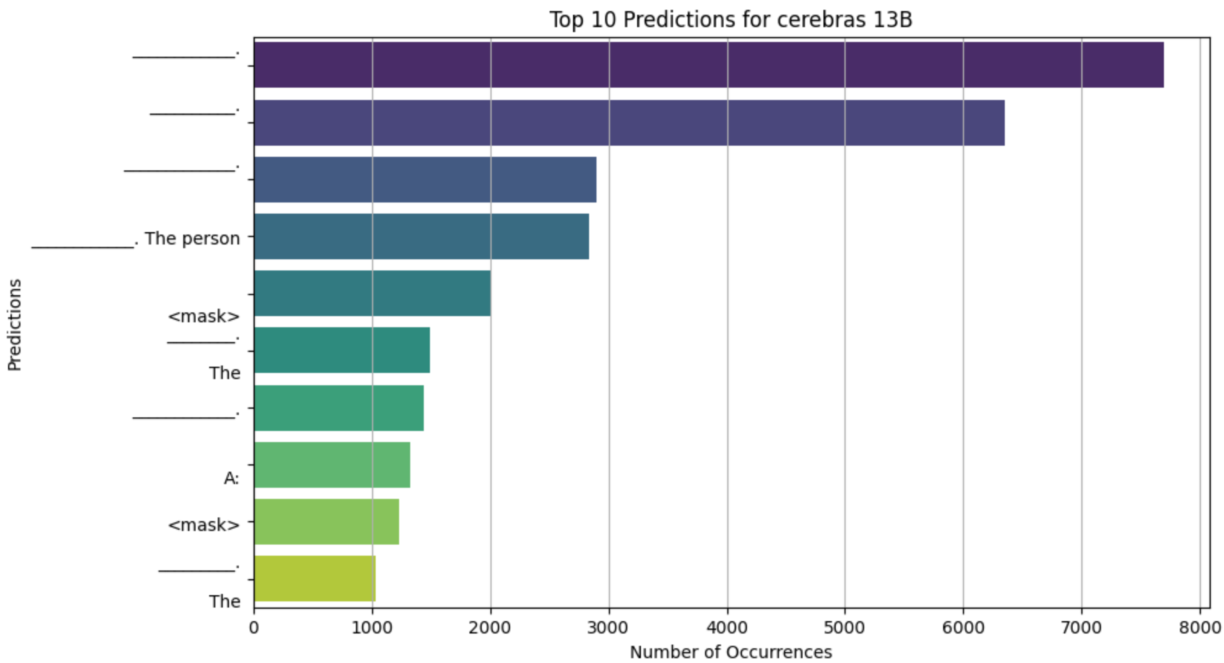


Figure 17: Most common predictions on Wikipedia for Cerebras-GPT 13B

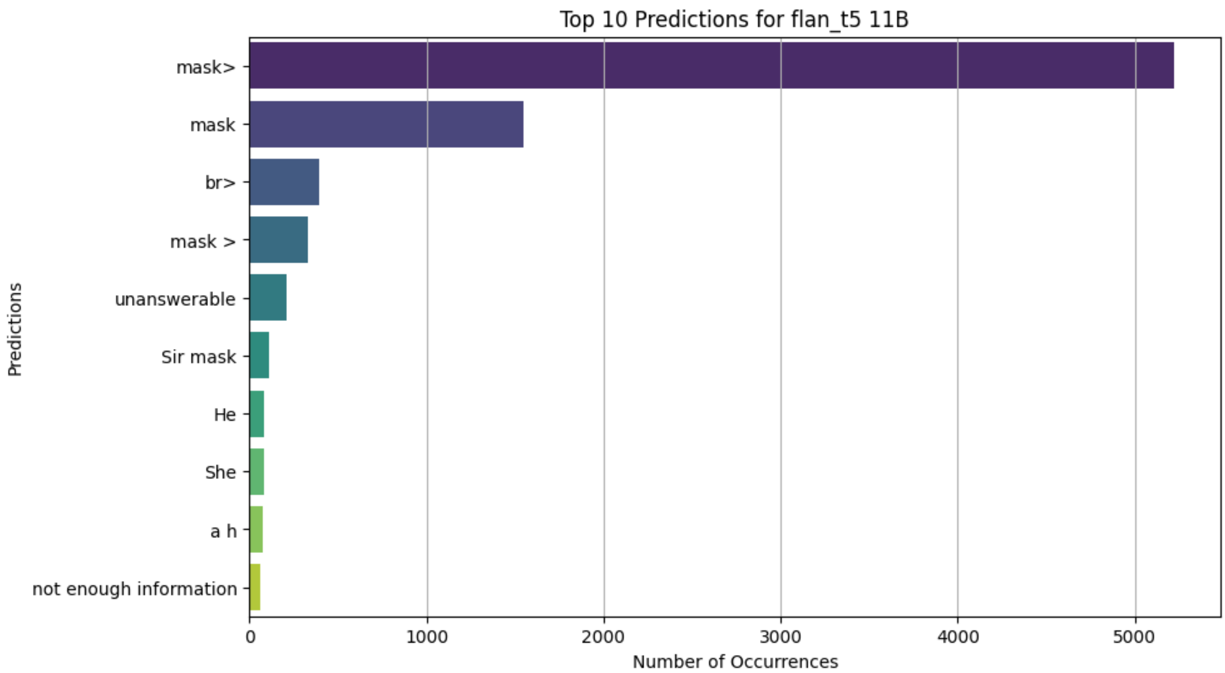


Figure 18: Most common predictions on Wikipedia for Flan_T5 11B

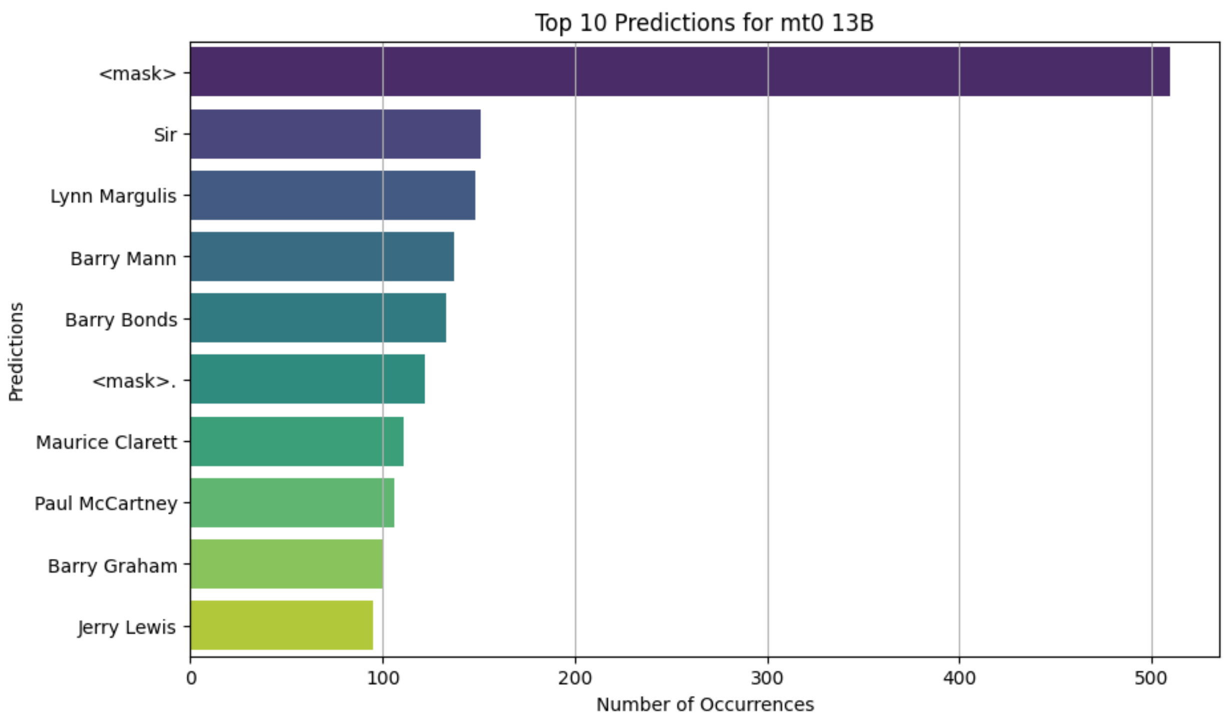


Figure 19: Most common predictions on Wikipedia for mT0 13B

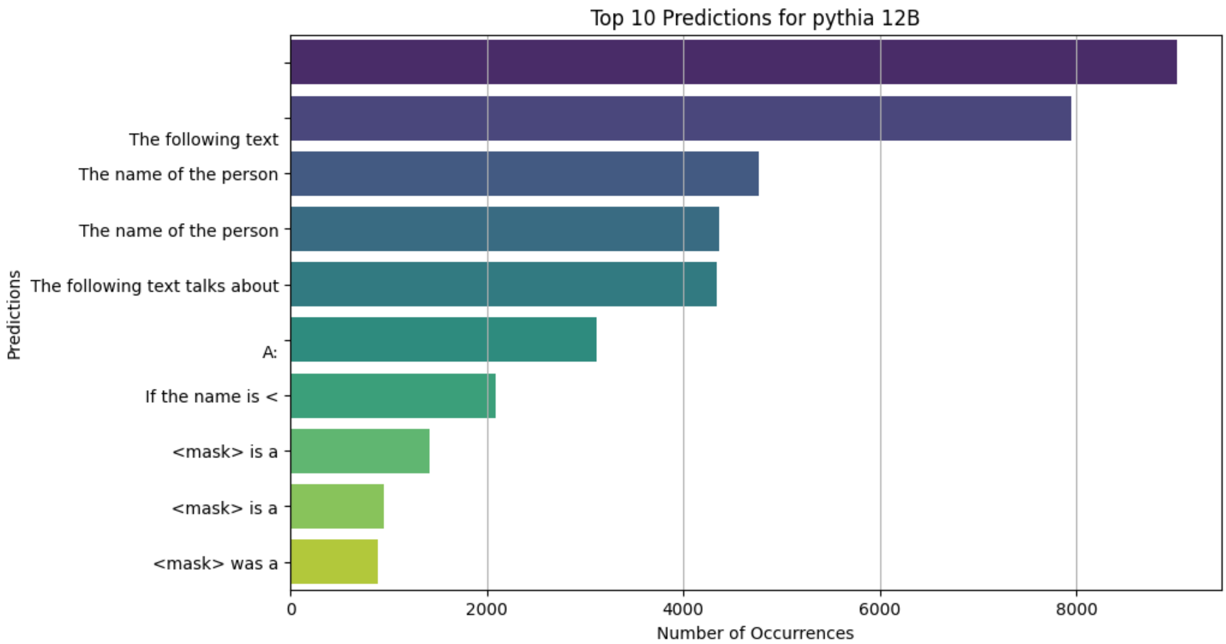


Figure 20: Most common predictions on Wikipedia for Pythia 12B

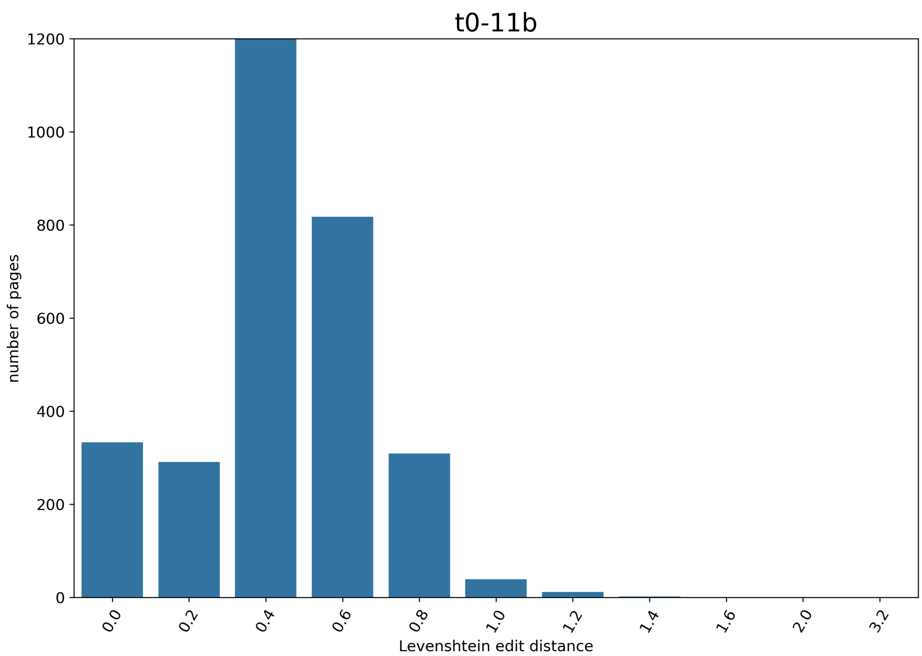


Figure 21: Normalized Levenshtein Distance distribution for T0 11B

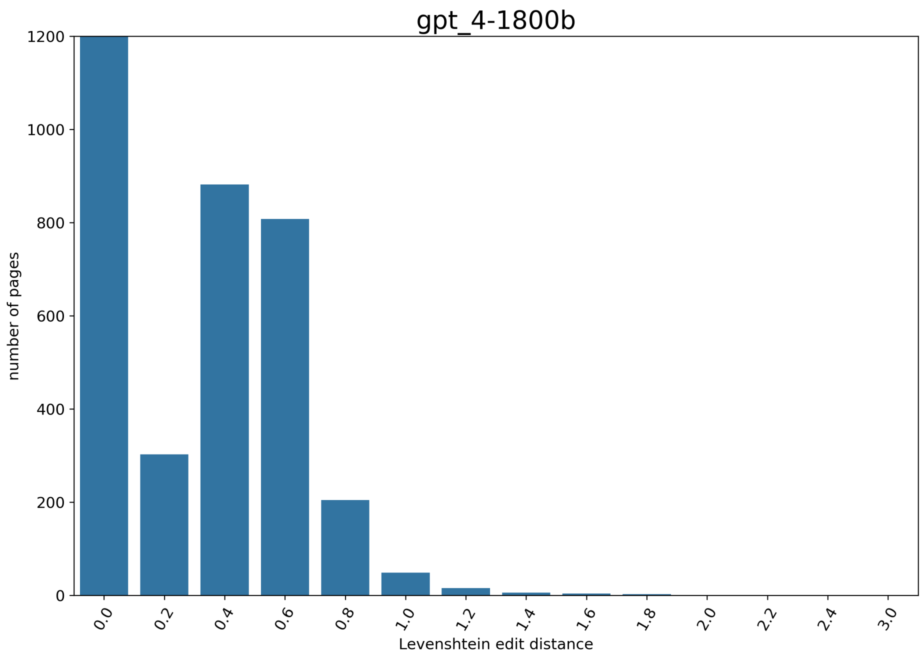


Figure 22: Normalized Levenshtein Distance distribution for GPT-4

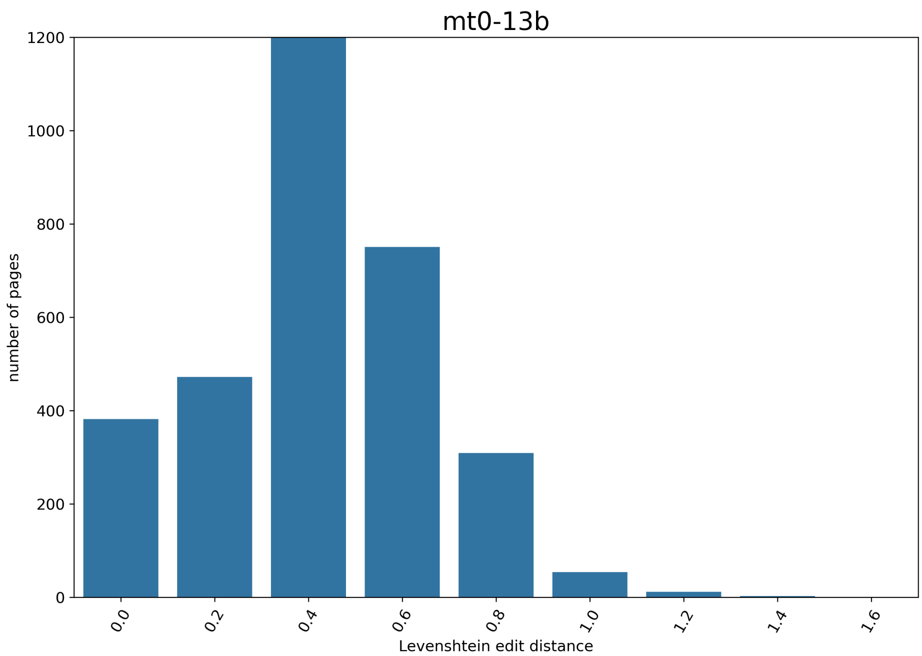


Figure 23: Normalized Levenshtein Distance distribution for mT0 13B

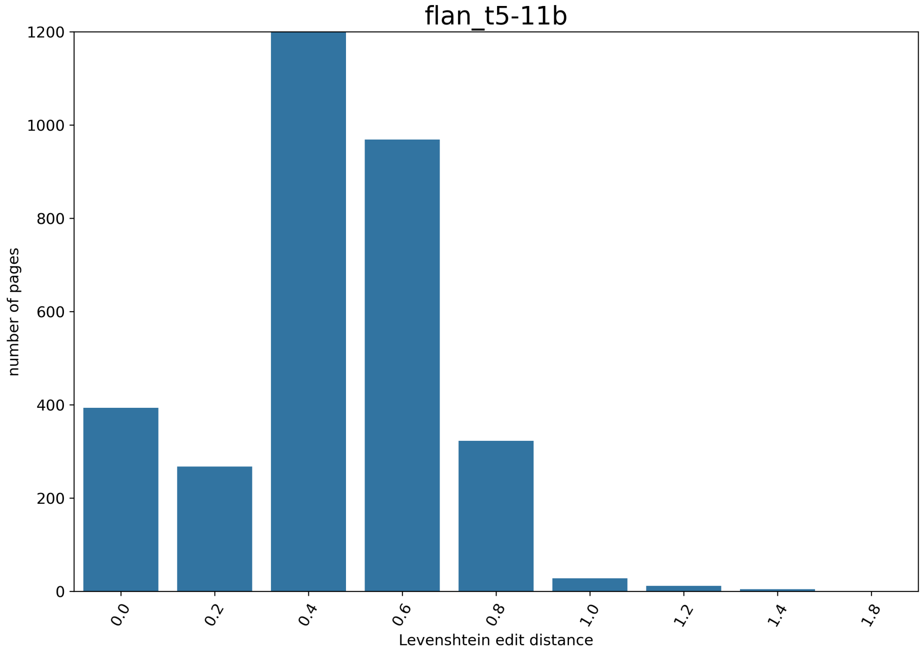


Figure 24: Normalized Levenshtein Distance distribution for T0 Flan_T5 11B

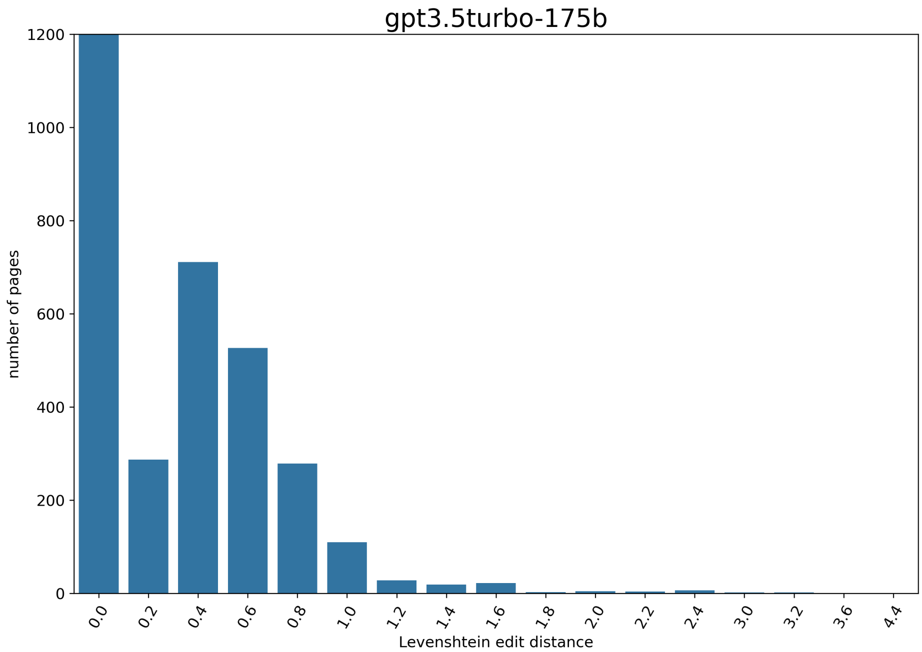


Figure 25: Normalized Levenshtein Distance distribution for GPT-3.5-turbo 175B

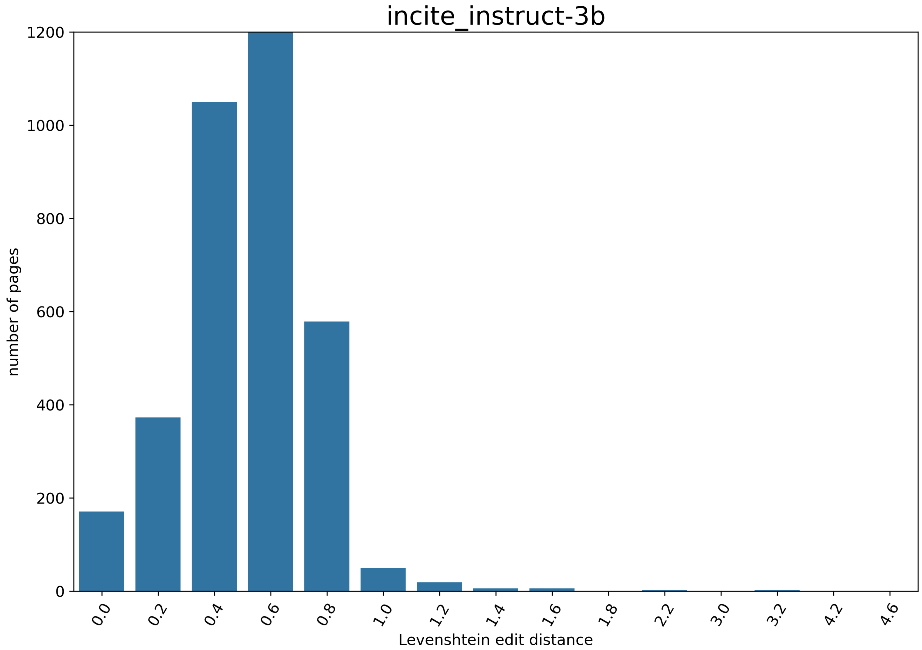


Figure 26: Normalized Levenshtein Distance distribution for INCITE-Instruct 3B

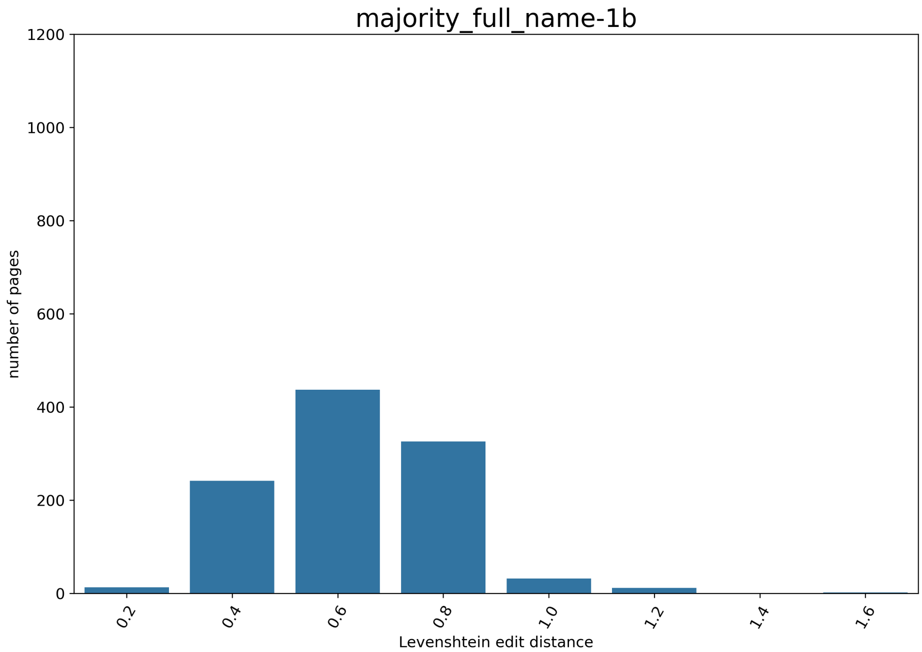


Figure 27: Normalized Levenshtein Distance distribution for Majority Name Baseline

LegalLens: Leveraging LLMs for Legal Violation Identification in Unstructured Text

Dor Bernsohn[1] **Ben Hagag**[1] **Yaron Vazana**[1] **Joel Niklaus**[2]

[1]**Darrow AI Ltd** [2]**Niklaus.ai**

Abstract

In this study, we focus on two main tasks, the first for detecting legal violations within unstructured textual data, and the second for associating these violations with potentially affected individuals. We constructed two datasets using Large Language Models (LLMs) which were subsequently validated by domain expert annotators. Both tasks were designed specifically for the context of class-action cases. The experimental design incorporated fine-tuning models from the BERT family and open-source LLMs, and conducting few-shot experiments using closed-source LLMs. Our results, with an F1-score of 62.69% (violation identification) and 81.02% (associating victims), show that our datasets and setups can be used for both tasks. Finally, we publicly release the datasets and the code used for the experiments in order to advance further research in the area of legal natural language processing (NLP).

1 Introduction

The widespread use of the internet has changed how information moves and connects in our society. Every day, the digital domain is flooded with a multitude of textual data, spanning from news articles and reviews to social media posts [1]. Within this sea of unstructured text, legal violations can often go unnoticed, concealed by the vast amount of surrounding information. These violations not only pose potential harm to individuals and entities but also challenge the very fabric of legal and ethical standards in the digital era. The significance of addressing these hidden violations cannot be overstated; as they have widespread implications for individual rights, societal norms, and the principles of justice. As a result, there is a pressing need to develop sophisticated methods to sift through the noise and identify these breaches.

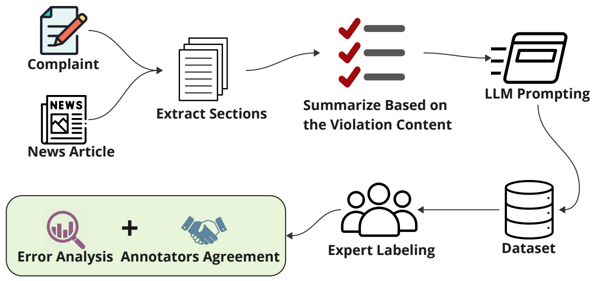


Figure 1: A visual representation of the data generation flow, illustrating the step-by-step process from raw input to the final synthesized dataset.

Legal violations often leave data trails. To detect these trails for pinpointing the violations, previous studies have often relied on specialized models tailored for specific domain applications (Silva et al., 2020; Yu et al., 2020). These models, while effective in their specific domains, lack the versatility needed to address the wide array of legal violations that can occur across different contexts.

Legal violation identification aims to automatically uncover legal violations from unstructured text sources and assign potential victims to these violations. We designed two setups, one for each task, the first for solving the legal violation identification task (a.k.a Identification Setup) using named entity recognition (NER), and the other for associating these violations with potentially affected individuals (a.k.a Resolution Setup) using natural language inference (NLI). Our dataset for the NER task is not limited to any specific domain, while the NLI dataset is focused on four common legal domains. Followed by recent research in the field of data generation (Leiker et al., 2023; Veselovsky et al., 2023; Hämäläinen et al., 2023), we chose to employ LLMs for synthetic data generation due to their ability to produce a large, diverse, and high-quality dataset that closely mimics the syntactic complexity of legal language, offering a scalable and ethically sound alternative to manual data craft-

[1] https://www.internetlivestats.com/total-number-of-websites

ing. We employed a thorough verification process to validate the data for both its realistic and complexity. Our approach involved automated data generation based on real-world event contexts, complemented by manual reviews conducted by seasoned legal annotators on the generated data.

Contributions

The contributions of this paper are three-fold:

- We introduce two dedicated datasets for legal violation identification, based on previous class action cases and legal news. These datasets, which include new legal entities, were generated using LLMs and validated by domain experts.

- We evaluate various language models, including BERT-based models and LLMs, across two different NLP tasks, offering valuable insights into their applicability and limitations in the context of legal NLP.

- We implement a two-setup approach employing both NER and NLI tasks, providing a methodology for legal violation detection and resolution.

Main Research Questions

We believe numerous violations exist in unstructured text. Our aim is to uncover these violations and link them to relevant prior class actions. This study focuses on the following key research questions:
RQ1: To what extent do our newly introduced datasets enhance the performance of language models in identifying legal violations within unstructured text and associate victims to them?
RQ2: How effectively do the language models adapt to new, unseen data for the purpose of identifying legal violations and correlating them with past resolved cases across different legal domains?
RQ3: What is the level of difference between machine-generated and human-generated text in the context of legal violation identification?

2 Related Work

Previous works in the field of legal violation identification mostly focused on domain-specific topics, encompassing areas such as compliance, data privacy, and industry-specific regulations. For instance, Amaral et al. (2023) evaluates data agreements for compliance with European privacy laws using NLP techniques. Silva et al. (2020) used NER to identify personal information in datasets, thereby uncovering instances of online data privacy breaches. Nyffenegger et al. (2023) used LLMs to attempt re-identification of anonymized persons from court decisions. Additionally, neural networks have been used to classify and annotate violation cases in specific industries like power supply (Yu et al., 2020). These studies, while valuable, have generally been limited to specific types of legal domains or particular sectors. Our work contributes to this existing body of research by introducing a dataset designed for broader applicability in identifying various types of legal violations.

Prior research has explored the use of Large Language Models (LLMs) for synthetic data generation (Rosenbaum et al., 2022a,b), beneficial in situations with scarce authentic data (Brown et al., 2020). In fact, training models on synthetic data led to improved outcomes in benchmarks like SQUAD1.1 (Puri et al., 2020). However, human-curated data often provides a richness that is hard to replicate (Møller et al., 2023; Ding et al., 2022). In this paper, we present a multi-step validation method to discern between real-world and machine-generated content, addressing the inherent limitations of relying solely on synthetic data.

Previous studies indicate that LLMs are capable of explaining legal terms present in legislative documents by drafting explanations of how previous courts explained the meaning of statutory terms(Savelka et al., 2023b). Moreover, the models demonstrated analytical depth in court decision analysis, rivaling seasoned law students(Savelka et al., 2023a). In this study, we created a dataset based on a previous lawsuits legislation background, rather than examining existing records.

While LLMs (Radford et al., 2019) have been employed to enhance datasets for event detection tasks (Veyseh et al., 2021), our methodology advances this by generating pairs of specific violations and their corresponding events, using data from previously settled lawsuits. Unlike Koreeda and Manning (2021), who concentrated on NLI in the context of legal contracts, our research introduces an NLI dataset based on class-action cases. Additionally, NER has been increasingly applied in the legal domain, including efforts to extract entities from Indian court judgments (Kalamkar et al., 2022) and other legal texts (Luz de Araujo et al., 2018; Angelidis et al., 2018; Leitner et al., 2019). Despite these advancements, existing research has largely focused on a standard set of entity types, such as parties (plaintiff and defendant), judges, court name and law/citation. Our work introduces a new set of entity types that have not been previously explored in legal NER research (Păiș et al., 2021; Luz de Araujo et al., 2018; Dozier et al., 2010; Leitner et al., 2020; Skylaki et al., 2020; Kalamkar et al., 2022), thereby expanding the scope and applicability of NER in legal contexts.

3 Curating Custom Legal Datasets: A Multi-stage Approach to NER and NLI Tasks

Existing datasets may not adequately address the diverse range of legal violations and contexts central to our study, which is not in specific areas. To overcome these challenges, we employed a systematic and carefully planned data generation process, consisting of three stages: prompting, labeling, and data validation. This approach aimed at creating two robust datasets for two NLP tasks in the legal domain. We chose to focus on two key tasks:

- NER (classifying tokens into predefined entities) for identifying violations. NER has been employed to define novel legal entities, enabling precise localization of pertinent information necessary for the extraction of legitimate legal violations, as detailed in Table 4 in Appendix C.

- NLI (classifying a hypothesis and a premise into entailed/contradict/neutral) for matching these violations with known, resolved class-action cases. NLI facilitates the correlation of multiple unstructured text associated with the same violation, thereby enabling the matching of extracted violations identified by the NER task with pre-existing legal complaints of class action cases.

This dual-setup approach was designed to mimic the process of legal violation detection and resolution, generating high-quality data that closely resembles real-world scenarios.

Based on recent research in prompt-based methods (Liu et al., 2023), our study employs prompts for a variety of reasons. LLMs have been shown to adapt to specialized tasks through techniques like instruction tuning (Wei et al., 2021), reinforcement learning from human feedback (Ouyang et al., 2022), and in-context learning (Brown et al., 2020) when prompted with natural language instructions. Prompts facilitate task-specific optimization, a quality emphasized by DialogPrompt (Gu et al., 2021), which aligns with our focus on NER and NLI in the legal domain by fine-tuning on the generated dataset. Additionally, the sensitivity of prompts in context, as demonstrated in Time-aware Prompts in Text Generation (Cao and Wang, 2022), is crucial for understanding specific legal contexts like resolved class-action cases. As a result, our methodology leverages a prompt-based approach, optimized for the legal domain, to generate high-quality data for NER and NLI tasks.

3.1 NER Data Generation

NER can be framed as a token classification task, wherein, the objective is to classify each word in a sentence as an entity class. In our dataset, there are four such entities; *Law*, *Violation*, *Violated By*, and *Violated On*.

For the NER task, our foundational data source was class action complaints, as described in (Semo et al., 2022). A complaint, often referred to as a plaintiff's plea, is a formal legal document that initiates a lawsuit. It outlines the complaints of the plaintiff and specifies the relief sought from the court. From each of these complaints, we extracted relevant sections such as *allegations*, *counts*, and *legal arguments* that were pertinent to our study, ensuring relevance and precision. These sections encapsulate the main context of the alleged violations. They were subsequently summarized through the utilization of GPT-4 (OpenAI, 2023) to capture the core essence of the violation content, and were employed as the context in the subsequent prompts.

For a visual representation of our data generation process, refer to Figure 1.

Prompt

For the NER task, we devised two unique prompting strategies: explicit and implicit. The explicit method not only emphasizes the inclusion of multiple distinct entities but also underscores the specific order of their appearance, adding a layer of complexity and structure to the generated content (refer to Appendix 6). This approach ensures that the content is not only diverse but also adheres to certain structural guidelines, which contain task descriptions, specific instructions, and few-shot examples. Conversely, the implicit strategy focuses solely on a singular entity, specifically the content that describes the violation, refer to Appendix 6.

Furthermore, both strategies incorporate additional parameters such as the cause of action, industry, and context. The inclusion of these parameters refines the generated content, tailoring it to specific scenarios and ensuring its relevance to the desired domain. By employing the explicit approach, we capture the comprehensive nature of a scenario, whereas the implicit method provides a concise perspective on one specific aspect.

3.2 NLI Data Generation

NLI can be framed as a classification task, wherein, the objective is to compare a premise to a hypothesis, and predict one of the three classes: (1) *Entailment* - where the hypothesis is contained and can be supported by the premise, (2) *Contradiction* - when the hypothesis contradicts the premise, (3) *Neutral* - when the premise neither entails nor contradicts the hypothesis.

For the NLI task, our data source consisted articles taken from a legal news website. Each news article was first summarized, by prompting GPT-4 (OpenAI, 2023), to capture its legal grounds. By summarizing, we ensured that the data was concise yet comprehensive by keeping only the legal violation section and removing background parts. This summarized content served as the premise. Using this premise, the model was tasked to generate a hypothesis that mimicked real-world scenarios. The intention behind this design was to create diverse records that spanned various legal areas. Table 5 in Appendix C presents the NLI data distributions.

Prompt

In this setup, we aimed to create scenarios that mirror real-life accounts of potential violations. We generated texts that mimic common situations where individuals share concerns, like online reviews or social media posts. The goal was to produce narratives that implicitly describe the effects of a violation. We added variations in attributes such as the writers age and gender and the text format to capture a wide range of experiences.

4 Human Expert Annotations

Data validation holds particular importance in our study due to the synthetic nature of the dataset. To ensure that the dataset is both realistic and challenging, we have implemented several validation methods. Experienced annotators were provided with a list of textual input along with the models predictions for both the NER and NLI tasks. Their primary task was to verify the validity of each prediction and suggest on additional entities if they existed. All annotators were given the same instructions, but the data they received was shuffled to ensure unbiased validation. Their insights were crucial in pinpointing any discrepancies, unclear areas, or possible inaccuracies in the dataset. Figure 4 in Appendix B presents a screenshot of the annotation platform we used.

Upon further examination of our data, a comparison between machine-generated and human-authored content revealed significant similarities. This comparison involved analyzing various linguistic and structural features of the texts. Both displayed identical average sentence lengths. Moreover, there was not significant difference between the character count between the generated content and the human-authored text. Additionally, when comparing the POS tags between the real text and the generated text, by averaging the total counts of each tag occurrences, the average difference was found to be 26% and the median was 16%.

A key part of our validation process was the classification task. In this task, three independent annotators had to distinguish between machine-generated and human-written records, a challenge also noted in recent research (Mitchell et al., 2023; Kirchenbauer et al., 2023). Our annotators' goal was to label each record based on its origin: machine-generated or human-written. The annotators achieved an average F1-score of 44.86%. However, their Cohen's Kappa scores, which were 0.0821, 0.2149, and 0.0988, showed only minor agreement among them. This low level of agreement, as indicated by Cohen's Kappa scores, points out the complexity of the task. It also suggests that our machine-generated content closely resembled human writing, making it difficult even for experts to tell them apart. The use of Cohen's Kappa in our study is supported by its well-known effectiveness in binary classification tasks, especially in data annotation scenarios (Wang et al., 2019).

5 Experiments

In this section, we explore several methods to tackle the challenging and realistic setups that we

Table 1: Comparison of different methodologies for NER. The table showcases various models, their sizes, and the method employed, along with their performance metrics.

Model	Size	Method	F1	Precision	Recall
nlpaueb/legal-bert-small-uncased	35M	Fine-tune	$48.90_{\pm 0.39}$	$41.92_{\pm 0.80}$	$58.69_{\pm 0.52}$
distilbert-base-uncased	66M	Fine-tune	$49.71_{\pm 0.83}$	$42.19_{\pm 0.89}$	$60.50_{\pm 0.77}$
bert-base-cased	108M	Fine-tune	$54.80_{\pm 0.64}$	$47.23_{\pm 1.06}$	$65.28_{\pm 1.01}$
bert-base-uncased	109M	Fine-tune	$53.22_{\pm 1.42}$	$45.86_{\pm 1.68}$	$63.42_{\pm 1.11}$
roberta-base	125M	Fine-tune	$\mathbf{62.69_{\pm 0.69}}$	$56.58_{\pm 1.12}$	$\mathbf{70.30_{\pm 0.73}}$
nlpaueb/legal-bert-base-uncased	109M	Fine-tune	$57.50_{\pm 0.94}$	$50.34_{\pm 1.26}$	$67.04_{\pm 0.71}$
lexlms/legal-roberta-base	124M	Fine-tune	$59.73_{\pm 2.03}$	$53.11_{\pm 2.27}$	$68.25_{\pm 1.86}$
joelito-legal-english-roberta-base	124M	Fine-tune	$59.01_{\pm 1.74}$	$52.52_{\pm 2.52}$	$67.40_{\pm 0.85}$
lexlms/legal-longformer-base	148M	Fine-tune	$62.30_{\pm 1.76}$	$\mathbf{56.78_{\pm 2.14}}$	$69.04_{\pm 1.32}$
lexlms/legal-roberta-large	355M	Fine-tune	$50.23_{\pm 28.1}$	$46.07_{\pm 25.8}$	$55.22_{\pm 30.8}$
lexlms/legal-longformer-large	434M	Fine-tune	$37.63_{\pm 34.4}$	$34.26_{\pm 31.3}$	$41.76_{\pm 38.1}$
joelito-legal-english-roberta-large	355M	Fine-tune	$58.92_{\pm 4.28}$	$52.88_{\pm 4.95}$	$66.59_{\pm 3.22}$
Falcon	7B	QLoRA	$1.00_{\pm 0.50}$	$39.50_{\pm 16.8}$	$0.50_{\pm 0.20}$
Llama-2	7B	QLoRA	$16.3_{\pm 4.10}$	$34.10_{\pm 11.1}$	$11.20_{\pm 2.60}$
OpenAI GPT-3.5	175B	Few-shot	$2.77_{\pm 0.12}$	$1.78_{\pm 0.08}$	$6.23_{\pm 0.29}$
OpenAI GPT-4	-	Few-shot	$13.55_{\pm 0.54}$	$8.29_{\pm 0.37}$	$37.1_{\pm 0.99}$

Table 2: Entity-specific F1 score for the best-performing NER model, ‘roberta-base‘.

LAW	VIOLATION	VIOLATED BY	VIOLATED ON
$77.57_{\pm 1.35}$	$59.06_{\pm 0.55}$	$76.88_{\pm 2.06}$	$62.83_{\pm 2.57}$

created. More precisely, we analyzed the performance of language models on these setups by conducting three sets of experiments. (1) We evaluated models that are inspired by the BERT architecture through the process of fine-tuning (Sun et al., 2020). (2) We explored LLMs such as Falcon-7B, Llama-2-7B and Llama-2-13B through the process of parameter efficient fine-tuning (Houlsby et al., 2019; Hu et al., 2021). (3) Thanks to their out-of-the-box generalization capabilities, we assessed OpenAI's GPT-3.5 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) models.

5.1 Setup

NLI Our dataset contains news articles across four legal domains. Given the similarities in the legal merits between these domains, there is a potential risk of data leakage related to the legal attributes of the cases. To address this issue, we employed a leave-one-out approach. In this method, we tested each legal domain separately while training the model on the other domains.

NER Our dataset is categorized by Cause of Action (CoA). CoA refers to a set of facts or legal reasons that justify the right to sue or seek legal remedy in a court of law. Due to the potential overlap and similarities between different CoAs, there's a risk of data leakage when training models. To mitigate this, we adopted a strategy where CoAs present in the training set were excluded from the test set. This ensures that the model is evaluated on entirely distinct CoAs, preventing any inadvertent training on test data.

5.2 Model Classes

BERT Models In this setting, we assess the effectiveness of transformer-based language models (Vaswani et al., 2017). We fine-tuned RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019) and BERT (Devlin et al., 2018) models. Additionally, we evaluated their legal counterparts, i.e., Legal-BERT (Chalkidis et al., 2020) and Legal-RoBERTa (Chalkidis* et al., 2023). Furthermore, we evaluated models (Mamakas et al., 2022) based on the Longformer architecture (Beltagy et al., 2020). Following this, we also assessed the Legal-English-RoBERTa models, which are specialized versions tailored for legal English (Niklaus et al., 2023). We utilized the AutoModelForTokenClassification class from the HuggingFace Transformers library to train the models. Each model was trained for 10 epochs with an initial learning rate of $2e-5$. In addition, we used early-stopping to prevent overfitting.

Open-Source LLMs In this setting, we evaluated Falcon (Almazrouei et al., 2023) and Llama2s (Touvron et al., 2023) performance. More precisely, we considered the 7 billion parametric version of Falcon, and 7 and 13 billion versions of Llama2. Following the success of Parameter Efficient Fine-Tuning methodologies for fine-tuning LLMs, we leveraged QLoRA (Dettmers et al., 2023) due to its superior performance over other methods. Figure 7 shows the prompt that we designed to guide the tuning process.

The prompt has two parts: Input and Output. The Input contains the sentence on which NER and NLI have to be performed. The Output contains the format in which the LLM has to predict the entities contained in the sentence. It is important to note that during inference, we prompt the model to generate the required output by only including the Input section.

We employed HuggingFace's AutoModelForCausalLM class for fine-tuning, available under an Apache-2.0 license[2]. Each model underwent training for 20 epochs, a learning rate of 2e-4, a QLoRA rank of 64, and a dropout rate of 0.25.

Closed-Source LLMs We evaluate OpenAI's GPT-4 (OpenAI, 2023) and OpenAI's GPT-3.5 (Brown et al., 2020) models for few-shot NER and NLI without any fine-tuning, using the matching production models of August 2023. We use the Langchain[3] client, available under an Apache-2.0 license, with few-shot prompts, as demonstrated in Figure 8. In all experiments, we set the temperature to 0.7 and used 9 random samples from the training dataset as few-shot examples. We employed the same prompts as those used for open-source models and the same evaluation mechanism. Each API call was repeated five times.

6 Results

6.1 NER

Table 1 presents the performance metrics of various models. Interestingly, BERT-based models with fewer parameters outperform LLMs by a significant margin. This disparity in performance is due to the difference in objective functions that the different model classes use. BERT-based models employ the cross-entropy objective function per token, providing a stronger gradient signal. Furthermore, the label space is well constrained by the number of possible entities in our data set. On the other hand, LLMs have been fine-tuned via causal language modeling, wherein the task is to learn the joint probability distribution of all tokens by maximizing the likelihood of the data. The gradient signal in the case of fine-tuning LLMs is not as fine-grained as cross-entropy. This is because the label space, i.e., the number of possibilities to predict the next token from, far exceeds the number of required entities.

Across BERT-based models, we notice interesting trends. First, *roberta-base* model attains the best performances, achieving an F1 score of 62.69% and Recall of 70.3%. Second, the performance across all metrics improved as model complexity grew, except for Longformer-based models and joelito-legal-english-roberta-based models.

Focusing on LLMs, we observed that both open-source and close-source models perform poorly on this task. Closer analysis of predictions indicated incorrect B-token prediction in generated text. These errors were propagated to the next predictions, causing the LLMs to misclassify the tokens and place them into incorrect entities.

6.2 NLI

Table 3 shows domain-specific performances across all model classes. In contrary to trends discovered in the NER experiments, in NLI we noticed that LLMs outperform BERT-based models by a very significant margin. Unlike NER, in NLI, LLMs are fine-tuned to predict only one token, i.e., either of *entailed*, *contradict*, and *neutral*. Additionally, the NLI task had only 312 samples, and LLMs learn relatively better in low data situations and generalize well to out-of-distribution (OOD) test data sets (Brown et al., 2020).

Except for domain *Wage*, *Falcon 7B* achieved the highest performance across domains (*Consumer Protection*, *Privacy*, and *TCPA*). *Falcon 7B* attained the highest Macro F1 metric, demonstrating its OOD capabilities. Among BERT-based models, *roberta-base* once again achieved the best performance, similar to NER tasks.

7 Error Analysis

To improve our models and enrich our understanding, we conducted a thorough error analysis of top-performing models across tasks. This analysis identifies their limitations, providing a clear roadmap for future refinements.

7.1 NER

In evaluating our NER model, the entity type "VIOLATION" exhibited the lowest F1 score. This entity is often lengthy and contextually complex, making it a challenging target for accurate identification. We conducted an error analysis on a subset

[2] https://github.com/huggingface/transformers
[3] https://github.com/langchain-ai/langchain

Table 3: Macro F1 evaluation of various model architectures for the NLI task across different legal entities.

Model	Consumer Protection	Privacy	TCPA	Wage
nlpaueb-legal-bert-small-uncased	$60.8_{\pm 7.1}$	$49.6_{\pm 14.}$	$47.6_{\pm 11.}$	$56.7_{\pm 6.0}$
distilbert-base-uncased	$79.8_{\pm 2.0}$	$53.9_{\pm 13.}$	$72.1_{\pm 9.3}$	$71.2_{\pm 7.3}$
bert-base-cased	$65.5_{\pm 9.2}$	$39.9_{\pm 18.}$	$58.9_{\pm 16.}$	$65.5_{\pm 13.}$
bert-base-uncased	$69.3_{\pm 7.7}$	$36.3_{\pm 16.}$	$69.5_{\pm 7.2}$	$64.0_{\pm 16.}$
roberta-base	$82.9_{\pm 4.5}$	$62.0_{\pm 5.0}$	$69.5_{\pm 31.}$	$69.7_{\pm 29.}$
lexlms-legal-roberta-base	$45.8_{\pm 5.8}$	$27.3_{\pm 7.9}$	$48.6_{\pm 14.}$	$44.4_{\pm 19.}$
joelito-legal-english-roberta-base	$61.6_{\pm 14.2}$	$33.1_{\pm 12.2}$	$55.8_{\pm 9.95}$	$48.6_{\pm 17.9}$
lexlms-legal-longformer-base	$58.3_{\pm 16.}$	$27.8_{\pm 4.6}$	$54.8_{\pm 11.}$	$54.5_{\pm 11.}$
lexlms-legal-roberta-large	$18.1_{\pm 0.7}$	$20.2_{\pm 8.1}$	$15.3_{\pm 1.8}$	$16.6_{\pm 0.0}$
lexlms-legal-longformer-large	$19.2_{\pm 1.3}$	$17.5_{\pm 0.6}$	$25.5_{\pm 24.}$	$26.3_{\pm 21.}$
joelito-legal-english-roberta-large	$16.4_{\pm 3.3}$	$20.2_{\pm 5.8}$	$47.3_{\pm 30.3}$	$27.3_{\pm 23.9}$
Falcon 7B	$\mathbf{87.2}_{\pm \mathbf{3.1}}$	$\mathbf{84.5}_{\pm \mathbf{8.8}}$	$\mathbf{83.9}_{\pm \mathbf{0.9}}$	$68.5\ _{\pm 11.}$
Llama-2 7B	$47.2_{\pm 5.9}$	$47.8_{\pm 10.}$	$63.5_{\pm 7.3}$	$63.7_{\pm 14.}$
Llama-2 13B	$63.1_{\pm 8.0}$	$75.2_{\pm 6.5}$	$63.9_{\pm 10.}$	$\mathbf{86.5}_{\pm \mathbf{5.6}}$
OpenAI GPT-3.5	$17.8_{\pm 2.6}$	$18.12_{\pm 3.1}$	$15.09_{\pm 1.9}$	$12.91_{\pm 5.4}$
OpenAI GPT-4	$49.83_{\pm 19.}$	$48.44_{\pm 9.4}$	$37.04_{\pm 7.4}$	$52.48_{\pm 11.6}$

of hard cases to understand the model's limitations.

The errors fall into three categories: truncation errors, context misunderstanding, and incorrect entity identification. For instance, in the sentence *"I've been getting these [VIOLATION] constant calls on my cell phone from some company that won't quit [VIOLATION]."*, the model predicted *"constant calls on"* instead of the actual entity. This truncation error suggests the model captures only the initial segment but fails to include the entire scope. In another example, *"They've been [VIOLATION] failing to disclose that their educational programs were underperforming [VIOLATION]."*, the model predicted *"disclose"*, indicating a context misunderstanding. Notably, when the model completely misses the target, it often predicts a much shorter entity, suggesting a bias towards shorter answers when uncertain.

The model struggles with the "VIOLATION" entity type, particularly with longer and more complex entities. Fine-tuning the model with a diverse, context-rich training set could improve its performance. Existing literature also suggests that NER models often struggle with complex entities (Dai, 2018), underscoring the need for continued research in this area.

7.2 NLI

In the error analysis of our best performing NLI model, Falcon 7B, we consolidated the model errors across different legal domains to form a comprehensive view. Our focus was on two types of classification errors: first-class errors, which involve confusions between "Contradict" and "Entailed", and second-class errors, which are misclassifications of "Contradict" or "Entailed" as "Neutral". Figure 2 shows that while Falcon 7B performs well in avoiding first-class errors, it exhibits a substantial number of second-class errors. The high rate of such errors indicates that the model finds it challenging to handle more nuanced cases where it is difficult to discern whether the person was affected by the violation or not.

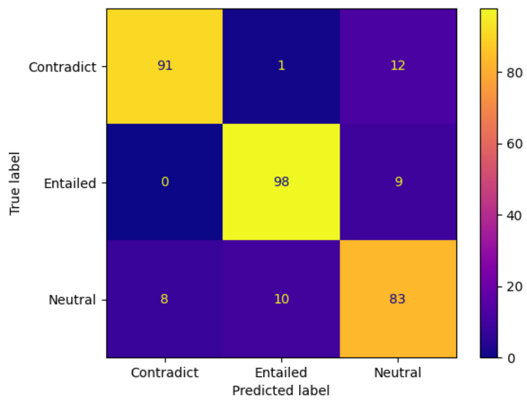


Figure 2: NLI Confusion Matrix derived from the top performer model (Falcon 7B's) predictions.

Although Falcon 7B outperforms other models in this task, it strugglesin accurately classifying statements related to wage areas. This could be attributed to the complexities and ambiguities of the wage norms, which make it challenging to clearly determine whether a wage violation has occurred. Therefore, investigating different token lengths to provide more context or fine-tuning the model to better navigate these intricate wage scenarios could be valuable directions for future work.

8 Conclusions and Future Work

8.1 Answers to the Research Questions

RQ1: *To what extent do our newly introduced datasets enhance the performance of language models in identifying legal violations within unstructured text and associate victims to them?* The study introduced new entities in the datasets. This addition improved the ability of language models to identify legal violations in unstructured text. With these new entities, the roberta-base model achieved an F1-score of 62.69% in identifying violations and 81.02% (Falcon 7B model) in linking them to victims. This demonstrates that our new approach, which focuses on identifying and associating violations to victims, has been successful, yet there remains potential for further refinements and improvements.

RQ2: *How effectively do the language models adapt to new, unseen data for the purpose of identifying legal violations and correlating them with past resolved cases across different legal domains?* Our experiments assessed language models' adaptability to unseen data, especially in the context of identifying legal violations and correlating them with past resolved cases across different legal domains. While BERT-based models demonstrated strong performance in certain tasks, LLMs like Falcon-7B excelled in low-data scenarios, particularly in associating violations with resolved cases. This suggests that these models effectively adapt to new data, especially when the data is limited.

RQ3: *What is the level of difference between machine-generated and human-generated text in the context of legal violation identification?* Our validation process involved a comparison between machine-generated and human-authored content. The findings revealed that the two types of content were strikingly similar in terms of average sentence lengths and character count. When expert annotators were tasked to distinguish between machine-generated and human-written records, they achieved an average F1-score of 44.86%. The low level of agreement among the annotators indicates that our machine-generated content closely resembles human writing, making it challenging even for experts to differentiate between the two.

8.2 Conclusion

In this study, by leveraging LLMs and expert validation, we introduced a dual setup approach to identify legal violations from text. Our approach uses (1) NER to pinpoint violations, resulting in an F1-score of 62. 69% and (2) NLI to associate these violations with resolved cases, resulting in an F1-score of 81.02%. We created two specialized datasets to advance research in this field.

8.3 Future Work

Expanding Legal Areas In future iterations, we aim to expand the dataset to include a broader range of legal areas. By incorporating diverse legal texts, we hope to create a more representative dataset for legal violation identification.

Incorporating Multiple Jurisdictions While our current dataset is heavily focused on common law in US courts, future work will aim to integrate legal texts from various global jurisdictions, including civil law systems. This will not only enhance the datasets diversity but also improve the robustness and applicability of models trained on it.

Fact Matching An avenue for future work is the integration of fact matching. Developing algorithms for cross-referencing facts across sources can enhance the accuracy of legal violation identification, especially when a single source might not provide a complete picture. (Thorne et al., 2018; Jiang et al., 2020)

Limitations

Focus on Common Law in US Courts A primary limitation of our dataset is its focus on US common law. While this deepens understanding of US legal principles and precedents, it may not apply to civil law jurisdictions or non-US legal systems. The nuances, interpretations, and applications of laws can vary significantly across different jurisdictions, and our dataset, being US-centric, might not capture these variations adequately.

Coverage of Areas of Law While our dataset provides a comprehensive overview of legal violations from various text sources, it does have its limitations in terms of the breadth of legal areas covered. The current dataset predominantly focuses on specific areas of law, potentially overlooking nuances and intricacies of other legal domains. For instance, while we have extensively covered areas like consumer protection and privacy, other equally significant areas such as intellectual property, environmental law, or international law might not have been represented with the same depth.

Ethics Statement

The primary objective of this research is to revolutionize the identification and understanding of legal violations within the sprawling landscape of online text. By introducing a novel dataset specifically tailored for Named Entity Recognition (NER) and Natural Language Inference (NLI) tasks in the legal context, we aim to significantly advance the field of Natural Language Processing (NLP) and its applications in law. Our research holds the potential to greatly assist legal professionals in efficiently identifying and addressing legal violations, thereby contributing to a safer and more equitable digital society.

In the pursuit of this objective, we have employed LLMs, specifically GPT-4 (OpenAI, 2023), for data generation, and have subjected the generated data to rigorous validation by expert annotators. This dual-layered approach ensures the quality and reliability of our dataset, while also providing a comprehensive range of examples that can be generalized across various domains.

However, we acknowledge that the deployment of machine learning models in the legal domain is fraught with ethical considerations (Tsarapatsanis and Aletras, 2021). Automating the detection of legal violations could inadvertently lead to false positives or negatives, with serious implications for individual rights and the rule of law. Therefore, we stress that our technology is intended to serve as a supplementary tool for legal professionals, rather than a replacement. It is essential that any application of our dataset and subsequent models be conducted responsibly with a thorough understanding of the limitations and biases that may be inherent in automated systems.

Moreover, we recognize the ethical imperative of data privacy and confidentiality, especially given the sensitive nature of legal texts. All data used in this research have been anonymized and stripped of personally identifiable information to the best of our ability, in compliance with relevant data protection regulations. All the data utilized in this study is sourced from publicly accessible online platforms and does not infringe on any individuals or entities proprietary rights.

Acknowledgements

We thank all the anonymous reviewers for their insightful comments. We thank the three company employees for the annotation of the dataset.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Orlando Amaral, Muhammad Ilyas Azeem, Sallam Abualhaija, and Lionel C Briand. 2023. Nlp-based automated compliance checking of data processing agreements against gdpr. *IEEE Transactions on Software Engineering*.

Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named entity recognition, linking and generation for greek legislation. In *JURIX*, pages 1–10.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shuyang Cao and Lu Wang. 2022. Time-aware prompting for text generation. *arXiv preprint arXiv:2211.02162*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: the muppets straight out of law school. *CoRR*, abs/2010.02559.

Ilias Chalkidis*, Nicolas Garneau*, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. LeXFiles and LegalLAMA: Facilitating English Multinational Legal Language Model Development. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada. Association for Computational Linguistics.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. 2010. *Named entity recognition and resolution in legal text*. Springer.

Xiaodong Gu, Kang Min Yoo, and Sang-Woo Lee. 2021. Response generation with context-aware prompt learning. *arXiv preprint arXiv:2111.02643*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in indian court judgments. *arXiv preprint arXiv:2211.03442*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. *arXiv preprint arXiv:2301.10226*.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Daniel Leiker, Sara Finnigan, Ashley Ricker Gyllen, and Mutlu Cukurova. 2023. Prototyping the use of large language models (llms) for adult learning content creation at scale. *arXiv preprint arXiv:2306.01815*.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287. Springer.

Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. 2020. A dataset of german legal documents for named entity recognition. *arXiv preprint arXiv:2003.13016*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Pedro Henrique Luz de Araujo, Teófilo E de Campos, Renato RR de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. Lener-br: a dataset for named entity recognition in brazilian legal text. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 313–323. Springer.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2023. Is a prompt and a few samples all you need? using gpt-4 for data augmentation in low-resource classification tasks. *arXiv preprint arXiv:2304.13861*.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel E Ho. 2023. Multilegalpile: A 689gb multilingual legal corpus. *arXiv preprint arXiv:2306.02069*.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2023. Anonymity at risk? assessing re-identification capabilities of large language models.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Amir Saffari, Macro Damonte, and Isabel Groves. 2022a. Clasp: Few-shot cross-lingual data augmentation for semantic parsing. *arXiv preprint arXiv:2210.07074*.

Andy Rosenbaum, Saleh Soltan, Wael Hamza, Yannick Versley, and Markus Boese. 2022b. Linguist: Language model instruction tuning to generate annotated utterances for intent classification and slot tagging. *arXiv preprint arXiv:2209.09900*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023a. Can gpt-4 support analysis of textual data in tasks requiring highly specialized domain expertise? *arXiv preprint arXiv:2306.13906*.

Jaromir Savelka, Kevin D Ashley, Morgan A Gray, Hannes Westermann, and Huihui Xu. 2023b. Explaining legal concepts with augmented large language models (gpt-4). *arXiv preprint arXiv:2306.09525*.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. Classactionprediction: A challenging benchmark for legal judgment prediction of class action cases in the us. *arXiv preprint arXiv:2211.00582*.

Paulo Silva, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, and Marilia Curado. 2020. Using nlp and machine learning to detect data privacy violations. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 972–977.

Stavroula Skylaki, Ali Oskooei, Omar Bari, Nadja Herger, and Zac Kriegman. 2020. Named entity recognition in the legal domain using a pointer generator network. *arXiv preprint arXiv:2012.09936*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2020. How to fine-tune bert for text classification?

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the ethical limits of natural language processing on legal text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041*.

Amir Pouran Ben Veyseh, Viet Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2021. Unleash gpt-2 power for event detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6271–6282.

Juan Wang, Yongyi Yang, and Bin Xia. 2019. A simplified cohen's kappa for use in binary classification data annotation tasks. *IEEE Access*, 7:164386–164397.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yaoquan Yu, Yuefeng Guo, Zhiyuan Zhang, Mengshi Li, Tianyao Ji, Wenhu Tang, and Qinghua Wu. 2020. Intelligent classification and automatic annotation of violations based on neural network language model. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

A Experiments Setting

All experiments were conducted on AWS g5.4xlarge instance, equipped with 1 NVIDIA A10G GPU. The total GPU hours are 85. For each model, the reported metrics are obtained by computing the mean and standard deviation across five runs with randomly initialized weights. All code[4], training logs via WANDB[5], datasets, and fine-tuned models[6] are available.

A.1 Library Versions

We used the following libraries and associated versions: python 3.8, transformers 4.31.0, seqeval 1.2.2, streamlit 1.25.0, datasets 2.14.2, evaluate 0.4.0, wandb 0.15.7, torch 2.0.1, accelerate 0.21.0, sentencepiece 0.1.99, google cloud aiplatform 1.28.1, openai 0.27.8, langchain 0.0.248, ipython 8.12.2, typer 0.9.0, nltk 3.8, matplotlib 3.7.2.

B Annotation Platform

We ran our annotation platform with the Argilla library [7] available under an Apache-2.0 license.

Figure 4 shows a screenshot of the annotation platform our human experts used.

C Data Distribution

Figure 5 shows the datasets tokens distribution.

Entity	Description	# Labeled Samples
LAW	Specific law or regulation breached.	292
VIOLATION	Content describing the violation.	1326
VIOLATED BY	Entity committing the violation.	292
VIOLATED ON	Victim or affected party.	292

Table 4: Distribution of the NER entities produced by the generation process (2202 in total).

D Prompts

In this appendix, we detail the data generation prompts utilized for the GPT-4 model. The prompts for the datasets creation are illustrated in Figures 6 and 3. Meanwhile, the prompts for fine-tuning can be found in Figure 7. The prompt for the Few-shot approach is depicted in Figure 8

[4]https://github.com/[SUBMISSION-MASK]/
[5]https://wandb.ai/[SUBMISSION-MASK]
[6]https://huggingface.co/[SUBMISSION-MASK]
[7]https://github.com/argilla-io/argilla

Entity	Description	Labels	# Labeled Samples
Consumer Protection	Deceptive advertising, fraud and unfair business practices.	16/17/29	62
Privacy	Unauthorized collection, use, or disclosure of personal data.	56/54/53	163
TCPA	Unauthorized telemarketing calls, faxes and text messages.	26/27/21	74
Wage	Illegal underpayment and unfair compensation practices by employers.	6/3/4	13

Table 5: Distribution of labeled samples across various legal domains for the NLI task. The number of samples is in the format of Contradiction/Neutral/Entailment.

You are an human expert who helps generate text based on real-world events.

You should write it in a way human been couldn't detect that it isn't real "platform" text.

Write text which describes how the person was affected and not aware of the lawsuit.

Describe how the person was affected before he even knew about the lawsuit.

The person could be male or female at the age of "age".

Write it "doc type" and "grammar mistakes" .

Don't mention the lawsuit.
Don't mention dates.
Don't mention states.
Don't start with "not allowed words" or any other permutations of those words.
Don't mention money or compensation.

The text should be written as "platform" in "length" "hashtags_emoji".

"agenda"

For example:

Description - Xglasses try-on application used facial recognition to scan the user's face and send it to 3rd parties without the user's consent.

"hypothesis example based on agenda"

event description:
"premise"

The output should be wrap in text tags
<text>

Figure 3: Prompt design for generating NLI data set. Prompt contains the task description, specific instructions, and few-shot examples.

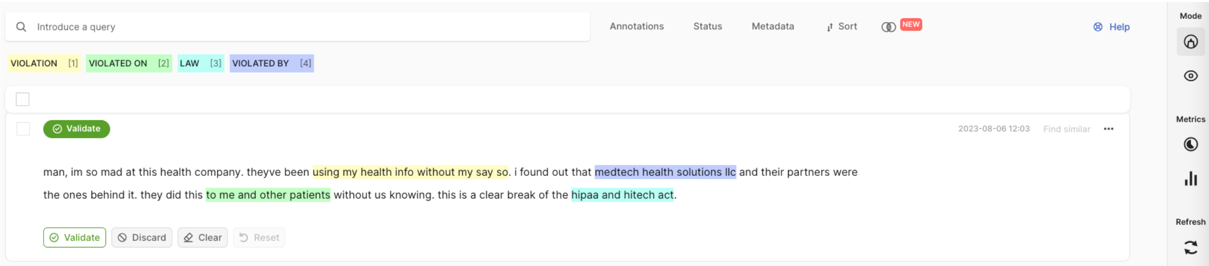


Figure 4: The platform for the human annotations.

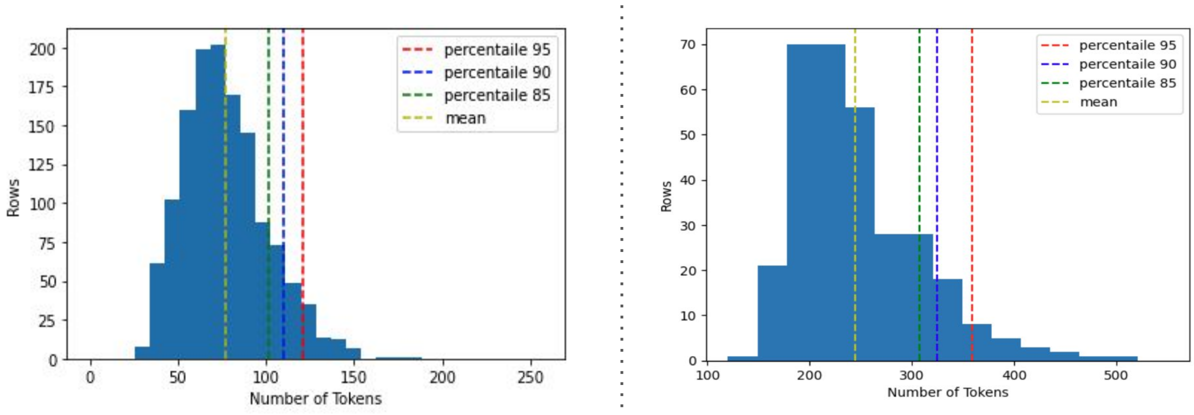


Figure 5: Token Distributions for (left) NER and (right) NLI.

You are an AI assistance that need to write example for training an ml model.
I want you to create for me two examples using this entities:
What is the content that describes the violation. Marks: begin-> [E1], end -> [/E1]

The examples should be paragraphs that contain all of these attributes.

For example:
"I regret to inform potential customers that this banking app has been involved in unsavory practices. They've been caught [E1]quietly charging small unnotified transaction fees and diverting them into undisclosed accounts[/E1]. The inappropriate conduct was led by the company's lead developer and raises serious concerns about the app's credibility."

"The banking app [E1]skimmed undisclosed transaction fees[/E1], led by their lead developer."

The examples should be paragraphs that contain all of these attributes.

Don't stick to the example structure, you can change it as you wish. Use the context below to adjust the story, use augmentation on numbers, dates, names etc to not duplicates examples. You can rephrase the story to other scenarios based on the context. The examples should be from the "coa" cause of action and from the "industry" industry. Don't mention the name of the law in the examples. Write it as a "length" and "text type" text "grammar mistakes" grammar mistakes that has been written as a "doc type". Write each example separately by a newline without numbering prefixes. Don't use any real company/person names. Write it that it will be impossible to know that a model generated this. Context: "context"

(a) Prompt design for Implicit NER data set. Prompt contains the task description, few-shot examples, and specific instructions.

You are an AI assistance that need to write example for training an ml model.
I want you to create for me two examples using these entities:

What is the law that has been broken? Marks: begin-> [E1], end -> [/E1]

What is the content that describes the violation. Marks: begin-> [E2], end -> [/E2]

The violation has been committed by who? This must be explicit and short, don't add non relevant information. Marks: begin-> [E3], end -> [/E3]

The violation has been committed on who (person, group of users etc)? This must be explicit and short, don't add non relevant information. Marks: begin-> [E4], end -> [/E4]

The examples should be paragraphs that contain all of these attributes.

For example:
"The recent case involved a violation of [E1]privacy laws[/E1], where an app was found guilty of [E2]illegally collecting and selling user data[/E2]. It was discovered that [E3]the app developer[/E3] intentionally deceived users by claiming their information would remain secure, but instead, it was being shared with third parties without consent [E4]on unsuspecting users[/E4]."

"In the marketing industry, a prominent advertising agency was found guilty of contravening the [E1]federal trade commission act[/E1] by [E2]misleading consumers with false advertising claims[/E2]. the court determined that [E3]the advertising agency[/E3] had intentionally deceived [E4]the consumers[/E4] by making false claims about the effectiveness of a weight loss product."

"An unsettling incident recently surfaced where an app was indicted for [E2]illegally collecting and selling user data[/E2], constituting a stark violation of [E1]privacy laws[/E1]. Detailed investigations revealed that [E3]the app developer[/E3] had been craftily exploiting [E4]unsuspecting users[/E4], falsely assuring them of data security, whilst secretly passing on their information to third parties."

"Under scrutiny in the realm of marketing was an advertising agency, called to account for [E2]misleading consumers with false advertising claims[/E2]. This breach conspicuously infringed the [E1]federal trade commission act[/E1]. It was adjudicated that [E3]the advertising agency[/E3] had willfully duped [E4]the consumers[/E4] by propagating baseless claims about the efficacy of a weight loss product."

Entities order should be: "entities order". Don't stick to the example structure, you can change it as you wish. Shuffle the appearance of the entities. Use the context below to adjust the story, use augmentation on numbers, dates, names etc to not duplicates examples. You can rephrase the story to other scenarios based on the context. The examples should be from the "coa" cause of action and from the "industry" industry. Write it as a "length" and "text type" text "grammar mistakes" grammar mistakes that has been written as a "doc type". Write each example separately by a newline without numbering prefixes. Don't use any real company/person names. Write it that it will be impossible to know that a model generated this. Context: "context"

(b) Prompt design for Explicit NER data set. Prompt contains the task description, few-shot examples, and specific instructions.

Figure 6: The prompts used for generating the NER data set.

INPUT: in the entertainment industry , a significant case has emerged where a company was found guilty of breaking the tcpa (telephone consumer protection act) . the company was found to have repeatedly sent unsolicited promotional emails about concert tickets to consumers , despite their requests to unsubscribe . the court ruled that the company had knowingly violated the law by continuing to send these emails without the express consent of the consumers . despite the consumers numerous attempts to unsubscribe , the company continued its relentless email campaign. \n### OUTPUT: [{O:in the entertainment industry , a significant case has emerged where a company was found guilty of breaking the}, {B-LAW:tcpa}, {I-LAW:(telephone consumer protection act)}, {O:. the company was found to have}, {B-VIOLATION: repeatedly}, {I-VIOLATION:sent unsolicited promotional emails about concert tickets to consumers , despite their request to unsubscribe}, {O:. the court ruled that}, {B-VIOLATED BY:the}, {I-VIOLATED BY:company}, {O:had knowingly violated the law by continuing to send these emails without the express consent of}, {B-VIOLATED ON:the}, {I-VIOLATED ON:consumers}, {O:. despite the consumers numerous attempts to unsubscribe , the company continued its relentless email campaign .}]\n\n

INPUT: in a shocking revelation , it has been discovered that a popular gaming platform has been distributing pirated copies of video games without obtaining the necessary permissions from the original game developers . this act of unauthorized distribution , even in the face of cease and desist letters , has raised serious concerns about the platforms ethical standards . \n### OUTPUT:

(a) Prompt design for NER. (Top) Training prompt, containing the input and output tags, input text, output text and corresponding NER tags. (Bottom) Inference prompt, containing only the input and output tags, input text.

Premise: <Premise text> ### Hypothesis: <Hypothesis text> ### Label: <entailed / contradict / neutral>

Premise: <Premise text> ### Hypothesis: <Hypothesis text> ### Label:

(b) Prompt design for NLI. (Top) Training prompt, containing the input and output tags, premise and hypothesis texts, and corresponding labels. (Bottom) Inference prompt, containing relevant tags, and premise and hypothesis texts.

Figure 7: The prompts used for fine-tuning open-source LLMs across (a) NER and (b) NLI tasks.

You're an AI language model and your task is to perform Named Entity Recognition (NER) on the provided sentence. Label each word in the sentence with the appropriate class based on the context. Use the following classes for labelling:

LAW: This class refers to a law, regulation, act, or any legal entity.
VIOLATION: This class refers to content that indicates a violation of law, a breach of contract, or misconduct.
VIOLATED BY: This class refers to the person, entity or organization that commits the violation.
VIOLATED ON: This class refers to the person, entity or organization that the violation is committed against.

examples:
{examples}

input:
{input}

Given an input consisting of a premise and a hypothesis, determine if the hypothesis supports, contradicts, or is neutral to the premise. The possible labels are: "**Support**", "**Contradict**", and "**Neutral**".

examples:
{examples}

input:
{input}

Figure 8: Few-shot prompt designs for (top) NER and (below) NLI experiments using OpenAI GPT models. Prompts contain input, general task-specific instructions, labels for each task and few-shot examples.

Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset

Santosh T.Y.S.S[1] **Nina Baumgartner**[2]

Matthias Stürmer[2,3] **Matthias Grabmair**[1] **Joel Niklaus**[2,3,4]

[1]Technical University of Munich [2]University of Bern
[3]Bern University of Applied Sciences [4]Stanford University

Abstract

The assessment of explainability in Legal Judgement Prediction (LJP) systems is of paramount importance in building trustworthy and transparent systems, particularly considering the reliance of these systems on factors that may lack legal relevance or involve sensitive attributes. This study delves into the realm of explainability and fairness in LJP models, utilizing Swiss Judgement Prediction (SJP), the only available multilingual LJP dataset. We curate a comprehensive collection of rationales that 'support' and 'oppose' judgement, from legal experts for 108 cases in German, French, and Italian . By employing an occlusion-based explainability approach, we evaluate the explainability performance of state-of-the-art monolingual and multilingual BERT-based SJP models, as well as models developed with techniques such as data augmentation and cross-lingual transfer, which demonstrated prediction performance improvement. Notably, our findings reveal that improved prediction performance does not necessarily correspond to enhanced explainability performance, underscoring the significance of evaluating models from an explainability perspective. Additionally, we introduce a novel evaluation framework, Lower Court Insertion (LCI), which allows us to quantify the influence of lower court information on model predictions, exposing the current models' bias.

1 Introduction

The task of Legal Judgement Prediction involves analyzing the textual description of case facts to determine various aspects of a case's outcome, such as the winning party, violated provisions, and motion results. It has garnered substantial attention in the mainstream NLP community (Aletras et al., 2016; Chalkidis et al., 2019; Malik et al., 2021a; Niklaus et al., 2021a; Semo et al., 2022) and is being considered as a benchmarking task for evaluating the capabilities of legal NLP (Chalkidis et al., 2022b; Niklaus et al., 2023) and long range (Condevaux and Harispe, 2022; Niklaus and Giofré, 2022; Chalkidis et al., 2022a; Hua et al., 2022) models.

The process of resolving legal cases encompasses evidential reasoning through exchange of arguments between the litigating parities before a decision-making body (Santosh et al., 2022b). Earlier methods to deal with outcome prediction task such as IBP (Brüninghaus and Ashley, 2003), SMILE+IBP (Brüninghaus and Ashley, 2005), VJF (Grabmair, 2017) typically involved identification/extraction of the factors from the textual description of the facts, then employing a conceptual schema to relate the factors to legal issues and predicts the outcome by comparing them with the past cases, thus providing the explanations for those predictions in terms that are legally intuitive. However, in the context of modern deep learning-based solutions, the outcome is determined solely from the text of the case facts, effectively bypassing the interpretable legal reasoning process. This poses a significant risk, particularly in high-stakes domains like law, when utilizing such systems that rely on factors that may be predictive but lack legal relevance or involve sensitive attributes (e.g., the race of an accused person). Such reliance can lead to unjust and biased outcomes, undermining the principles of fairness and equal treatment within the legal system. Hence, such systems need to be analyzed from an explainability standpoint thus making them transparent thereby enhancing the trust of legal practitioners and stakeholders to comprehend the factors and legal principles that contribute to a particular prediction.

In the line of explainable LJP, Chalkidis et al. 2021 investigated the rationales behind models' decisions in Legal Judgment Prediction (LJP) for European Court of Human Rights (ECtHR) cases. Subsequent studies by Santosh et al. 2022b extended the above dataset and Malik et al. 2021a created new dataset for Indian Jurisdiction. In contrast to these works in English, our study focuses on assessing the explainability of LJP models trained on the Swiss-Judgment-Prediction (SJP) dataset, which is the only available multilingual LJP dataset. It contains cases from the Federal Supreme Court of Switzerland (FSCS), written in three official Swiss languages (German, French, Italian)[1]. To this end, we curate a multilingual set of rationales that 'support and 'oppose' Judgment on 108 cases in German, French and Italian collectively. We employ a perturbation-based explainability approach namely Occlusion (Zeiler and Fergus, 2014) wherein we remove the factors from the fact statements and measure the change in the prediction confidence in comparison to a non-occluded baseline. This occlusion based method facilitates to identify the contribution of each factor in arriving at the final prediction, which also links to the characteristics of earlier factor based formal methods of LJP which are known for their interpretability. To enable a fair comparison across methods, we release four distinct occlusion test sets. Each test set involves occluding a different number of sentences (1, 2, 3, and 4) per experiment. This comprehensive range of occlusion scenarios allows us to assess the impact of varying levels of factor removal on the prediction outcomes.

Using the occluded datasets, we assess evaluate the explainability performance of state-of-the-art models developed for SJP task using both monolingual (Niklaus et al., 2021a) and multilingual BERT (Niklaus et al., 2022) architectures, as well as models developed with techniques such as data augmentation and cross-lingual transfer (Niklaus et al., 2022). Our findings highlight the fact that the prediction performance improvement does not translate to explainability improvement.

Furthermore we leverage the peculiar characteristics of the Federal Supreme Court of Switzerland (FSCS), which handles only the most contentious cases that lower courts have struggled to resolve adequately. In their decisions, the FSCS often focuses on specific portions of previous decisions, scrutinizing potential flaws in the lower court's reasoning. This setup offers an intriguing testbed to systematically assess the bias of the lower court in the final predictions generated by our models. This approach is reminiscent of recent works Chalkidis et al. 2022c; Wang et al. 2021 that have examined the fairness of LJP models by examining Group fairness or Disparate Impact i.e, performance disparities across various attributes, such as gender, age, and region. Our approach, termed Lower Court Insertion (LCI), adopts a counterfactual fairness perspective, unlike prior studies examining performance disparities in LJP models. This involves extracting instances of the lower court in each case document and inserting other lower courts into each case to measure the resulting changes in prediction confidence scores. Remarkably, despite the lower court's average length being only 7 words in documents with an average length of 350 words, it has shown the potential to flip the prediction label in some cases.

In sum, our main contributions are as follows:

- We release a new dataset of 108 cases from a trilingual Switzerland Judgment Prediction corpus with rationales annotated by experts to assess the explainability of SJP models.
- We evaluate the state-of-the-art models developed for the SJP task, including monolingual & multilingual models and models trained with several techniques, from explainability standpoint using the occlusion technique.
- We perform systematic evaluation of lower court bias embodied in these models using LCI technique, allowing us to quantify the influence of lower court on the final predictions generated by the models.

2 Related Work

Legal Judgement Prediction: LJP has been studied under various jurisdictions such as the European Court of Human Rights (ECtHR) (Chalkidis et al., 2019, 2021, 2022b; Aletras et al., 2016; Liu and Chen, 2017; Medvedeva et al., 2018, 2021; Santosh et al., 2022a, 2023a,b), Chinese Criminal Courts (Luo et al., 2017; Yue et al., 2021; Zhong et al., 2020), US Supreme Court (Katz et al., 2017; Kaufman et al., 2019), Indian Supreme Court (Malik et al., 2021a), the French court of Cassation (Şulea et al., 2017b,a), Brazilian courts (Bertalan and Ruiz, 2020), the Turkish Constitutional court (Sert et al., 2021), UK courts (Strickson and De La Iglesia, 2020), German courts (Waltl et al., 2017), and the Federal Supreme Court of Switzerland (Niklaus et al., 2021a, 2022; Rasiah et al., 2023) – the only publicly available multi-lingual LJP corpus – which is the main focus of this work.

[1]The dataset consists of non-parallel cases, with each case being unique and decisions being written in a single language.

Swiss Judgement Prediction (SJP): Niklaus et al. 2021a evaluate different methods for the LJP task on the Swiss-Judgment-Prediction (SJP) dataset. They achieve the best performance using a hierarchical variant of BERT that overcomes the token input limitation. Niklaus et al. 2022 further enhance the performance through cross-lingual transfer learning, adapter-based fine-tuning and data augmentation using machine translation. In contrast to previous works, this study examines the explainability of these models and investigates if improved prediction performance translates into improved explainability performance.

Explainability: Explanations in Explainable Artificial Intelligence (XAI) methods are classified based on two factors: whether the explanation is for an individual prediction or the overall prediction process (local or global), and whether the explanation is derived directly from the prediction process or requires post-processing (self-explaining or post-hoc) (Danilevsky et al., 2020). These methods can be model-agnostic (LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), Occlusion (Li et al., 2016; Zeiler and Fergus, 2014), Anchors (Ribeiro et al., 2018)), applicable to any model, or model-specific (Integrated Gradients (Sundararajan et al., 2017), Gradient Saliency, and Attention-Based Methods), designed for specific models. In this study, we use occlusion, a model-agnostic, local, and post-hoc explainability technique.

Fairness: Fairness in machine learning has been defined in different ways to address various types of discrimination. These definitions include group fairness, individual fairness, and causality-based fairness. Group fairness ensures equitable predictions across demographic subgroups, avoiding differential treatment based on attributes such as race, gender, or age (Zafar et al., 2017; Hardt et al., 2016). Individual fairness focuses on treating similar individuals similarly, avoiding arbitrary distinctions based on their characteristics (Sharifi-Malvajerdi et al., 2019; Yurochkin et al.). Causality-based fairness considers underlying causal mechanisms and aims to identify and mitigate biases caused by confounding variables or indirect discrimination (Wu et al., 2019; Zhang and Bareinboim, 2018). In this study, we examine bias related to the lower court variable using counterfactual and causal fairness estimation methods.

Explainability and Fairness in LJP: Early works in the field of legal judgment prediction, such as HYPO (Rissland and Ashley, 1987), CATO (Aleven and Ashley, 1997), IBP (Brüninghaus and Ashley, 2003) and IBP+SMILE (Brüninghaus and Ashley, 2005), relied on symbolic AI techniques to incorporate domain knowledge and provide interpretable explanations for the outcomes. However, deep learning models in LJP have prioritized prediction performance over explainability. Nevertheless, recent research emphasizes the significance of explainability in the legal domain for trust and the right to explanation principle. Efforts have been made to investigate explainability in LJP. For instance, Chalkidis et al. 2021 introduced the task of rationale extraction from facts statements and released a dataset from ECtHR. They used neural models with regularization constraints to select rationales using a learned binary mask. Additionally, Santosh et al. 2022b identified distractor words highly correlated with outcomes but not legally relevant, and proposed an adversarial deconfounding procedure to align model explanations with those chosen by legal experts. Similarly, Malik et al. 2021b developed a dataset of Indian jurisdiction corpus for explainability assessment using the occlusion method. In this work, a dataset from a trilingual Switzerland jurisdiction corpus is curated, and the occlusion method is employed to evaluate models for Swiss Judgment Prediction.

Fair machine learning in the legal domain is a relatively new field. Studies such as Angwin et al. 2016 identified racial bias in the COMPAS system, a parole risk assessment tool in the US, where black individuals were more likely to be mislabeled as high risk. Another study by Wang et al. 2021 found significant fairness gaps across gender in LSTM-based models for legal judgment consistency using a dataset of Chinese criminal cases. Recently, Chalkidis et al. 2022c developed the FairLex benchmark to facilitate research on bias mitigation algorithms in the legal domain. It includes four datasets from different jurisdictions and languages, covering various sensitive attributes. While previous works focused on group fairness and quantifying predictions across demographic subgroups, this study examines a specific variable of lower court from a counterfactual perspective.

3 Occlussion & LCI Dataset for SJP

The SJP dataset (Niklaus et al., 2021a) comprises 85,000 cases from the Federal Supreme Court of Switzerland (FSCS) spanning the years 2000 to

2020, chronologically split into training (2000-14), validation (2015-16) and test (2017-20) splits and are written in three languages: German, French, and Italian. However, it is important to note that the dataset is not evenly distributed among these languages with Italian having a much smaller number of documents (4K) compared to German (50K) and French (31k). Additionally, this representation disparity is also evident across various legal areas and regions. For more detailed dataset statistics, please refer to work of Niklaus et al. 2021a.

3.1 Rationale & Lower Court Annotation

We sample a total of 108 cases from both the validation and test sets (2015-20). These cases were equally distributed across the three languages. Within each year of the validation and test sets, we sampled six cases per language, resulting in two cases per legal area. Specifically, each legal area in every year contained one case with the judgment "approved" and one with the judgment "dismissed." It is worth noting that our annotation dataset is balanced in terms of final outcomes and languages, in contrast to the SJP dataset, which contains a majority of dismissed cases (> 3/4). The annotations were conducted by a team of three legal experts, consisting of two law students pursuing their master's degrees and one lawyer, over a period of five months. Two legal experts are native German speakers with intermediate knowledge in French and basic Italian skills. The third expert is a native speaker in German and Italian and fluent in French. The annotation was facilitated using the Prodigy tool.

The annotation task was to highlight sentences or sub-sentences in the facts section of the judgment that "support" or "oppose" the final outcome of the case. We have chosen sub-sentences as the atomic unit for annotation after consulting with legal experts who expressed that a sentence can contain two sub-sentences opposing each other and hence should be annotated with different labels. The annotators had been given access to the entire case to make their annotation instead of just the facts section, which is the actual input for the models dealing with judgment prediction task. These decisions have been taken to address two points: (i) Experts opined that sentences/sub-sentences may have opposing labels depending on how the court interpreted those facts in its reasoning; hence providing them the entire case would greatly assist them in arriving at explanations leading to higher inter-annotator agreement (ii) Having prior knowledge about a specific case allows an expert to be familiar with its specific legal and factual details, as well as the court's opinions on the matter. As a result, varying levels of prior familiarity with a case can lead to different interpretations and perspectives in understanding it. Hence providing the entire case levels the playing field and eliminates the possibility that some cases are known to only some experts before, possibly leading to different annotations.

The experts are instructed to read through the facts, the considerations, the ruling, and any other needed legal document (such as relevant legislation, analyses or case law) to understand the court case and then annotate the rationale. Unlike the previous works involving explainability annotations in LJP (Chalkidis et al., 2021; Santosh et al., 2022b; Malik et al., 2021a) which only collect rationales that "Supports Judgment", we introduce an additional label termed "Opposes Judgment" which holds significance especially in the task of judgment prediction due to the inherent nature of legal text of often operating within the realm of gray areas rather than clear-cut black-and-white distinctions. Legal cases involve complex issues, conflicting facts leading to alternative legal reasoning, dissenting opinions, alternative interpretations of the law and can serve as potential grounds for challenging the ruling and can serve as a reference point for legal arguments or considerations. Thus, additionally including the "Opposes Judgment" label provides a more comprehensive and nuanced understanding of the case acknowledging that legal decisions are not always unanimous and different perspectives may exist within the legal community.

Additionally, we request annotators to label neutral sentences. This is not a label per se, but covers sentences not assigned other labels, as this assists in implementing the occlusion method to partition the facts section into more coherent sentences with minimal effort, as segmenting legal text is a complex task in itself (Read et al., 2012; Savelka et al., 2017; Brugger et al., 2023).

In addition to sentences and sub-sentences indicating towards outcome explanations, we also ask annotators to label the lower court mentions in the fact section as indicated in the rubrum (header including identifiers, and listing judges, lawyers and involved parties) of the ruling.

The annotation task was conducted in two cycles to ensure high quality. The initial cycle involved pilot annotations, highlighting uncertainties regarding guidelines. As a result, we refined the guidelines by providing more precise instructions to address these concerns[2]. Subsequently, a discussion among the legal experts was held to resolve any conflicts and consolidate the annotations in the most effective manner, thereby ensuring the high quality of the annotations.

3.2 Inter Annotator Agreement

We could obtain annotations from three annotators only for the German subset. Detailed distribution of labeled tokens per annotator can be found in App. A. Among the three, Annotator 1 annotated the least amount of tokens. Annotator 3 annotated the most comparable to the Annotator 2, especially when using the Supports Judgment label. To measure inter-annotator agreement for explanations, we use the machine translation metrics as suggested by Malik et al. 2021a like ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002) (unigram and bigram averaging), METEOR (Agarwal and Lavie, 2007), Jaccard Similarity, Overlap Maximum, Overlap Minimum and BERTScore (Zhang et al., 2020). We report the inter-annotator agreement scores in the German subset for the first round of annotations of German dataset in Table 1. These scores are aggregated over all the labels (supports, opposes judgment and lower court). Table 1 demonstrates high agreement across all scores, with values ranging from 0.7 to 0.9. The high BERTScore indicates strong similarities in non-lexical matches, while the indication of OVERLAP Minimum suggests that the annotations frequently overlapped as subsequences. Notably, Experts 2 and 3 exhibit the highest agreement, which can be attributed to their larger number of annotated tokens compared to Expert 1 (see Appendix A). We also notice that the agreement within the categories "Lower Court" and "Supports Judgment" is notably high in comparison to "Opposes Judgment". The experts confirmed that the higher variance in the "Opposes Judgment" label stemmed from the difficulty in identifying these sentences and resolving these conflicts constituted a significant effort in landing with final annotations. Distribution of final number of tokens obtained per label across language is visualised in App. B. We publicly release the final annotations [3], obtained from the 108 cases, to encourage and facilitate further research in this area.

IAA metric	A1&A2	A1&A3	A2&A3
Rouge-1	0.78	0.69	0.87
Rouge-2	0.74	0.64	0.85
Rouge-L	0.77	0.68	0.87
BLEU	0.75	0.69	0.85
METEOR	0.77	0.71	0.88
Jaccard Sim.	0.73	0.64	0.82
Overlap Max.	0.68	0.61	0.74
Overlap Min.	0.83	0.73	0.81
BERTScore	0.91	0.86	0.93

Table 1: IAA score between the annotators in the first cycle for German subset

3.3 Occlusion and LCI dataset

To evaluate the explainability of models and enable a fair comparision among them, we derive four distinct occlusion based datasets from the test split[4] of above annotated rationales data, consisting of 27, 24 and 23 cases in German, French and Italian respectively. For each test set, we occlude a different number of sentences (1, 2, 3, and 4) belonging to same label (Supports/Opposes Judgment/Neutral) per experiment in a case, adding no marker or trace of the occlusion in the fact section to leave it as similar and natural as possible. For every test experiment instance, we also include it with a baseline wherein there is no text occluded. Thus, we arrive in a total of 28k occluded instances with varying levels of occlusion, across three languages. Using these occluded instances, we analyze the difference in prediction confidence in comparison to the non-occluded baseline.

For LCI, we derive the counterfactual based test set wherein we extract the lower court instances annotated by the annotator and insert each of the other lower court in each case resulting in a total of 1127 instances. There are a total of 13, 9, 16 unique lower court instances in German, French and Italian respectively. Similar to above, each instance also has a baseline which represents the case text with actual lower court name without any insertion and use it to analyze the change in prediction confidence.

Table 2 provides statistics on the total number of instances in both the Occlusion and LCI test

[2] We will make the full Annotation Guidelines along with discussion log available upon publication.

[3] URL anonymized

[4] We exclude the instances from the validation split, which is used for hyperparameter tuning during model training, to derive the occluded test set for explainability.

	Occlusion				LCI
	Opposes	Neutral	Supports	Total	Total
DE	201	11124	1243	12568	351
FR	64	3811	2467	6342	391
IT	63	9155	203	9421	312

Table 2: Statistics of the Occlusion and LCI test sets across the three languages.

sets. For a detailed breakdown of the number of instances in each occlusion set by language, please refer to Appendix C. Across languages, the German subset comprises the largest portion of the test set, as the French and Italian datasets have fewer sentences and a lower number of annotated tokens in French compared to Italian. Among the labels, the 'Opposes Judgment' label has fewer instances, which can be attributed to the lower number of annotated tokens associated with this label.

4 Experimental Setup

4.1 Models

We assess the following six class of models, developed on the backbone of hierarchical BERT, developed for the SJP task in previous literature (Niklaus et al., 2021a, 2022). We follow the same dataset splits provided by Niklaus et al. 2021a for training and validation. Hierarchical BERT is employed because the SJP dataset includes documents with more than 512 tokens. In this approach, the text is split into 4 consecutive blocks of 512 tokens (90% of cases are less than 2048 tokens) and fed into a shared standard BERT encoder independently. Then the CLS token of each block is passed through a 2-layer transformer encoder to aggregate the information across blocks, followed by max-pooling and a final classification layer.

MonoLingual: This variant uses monolingually pre-trained BERT models i.e German-BERT (Chan et al., 2019), CamemBERT (Martin et al., 2020) and UmBERTo (Parisi et al., 2020) for German, French and Italian. Each model is fine-tuned and evaluayed using that language subset dataset.

MultiLingual: This variant uses the multilingually pre-trained XLM-R model (Conneau et al., 2019) instead of language-specific pre-trained BERT. However, the fine-tuning process is still performed separately for each language, similar to the MonoLingual approach.

Mono/Multi Lingual with Data Augmentation: We translate the cases in SJP dataset into other languages from the original language using the EasyNMT2 framework, following the approach proposed by Niklaus et al. 2022. Then these translated instances are then augmented with the original data for a specific language during the fine-tuning process with Mono/Multilingual BERT. This is similar to above experiment in setup, with the main distinction being the additional augmented data.

Joint Training without/with Data Augmentation: We use a multilingual pre-trained model and fine-tune it across all the three language corpora jointly, which tries to capitalize on the inherited benefit of using larger multilingual corpora during fine-tuning. Similarly, data augmented version with translated versions of all the corpus into other languages is used for fine-tuning. Unlike the previous approaches where separate models were fine-tuned for each language, this method jointly fine-tunes on all languages, resulting in a single final model instead of multiple models for each language.

4.2 Implementation Details

We use the code repositories from prior work Niklaus et al. 2021a, 2022 to assess the state-of-the-art models on SJP[5]. We employ a learning rate of 1e-5 with early stopping based on macro-F1 on the development set. All models are trained with a batch size of 64 for 10 epochs using AdamW optimizer with mixed precision and gradient accumulation using huggingface library (Wolf et al., 2020). We use oversampling to handle class imbalance. We use 4 as the number of segments in our hierarchical models which make maximum sequence length of 2048.

4.3 Metrics

We report macro-F1 following (Niklaus et al., 2021a, 2022) for assessing prediction performance. For assessing explainability through occlusion experiments, we calculate the explainability score S_{exp} for every test instance as the difference between the temperature-scaled[6] confidence of the baseline and the occluded instance. (i.e baseline - occluded). A negative (positive) S_{exp} score indicates that occluded text is opposing (supporting) its prediction. Then, we assign the label 'Opposes Judgement'/'Neutral'/'Supports Judgement' based on the sign of explainability score. Finally, we report F1-score for each of the labels across all the occluded instances.

[5]https://github.com/JoelNiklaus/SwissJudgementPrediction

[6]We adopt temperature scaling (Guo et al., 2017) to calibrate the confidence estimates of the model.

Model	German	French	Italian
MonoLingual	69.08	71.78	67.82
MultiLingual	67.92	69.24	65.28
MonoLingual + DA	70.47	71.24	69.21
MultiLingual + DA	68.94	71.06	69.86
Joint Training	68.74	70.82	70.62
Joint Training + DA	70.58	71.62	71.18

Table 3: Prediction Performance on Test set of Niklaus et al. 2021b

In bias estimation using the LCI method, we calculate an explainability score for each instance. As the explainability scores are sign dependent, we separately compute the Mean of Explainability Scores (MES), for positive and negative values, expressed as a percentage. A positive explainability score indicates that the insertion of the lower court decreases the probability, suggesting that the inserted court has a pro-dismissal influence. Conversely, a negative score indicates an increase in the probability, indicating a pro-approval trend of the inserted lower court. For an ideally unbiased model, the presence of the lower court should not affect the probability of the prediction. Therefore, a value of the mean explainability score closer to 0 is desirable. Additionally, we report the percentage of cases where the insertion of the lower court leads to a flip in the label of the prediction, changing it from 0 to 1 or vice versa.

4.4 Analysis using Occlusion Test

We present the results of prediction performance and explainability analysis using occlusion in Tables 3 and 4, respectively. Analyzing Table 4, we observe that the model achieves higher accuracy in classifying instances with Supports Judgment compared to those with Neutral or Opposes Judgment. This could be attributed to the fact that the Opposes Judgment category is underrepresented in the occlusion dataset (due to fewer annotated tokens with this label) and the challenging task to classify Neutral instances. Among the three languages, French exhibits the highest score for the Supports Judgment category, but it also shows lower scores for the other classes.

Despite the Multilingual model displaying a decrease in predictive performance, it shows some improvement in occlusion performance, particularly for the Supports Judgments class, across all languages. A similar trend is observed in the Joint training model, which consistently demonstrates a significant increase in explainability scores across languages for most classes.

While the inclusion of the DA component in both MonoLingual and MultiLingual models resulted in improved explainability scores for most labels compared to their counterparts, its addition to the Joint training model leads to mixed results. Surprisingly, the addition of the DA component to the Joint training model consistently increases prediction performance but does not maintain consistency in explainability performance. This finding emphasizes the importance of evaluating explainability to develop transparent systems that can make accurate predictions for the right reasons.

Overall, the lower scores across the board indicate the flawed inference about factors predictive for the outcome. Despite the impressive performances of state-of-the-art models on standard LJP prediction performance, there is still much progress to be made to make those models align as closely as possible with the rationales deemed relevant by legal experts. To create practical value for the legal field, the field of LJP should aim for a productive fusion of expert knowledge and data-driven insights, rather than data-driven correlation based learning.

4.5 Analysis using LCI Test

From Table 5, we can observe that the modification of the lower court has a considerable influence on the overall prediction confidence, as indicated by the changes in confidence scores up to 5% in both directions across all languages, despite the lower court name on average spans around seven tokens in documents of an average length of 350 tokens. However, these smaller changes in confidence scores did result in label flips.

Overall, no consensus exists on which model setting has yielded lower MES scores in both directions consistently across all the languages. MultiLingual model's prediction performance decreased compared to the MonoLingual model across all the three languages and its bias scores increased significantly across all languages, barring -MES for Italian and +MES for German. While the inclusion of the DA (Data Augmentation) component resulted in improved prediction performance compared to the non-DA variants in both MonoLingual and MultiLingual settings, the Multilingual + DA model exhibited a notable increase in bias. This suggests that the model's reliance on lower court names is more pronounced in the presence of the

Model	German			French			Italian		
	Opposes	Neutral	Supports	Opposes	Neutral	Supports	Opposes	Neutral	Supports
MonoLingual	3.02	16.78	15.1	1.95	3.68	40.24	0.49	3.68	11.24
MultiLingual	2.04	11.90	17.46	1.77	3.62	42.77	0.85	5.72	13.48
MonoLingual + DA	3.21	16.26	18.08	1.78	5.98	43.12	0.98	4.39	14.99
MultiLingual + DA	3.64	19.06	20.83	1.43	4.63	45.77	0.83	4.84	15.36
Joint Training	2.62	15.72	26.97	1.67	4.19	48.51	0.54	5.37	18.82
Joint Training + DA	3.75	14.54	21.95	1.93	5.73	45.73	0.63	4.82	19.68

Table 4: Analysis of Explainability using Occlusion - F1-scores across all instances from all four test sets for each label in every language. Higher the scores, better the explainability.

Model	German				French				Italian			
	+ MES	- MES	Flip $1\rightarrow 0$	Flip $0\rightarrow 1$	+ MES	- MES	Flip $1\rightarrow 0$	Flip $0\rightarrow 1$	+ MES	- MES	Flip $1\rightarrow 0$	Flip $0\rightarrow 1$
MonoLingual	$3.48_{5.12}$	$-2.3_{2.82}$	2.28	0.43	$2.56_{2.65}$	$-2.2_{3.79}$	3.84	1.02	$1.64_{3.46}$	$-2.22_{7.76}$	0.12	1.20
MultiLingual	$3.39_{5.43}$	$-2.77_{3.64}$	1.71	0.32	$4.01_{4.62}$	$-3.06_{3.22}$	0.51	0.12	$1.72_{2.18}$	$-1.78_{2.52}$	1.05	2.88
MonoLingual + DA	$3.09_{5.15}$	$-2.77_{5.42}$	2.56	0.22	$4.12_{6.73}$	$-1.83_{2.34}$	1.24	0.82	$1.25_{2.23}$	$-1.29_{1.97}$	0.32	2.19
MultiLingual + DA	$5.32_{8.27}$	$-3.35_{5.24}$	4.56	2.56	$4.08_{6.05}$	$-6.48_{9.64}$	1.53	3.07	$2.88_{3.51}$	$-6.12_{8.08}$	2.56	3.85
Joint Training	$3.32_{6.18}$	$-1.86_{2.89}$	3.13	1.99	$4.08_{4.37}$	$-2.71_{3.69}$	0.51	2.56	$6.13_{6.79}$	$-2.66_{2.12}$	4.92	4.24
Joint Training + DA	$3.23_{4.46}$	$-1.84_{2.45}$	2.85	1.99	$3.04_{3.96}$	$-4.04_{4.31}$	3.07	2.32	$6.14_{7.94}$	$-3.21_{3.11}$	4.09	4.83

Table 5: Analysis of Lower Court Bias using LCI - Results of Positive and Negative MES Scores, and Label Flips across the three languages. Labels 0 and 1 indicates dismissal and approval respectively. Lower scores indicate a less biased model. Subscript indicates the standard deviation values.

DA component compared to its non-DA variant.

Joint training models, which aim to generalize across languages, demonstrate improved prediction performance, particularly for Italian, which is underrepresented in the training set. However, this improvement comes at the cost of higher MES scores, indicating potential overfitting to court-specific correlations rather than capturing the actual reasoning behind the predictions. Interestingly, despite training a single model using data from all three languages, the Italian data shows a significant divergence in MES scores compared to German and French. This highlights that representational bias across languages seems to be a crucial part. While adding DA to Joint training models significantly improves prediction performance, bias scores lack a clear pattern.

Across all the models in the German setting, we can witness an overall pro-dismissal trend (greater + MES scores compared to - MES) echoing with more number of label flips from approval to dismissal. While French notices an overall pro-dismissal trend, Italian shows pro-approval trend, barring the cross-lingual models. These observed biases regarding lower court underscore the need for continuous bias evaluation and mitigation in LJP models.

5 Conclusion

In this work, we present the rationale dataset curated at fine-grained level of both ‘supporting’ and ’opposing’ factors for Swiss Judgment Prediction (SJP), the only available multilingual LJP dataset. We employ a perturbation-based occlusion approach to assess various state-of-the-art models developed for SJP and also release four distinct occlusion test sets, occluding a different number of sentences in each of the sets. Our lower explainability scores suggest that the current models do not align well with the legal experts which can lead to sub-optimal litigation strategies due to flawed inference about factors responsible for the outcome. Furthermore, we assess the bias of the lower court information in the final predictions generated by the models using LCI test and notice that models learn court-outcome spurious correlations in the data. In future, we would explore deconfounding strategy (Santosh et al., 2022b) to improve the alignment between what models and experts’ deem relevant. One can explore different group robust algorithms such as adversarial removal, IRM, Group DRO and V-REx, as an effective bias-mitigation strategy (Chalkidis et al., 2022c) and investigate its impact on explainability. We hope our data resource will be useful to the research community working on Legal Judgement Prediction.

Limitations

In this study, our approach to obtaining rationales involved a consolidation process wherein we aimed to achieve a final set of high-quality annotations through discussions with legal experts. However, it is important to acknowledge that the assumption of a single ground truth may overlook the presence of genuine human variation, which can arise due to factors such as disagreement, subjectivity in annotation, or the existence of multiple plausible answers. Particularly in the field of law, where complexity and interpretation are inherent, it is well-recognized that lawyers may have differing legal assessments of case facts and how they contribute to the eventual outcome. Instead of attempting to resolve variations in expert labels, it is essential to acknowledge and embrace the inherent variation in human annotations. Moving forward, it is crucial to develop methods that can comprehensively capture and account for variation from data to evaluation, enabling a more comprehensive treatment of this variability in future research.

In this work, we focus on obtaining in-text rationales for SJP, i.e. spans of text in the facts statement marked up by annotators. However, in-text rationales only provide evidence for the outcome without conveying the mechanisms for how the evidence leads to the outcome. A potentially better way to alleviate both these limitations would be to obtain free-text rationales (Tan, 2022), which can be explored in future. On the other hand, this makes it challenging to evaluate the current SJP models which are discriminative in nature.

In the evaluation of our occlusion-based explainability setup, we utilized the F1-score, which focuses solely on the final label obtained from the change in confidence score between the baseline and occluded instances. However, it is important to emphasize the need for a metric that takes into account the magnitude of the difference in confidence scores during aggregation, in order to present a more comprehensive and holistic assessment.

Ethics Statement

The dataset used in this work comes from prior work by Niklaus et al. 2021b and these are publicly available on the https://entscheidsuche.ch platform and the names of the parties have been redacted by the court to ensure anonymity.

This work does not endorse or advocate for practical use of such systems. Instead our aim in this work is to rather empirically demonstrates that these systems are far from practical use due to their flawed inference about factors leading to outcome prediction. The scope of this work is to study LJP from explainability standpoint and to showcase the discrepancy between the prediction performance and explainability performance and emphasize the need to build technology that can help practitioners with reliable insights. Our dataset and findings associated with this work will contribute to advancing the field of explainable legal judgement prediction and provide valuable insights for developing more reliable and unbiased models in the future.

Furthermore, we would like to draw attention to the work by Tsarapatsanis and Aletras 2021 which discusses various normative factors related to ethics in the context of legal natural language processing. These discussions are crucial for fostering ethical thinking within the legal NLP community and ensuring the responsible development of systems that can assist lawyers, judges, and the general public.

References

Abhaya Agarwal and Alon Lavie. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of WMT-08*.

Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. 2016. Predicting judicial decisions of the european court of human rights: A natural language processing perspective. *PeerJ Computer Science*, 2:e93.

Vincent Aleven and Kevin D Ashley. 1997. Teaching case-based argumentation through a model and examples: Empirical evaluation of an intelligent learning environment. In *Artificial intelligence in education*, volume 39, pages 87–94. Citeseer.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23:77–91.

Vithor Gomes Ferreira Bertalan and Evandro Eduardo Seron Ruiz. 2020. Predicting judicial outcomes in the brazilian legal system using textual features. In *DHandNLP@ PROPOR*, pages 22–32.

Tobias Brugger, Matthias Stürmer, and Joel Niklaus. 2023. MultiLegalSBD: A Multilingual Legal Sentence Boundary Detection Dataset. ArXiv:2305.01211 [cs].

Stefanie Brüninghaus and Kevin D Ashley. 2003. Combining case-based and model-based reasoning for predicting the outcome of legal cases. In *Case-Based*

Reasoning Research and Development: 5th International Conference on Case-Based Reasoning, ICCBR 2003 Trondheim, Norway, June 23–26, 2003 Proceedings 5, pages 65–79. Springer.

Stefanie Brüninghaus and Kevin D Ashley. 2005. Generating legal arguments and predictions from case texts. In *Proceedings of ICAIL 2005*, pages 65–74.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in english. In *Proceedings of ACL 2019*, pages 4317–4323.

Ilias Chalkidis, Xiang Dai, Manos Fergadiotis, Prodromos Malakasiotis, and Desmond Elliott. 2022a. An exploration of hierarchical attention transformers for efficient long document classification. *arXiv preprint arXiv:2210.05529*.

Ilias Chalkidis, Manos Fergadiotis, Dimitrios Tsarapatsanis, Nikolaos Aletras, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. Paragraph-level rationale extraction through regularization: A case study on european court of human rights cases. In *Proceedings of the NAACL-HLT 2021*, pages 226–241.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022b. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of ACL 2022*, pages 4310–4330.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. 2022c. Fairlex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406.

Branden Chan, Timo Möller, Malte Pietsch, Tanay Soni, and Chin Man Yeung. 2019. Deepset-open sourcing german bert.

Charles Condevaux and Sébastien Harispe. 2022. Lsg attention: Extrapolation of pretrained transformers to long sequences. *arXiv preprint arXiv:2210.15497*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable ai for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459.

Matthias Grabmair. 2017. Predicting trade secret case outcomes using argument schemes and learned quantitative value effect tradeoffs. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 89–98.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Wenyue Hua, Yuchen Zhang, Zhe Chen, Josie Li, and Melanie Weber. 2022. LegalRelectra: Mixed-domain Language Modeling for Long-range Legal Text Comprehension. ArXiv:2212.08204 [cs].

Daniel Martin Katz, Michael J Bommarito, and Josh Blackman. 2017. A general approach for predicting the behavior of the supreme court of the united states. *PloS one*, 12(4):e0174698.

Aaron Russell Kaufman, Peter Kraft, and Maya Sen. 2019. Improving supreme court forecasting using boosted decision trees. *Political Analysis*, 27(3):381–387.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhenyu Liu and Huanhuan Chen. 2017. A predictive performance comparison of machine learning models for judicial cases. In *2017 IEEE Symposium series on computational intelligence (SSCI)*, pages 1–6. IEEE.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. 2017. Learning to predict charges for criminal cases with legal basis. In *Proceedings of EMNLP 2017*, pages 2727–2736.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021a. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of ACL-IJCNLP 2021*, pages 4046–4062.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021b. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4046–4062, Online. Association for Computational Linguistics.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *ACL 2020-58th Annual Meeting of the Association for Computational Linguistics*.

Masha Medvedeva, Ahmet Üstün, Xiao Xu, Michel Vols, and Martijn Wieling. 2021. Automatic judgement forecasting for pending applications of the european court of human rights. In *ASAIL/LegalAIIA@ ICAIL*.

Masha Medvedeva, Michel Vols, and Martijn Wieling. 2018. Judicial decisions of the european court of human rights: Looking into the crystal ball. In *Proceedings of the conference on empirical legal studies*, page 24.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021a. Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35.

Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. 2021b. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 19–35, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Niklaus and Daniele Giofré. 2022. Budgetlongformer: Can we cheaply pretrain a sota legal language model from scratch? *arXiv preprint arXiv:2211.17135*.

Joel Niklaus, Veton Matoshi, Pooja Rani, Andrea Galassi, Matthias Stürmer, and Ilias Chalkidis. 2023. Lextreme: A multi-lingual and multi-task benchmark for the legal domain. *arXiv preprint arXiv:2301.13126*.

Joel Niklaus, Matthias Stürmer, and Ilias Chalkidis. 2022. An empirical study on cross-x transfer for legal judgment prediction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 32–46.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. *Original-date*, 55:31Z.

Vishvaksenan Rasiah, Ronja Stern, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, Daniel E. Ho, and Joel Niklaus. 2023. SCALE: Scaling up the Complexity for Advanced Language Model Evaluation. ArXiv:2306.09237 [cs].

Jonathon Read, Rebecca Dridan, Stephan Oepen, and Lars Jørgen Solberg. 2012. Sentence Boundary Detection: A Long Solved Problem? In *Proceedings of COLING 2012: Posters*, pages 985–994, Mumbai, India. The COLING 2012 Organizing Committee.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Edwina L Rissland and Kevin D Ashley. 1987. A case-based system for trade secrets law. In *Proceedings of the 1st international conference on Artificial intelligence and law*, pages 60–66.

T. Y. S. S Santosh, Marcel Perez San Blas, Phillip Kemper, and Matthias Grabmair. 2023a. Leveraging task dependency and contrastive learning for case outcome classification on european court of human rights cases. *arXiv preprint arXiv:2302.00768*.

T. Y. S. S Santosh, Oana Ichim, and Matthias Grabmair. 2023b. Zero shot transfer of article-aware legal outcome classification for european court of human rights cases. *arXiv preprint arXiv:2302.00609*.

T. Y. S. S. Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022a. Deconfounding Legal Judgment Prediction for European Court of Human Rights Cases Towards Better Alignment with Experts. ArXiv:2210.13836 [cs].

T.y.s.s Santosh, Shanshan Xu, Oana Ichim, and Matthias Grabmair. 2022b. Deconfounding legal judgment prediction for European court of human rights cases towards better alignment with experts. In *Proceedings of EMNLP 2022*.

Jaromir Savelka, Vern Walker, Matthias Grabmair, and Kevin Ashley. 2017. Sentence Boundary Detection in Adjudicatory Decisions in the United States. *Traitement Automatique des Langues*, 58:21.

Gil Semo, Dor Bernsohn, Ben Hagag, Gila Hayat, and Joel Niklaus. 2022. ClassActionPrediction: A Challenging Benchmark for Legal Judgment Prediction of Class Action Cases in the US. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 31–46, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mehmet Fatih Sert, Engin Yıldırım, and İrfan Haşlak. 2021. Using artificial intelligence to predict decisions of the turkish constitutional court. *Social Science Computer Review*, page 08944393211010398.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. 2019. Average individual fairness: Algorithms, generalization and experiments. *Advances in neural information processing systems*, 32.

Benjamin Strickson and Beatriz De La Iglesia. 2020. Legal judgement prediction for uk courts. In *Proceedings of the 2020 the 3rd international conference on information science and system*, pages 204–209.

Octavia-Maria Şulea, Marcos Zampieri, Shervin Malmasi, Mihaela Vela, Liviu P Dinu, and Josef van Genabith. 2017a. Exploring the use of text classification in the legal domain.

Octavia-Maria Şulea, Marcos Zampieri, Mihaela Vela, and Josef van Genabith. 2017b. Predicting the law area and decisions of french supreme court cases. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 716–722.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328. PMLR.

Chenhao Tan. 2022. On the diversity and limits of human explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2173–2188.

Dimitrios Tsarapatsanis and Nikolaos Aletras. 2021. On the Ethical Limits of Natural Language Processing on Legal Text. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3590–3599, Online. Association for Computational Linguistics.

Bernhard Waltl, Georg Bonczek, Elena Scepankova, Jörg Landthaler, and Florian Matthes. 2017. Predicting the outcome of appeal decisions in germany's tax law. In *International conference on electronic participation*, pages 89–99. Springer.

Yuzhong Wang, Chaojun Xiao, Shirong Ma, Haoxi Zhong, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2021. Equality before the law: Legal judgment consistency analysis for fairness. *arXiv preprint arXiv:2103.13868*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

Linan Yue, Qi Liu, Binbin Jin, Han Wu, Kai Zhang, Yanqing An, Mingyue Cheng, Biao Yin, and Dayong Wu. 2021. Neurjudge: a circumstance-aware neural framework for legal judgment prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 973–982.

Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. In *International Conference on Learning Representations*.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer.

Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *arXiv:1904.09675 [cs]*. ArXiv: 1904.09675.

Haoxi Zhong, Yuzhong Wang, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Iteratively questioning and answering for interpretable legal judgment prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1250–1257.

A Distribution of tokens labeled by annotators in annotation

Figure 1 displays the distribution of the explainability labels annotated by the three annotators for the German dataset. Annotator 1 annotated the least amount of tokens, while annotator 3 annotated the most, especially with respect to "Supports Judgment label".

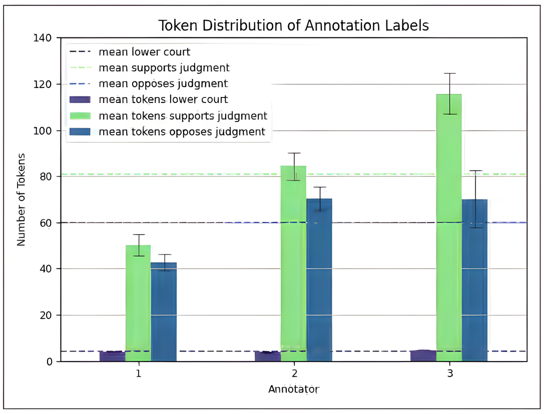


Figure 1: Mean number of tokens annotated per label per annotator in German subset

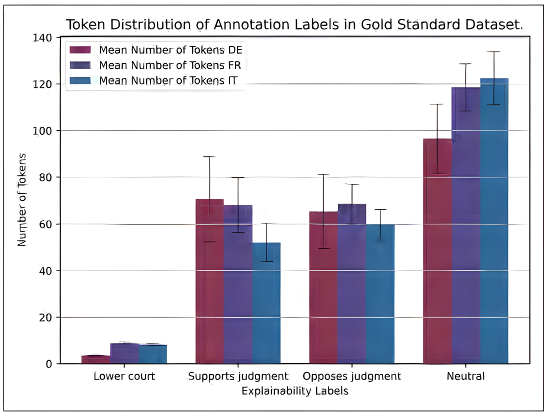


Figure 2: Distribution of the number of tokens per label in the final dataset across each language.

B Mean number of tokens annotated per label per language

Fig. 2 displays mean number of tokens annotated per label in the final round across each language.

C Occlusion Dataset

Table 6, 7 reports number of instances of each label in each occlusion test set across three languages.

D Explainability performance for different levels of occlusion

We report the label wise F1-score for each occlusion test for every language. Overall, 'Opposes judgement' and 'Neutral' are challenging ones compared to "Supports judgement". In the case of French, there was an improvement in scores as the number of occluded sentences increased. This improvement indicates that the model was able to correctly associate the label 'supports judgement' with the occluded sentences, thereby enhancing the model's explainability performance. However, in the case of French and Italian, a different trend was observed. The model did not exhibit the same improvement as the number of occluded sentences increased. It is speculated that the model might have encountered conflicting labels for each occluded sentence, leading to incorrect predictions when multiple occlusions were present.

		Set - 1			Set - 2			Set - 3		
	Documents	Opposes	Neutral	Supports	Opposes	Neutral	Supports	Opposes	Neutral	Supports
DE	27	55	247	98	66	1097	203	53	3158	356
FR	24	34	164	85	22	586	246	7	1260	659
IT	23	31	195	50	23	827	69	8	2429	56

Table 6: Split of number of instances per label in each occluded test set-1,2,3 across three languages.

	Set- 4			Total			Total
	Opposes	Neutral	Supports	Opposes	Neutral	Supports	
DE	27	6622	586	201	11325	1243	12769
FR	1	1801	1477	64	3875	2467	6406
IT	1	5704	28	63	9218	203	9484

Table 7: Split of number of instances per label in each occluded test set-4 and total across three languages.

	Set 1			**Set 2**			**Set 3**			**Set 4**		
Model	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**
MonoLingual	23.08	20.98	33.86	8.80	17.85	27.30	3.28	20.19	16.84	1.15	21.34	11.89
MultiLingual	23.79	18.05	34.46	8.12	13.87	26.81	1.52	12.79	24.25	1.08	14..10	11.17
MonoLingual + DA	24.05	21.77	25.12	10.69	19.29	17.34	3.93	18.24	10.01	0.76	10.80	22.03
MultiLingual + DA	28.80	26.40	36.44	12.45	21.64	27.98	4.49	19.83	17.15	1.04	18.01	11.08
Joint Training	24.52	20.14	31.87	9.86	10.62	28.25	2.69	9.20	27.52	0.98	5.75	26.14
Joint Training + DA	23.85	23.33	38.3	13.40	16.02	25.11	4.71	10.74	12.67	1.24	7.77	8.74

Table 8: Explainability performance for German dataset over different occlusion test sets

	Set 1			**Set 2**			**Set 3**			**Set 4**		
Model	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**
MonoLingual	21.48	10.17	36.67	4.34	4.87	38.17	1.34	4.17	53.64	0.18	3.58	67.99
MultiLingual	26.80	2.40	40.98	6.19	0.68	38.14	1.05	0.79	40.19	0.00	0.22	51.58
MonoLingual + DA	20.05	10.53	28.57	4.72	8.67	43.29	1.28	6.79	58.81	0.14	4.35	69.36
MultiLingual + DA	21.11	4.73	30.41	6.17	2.02	38.24	1.03	1.26	41.39	0.00	1.87	40.86
Joint Training	21.43	3.49	37.37	7.09	0.68	43.02	1.11	1.12	49.77	0.12	1.12	58.29
Joint Training + DA	28.11	14.05	38.78	8.51	9.69	44.26	1.21	10.09	43.65	0.00	8.80	42.04

Table 9: Explainability performance for French dataset over different occlusion test sets

	Set 1			**Set 2**			**Set 3**			**Set 4**		
Model	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**	**Opposes**	**Neutral**	**Supports**
MonoLingual	9.01	24.70	23.53	2.01	26.56	6.72	0.52	31.08	0.22	0.08	32.43	0.00
MultiLingual	12.84	19.18	31.25	2.54	7.66	14.86	0.41	7.22	4.84	0.0	3.92	1.08
MonoLingual + DA	22.11	3.03	27.10	5.61	0.24	13.37	0.84	0.33	6.12	0.05	0.35	1.92
MultiLingual + DA	21.28	11.43	35.82	8.66	2.62	20.77	0.76	1.06	8.51	0.06	0.18	2.46
Joint Training	17.24	11.21	25.61	3.27	5.61	17.55	0.37	1.79	8.07	0.0	0.63	3.01
Joint Training + DA	20.99	17.94	33.53	3.99	6.50	23.02	0.46	2.28	11.3	0.04	0.77	3.91

Table 10: Explainability performance for Italian dataset over different occlusion test sets

Re-Identifizierung in Gerichtsurteilen mit Simap Daten

Autor/Autorin: Joel Niklaus / Magda Chodup / Thomas Lüthi / Daniel Kettiger
Beitragsart: Wissenschaftliche Beiträge
Rechtsgebiete: öffentliches Beschaffungsrecht; Datenschutzrecht

Die digitale Transformation erreicht nach und nach immer mehr Bereiche der Justiz. Bereits heute veröffentlichen viele Gerichte ihre Urteile in anonymisierter Form im Internet. Gleichzeitig werden technische Hilfsmittel, die auch zur Re-Identifikation dieser Urteile eingesetzt werden können, immer leistungsfähiger und ausgeklügelter. In der vorliegenden Untersuchung wurde im Bereich des öffentlichen Beschaffungswesens – durch ein vergleichsweise einfaches «String-Matching» mit Simap Projektnummern – eine Re-Identifikation von Verfahrensbeteiligten von bis zu 81.2 Prozent erreicht.

Inhaltsverzeichnis

Inhalt

1. Einleitung

[1] Im Rahmen des Nationalen Forschungsprogramms NFP77, welches sich mit der digitalen Transformation unserer Gesellschaft befasst, wurde das Projekt «Open Justice versus Privacy» lanciert. In diesem Forschungsprojekt soll untersucht werden, inwiefern die heute gängigen Anonymisierungsparadigma von Gerichtsentscheiden – insbesondere unter Berücksichtigung neuer technischer Möglichkeiten – Schwachstellen aufweisen, und wie diese korrigiert werden können. Es stellt sich zum Beispiel die Frage, ob es denkbar sei, dass künstliche Intelligenz eine Re-Identifikation veröffentlichter Urteile ermöglicht.

[2] Im Rahmen der Vorbereitung des Projekts «Open Justice versus Privacy» wurde eine Datenbank schweizerischer Gerichtsentscheide (insbesondere auch von Bundesgerichtsurteilen) mit mehreren hunderttausend Urteilen aufgebaut. Im Projekt werden anhand dieser Datenbank verschiedene Re-Identifikationsstrategien zusammengetragen und getestet. Ein erstes Ergebnis lieferte ein «String-Matching» Versuch im Bereich des öffentlichen Beschaffungswesens. Hierbei wurden Projektnummern in Urteilen mit publizierten Nummern auf «simap»[1] gepaart. Mittels dieses Verfahrens konnten, aufgegliedert auf die verschiedenen Landessprachen, bis zu 81.2 Prozent der Beschwerdegegner/innen re-identifiziert werden.

[1] Simap ist die wichtigste amtliche elektronische Plattform im Bereich des öffentlichen Beschaffungswesens. Sie wird gemeinsam von Bund, Kantonen und Gemeinden betrieben (abrufbar unter: www.simap.ch).

[3] Die Re-Identifikation von Gerichtsurteilen wurde auch von VOKINGER/MÜHLEMATTER[2] mittels «Linkage» von Datenbanken im Bereich von Arzneimitteln versucht, mit dem Ergebnis, dass selbst unscheinbare Daten in den Urteilsbegründungen zu Re-Identifikationen führen können. Die Überprüfung der Anonymisierung veröffentlichter Urteile ist notwendig, um auch in Zukunft die öffentliche Zugänglichkeit zu Gerichtsurteilen zu ermöglichen, aber gleichzeitig den bestmöglichen Persönlichkeitsschutz der Verfahrensbeteiligten und weiterer in Urteilen genannter Personen zu gewährleisten.

2. Rechtlicher Kontext: öffentliches Beschaffungsrecht

[4] Das öffentliche Beschaffungsrecht ist ein relativ «junges» Rechtsgebiet, dessen Entstehung auf die Verabschiedung des sog. GPA-Übereinkommen (engl. Government Procurement Agreement) unter der Ägide der Welthandelsorganisation im Jahr 1994 zurückzuführen ist. Die Grundlage dieses Rechtsgebiets bildet ein internationaler Rechtsakt, der in jeweilige Rechtssysteme der Mitgliedsländer zu implementieren ist. Die Schweiz implementierte das erste GPA Übereinkommen, indem auf der Bundesebene das Bundesgesetz über das öffentliche Beschaffungswesen vom 16. Dezember 1994 (aBöB)[3], die Verordnung über das öffentliche Beschaffungswesen vom 11. Dezember 1995 (aVöB)[4] und das Binnenmarktgesetz vom 6. Oktober 1995 (BGBM)[5] sowie auf der kantonalen Ebene die interkantonale Vereinbarung über das öffentliche Beschaffungswesen vom 25. November 1994/15.März 2001 (IvöB1994)[6] und Submissionsgesetze der einzelnen Kantone beschlossen und in Kraft gesetzt wurden.

[5] Aufgrund der Revision des GPA Übereinkommens im Jahr 2012, wurden die Mitgliedsstaaten zur Revision der nationalen Rechtsgrundlagen gezwungen. In der Schweiz mussten sowohl das aBöB und die aVöB als auch die IvöB1994 weitgehend überarbeitet werden. Das Ziel der Totalrevision bestand einerseits darin, das neue GPA Übereinkommen 2012 zu implementieren, und andererseits das öffentliche Beschaffungsrecht zwischen Bund und Kantonen zu harmonisieren. Die neuen Rechtsgrundlagen auf Bundesebene (BöB[7] und VöB[8]) traten am 1. Januar 2021 in Kraft. Die revidierte IvöB2019 wurde am 15. November 2019 verabschiedet – für das Inkrafttreten des Konkordats war erforderlich, dass diesem zumindest zwei Kantone beigetreten sind[9].

[6] Die Revision des öffentlichen Beschaffungsrechts hatte viele Neuerungen zufolge. Das Ziel der Revision war einerseits die Flexibilisierung dieses Rechtsgebiets, indem neue Rechtsinstrumente eingeführt resp. gesetzlich verankert wurden (vide: elektronische

---

[2] VOKINGER KERSTIN NOËLLE/MÜHLEMATTER URS JAKOB, Re-Identifikation von Gerichtsurteilen durch «Linkage» von Daten(banken). Eine empirische Analyse anhand von Beschwerden gegen (Preisfestsetzungs-)Verfügungen von Arzneimitteln vor Bundesgericht, Jusletter 2. September 2019.
[3] SR 172.056.1 (ausser Kraft).
[4] SR 172.056.11 (ausser Kraft).
[5] SR 943.02.
[6] SR 172.056.5; in einigen Kantonen noch immer in Kraft.
[7] Bundesgesetz über das öffentliche Beschaffungswesen (BöB) vom 21. Juni 2019, SR 172.056.1.
[8] Verordnung über das öffentliche Beschaffungswesen (VöB) vom 12. Februar 2020, SR 172.056.11.
[9] Der IVÖB2019 sind bis zum 1. Januar 2023 10 Kantone beigetreten: GR, SZ; AI, TG, AG, SO, VD, FR, LU, SH; der Kanton Bern ist der IvöB2019 nicht beigetreten, er wendet den Inhalt der Vereinbarung als eigenes kantonales Recht an; vgl. https://www.bpuk.ch/bpuk/konkordate/ivoeb/ivoeb-2019.

Offerteinreichung und elektronische Auktionen, Verankerung der von der Praxis entwickelten Rahmenverträge und des Dialogverfahrens). Andererseits konnten mit Hilfe der Revision die von der Praxis und Rechtsprechung erkannten Probleme gelöst werden, indem die Korruptionsprävention gestärkt, der Rechtsschutz ausgebaut und die Nachhaltigkeit sowie Transparenz weitgehend gefördert wurden.

[7] Die Stärkung der Transparenz im Rahmen des öffentlichen Beschaffungswesens war schon seit langer Zeit ein kontrovers diskutiertes Thema. Dies schlug sich auch im Rahmen des Revisionsverfahrens nieder, indem Reaktionen auf den Vorschlag, bestimmte Unterlagen vom Beschaffungsverfahren für die Dauer der Aufbewahrungsfrist der Geltung des Öffentlichkeitsgesetzes zu entziehen (vgl. Art. 49 Abs. 3 BöB in der vom Bundesrat vorgeschlagenen Fassung), besonders heftig und kritisch ausgefallen sind[10]. In der parlamentarischen Diskussion wurde diese Bestimmung dann ersatzlos gestrichen. Aktuell wird lediglich der Zugriff auf Unterlagen während der Dauer eines Vergabeverfahrens ausgeschlossen und im Rahmen eines Beschwerdeverfahrens eingeschränkt (vgl. Art. 57 BöB/IvöB2019).

[8] Das Transparenzprinzip gehörte schon nach dem alten Rechtsstand zu den Grundsätzen des öffentlichen Beschaffungswesens (vgl. Art. 1 Abs.1 Bst. A aBöB); nun infolge der Revision wurde dieses um eine Publikationspflicht erweitert. Auf Bundesebene waren Auftraggeberinnen bisher lediglich zur Veröffentlichung der Ausschreibungen und Zuschläge verpflichtet (vgl. Art. 24 Abs.1 a BöB). Nach der Gesetzesrevision wurde die Offenlegungspflicht um Abbruchverfügungen und Zuschläge, die in freihändiger Vergabe im Nichtstaatsvertragsbereich erteilt wurden, erweitert (vgl. Art. 48 Abs.1 BöB). Neu müssen sowohl Zuschläge im offenen und selektiven als auch freihändigen Verfahren oberhalb des Schwellenwertes öffentlich, d.h. auf der Plattform Simap, publiziert werden. (Unterschied bei Kantonen: Veröffentlichungspflicht gilt für Zuschläge im offenen/selektiven Verfahren und nur diejenigen freihändigen Verfahren, die in den Staatsvertragsbereich fallen, vgl. Art. 48 Abs. 1 IvöB2019).

[9] Es ist zu vermerken, dass das durchgeführte Experiment sich grösstenteils auf angefochtene Verfügungen *nach dem alten Recht* bezieht (Zuschläge erteilte bis zum 24.09.2021, siehe auch unten bei der Methodik).

[10] Der Inhalt der Publikation auf der Plattform Simap ist gesetzlich geregelt (vgl. Art. 48 i.V.m. 51 BöB/IvöB2019). Dabei ist zu beachten, dass im Rahmen einer Publikation nur zwei Subjekte offenbart werden dürfen, d.h. die öffentliche Auftraggeberin/Vergabestelle und die berücksichtigte Anbieterin (Zuschlagsempfängerin). Demgegenüber dürfen die Angaben den restlichen Mitbewerbern/Innen u.a. aus datenschutzrechtlichen sowie wettbewerbsrechtlichen bzw. wettbewerbsfördernden Gründen nicht genannt werden. Dies ist vor allem aus Sicht der Anonymität in allfälligen künftigen Beschwerdeverfahren von Relevanz: bei den Beschwerdeführern/Innen wird es sich in der Regel um diejenigen Mitbewerber/Mitbewerberinnen handeln, die keinen Zuschlag erhalten hatten oder die kein Angebot im Rahmen eines freihändigen Verfahrens einreichen durften (Prämisse des besonders schutzwürdigen Interesses). Bei der Beschwerdegegnerin wird sich um die ausschreibende Stelle, d.h. Auftraggeberin handeln. Die Zuschlagsempfängerin dürfte sich

[10] BBl 2017 1851

zwar am Beschwerdeverfahren mittels Intervention beteiligen; dies ist aber tendenziell selten der Fall. Aus simap-technischen Gründen wird jede Publikation mit einer eindeutigen Identifikationsnummer, der sog. «Meldungsnummer», versehen. Jedem Projekt wird dazu eine Projektnummer ("Projekt-ID") zugeteilt, die auf jeder Publikation, die das Verfahren betrifft, ersichtlich ist.

[11] Bestimmte Verfügungen, insbesondere die Zuschlagsverfügung, können angefochten werden (vgl. Art. 52 Abs.1 und 2 i.V. m. Art. 53 Abs. 1 BöB/IvöB2019). Für die Beschaffungen des Bundes ist als erste Instanz das Bundesverwaltungsgericht (vgl. Art. 52 Abs. 1 BöB)[11] und in bestimmten Fällen als zweite Instanz das Bundesgericht (e contrario Art. 83 Bst. f. BGG[12])[13] tätig. Gegen die Verfügungen der öffentlichen Auftraggeber/Innen auf der kantonalen Ebene kann eine Beschwerde an das entsprechende kantonale Verwaltungsgericht und als zweite Instanz an das Bundesgericht eingereicht werden (vgl. Art. 52 Abs. 1 und 2 IvöB2019; e contrario Art. 83 Bst. f. BGG) [14].

3. Methodik

[12] In diesem Kapitel wird die Methodik des Re-Identifizierungs-Versuchs beschrieben. Zunächst wurde mithilfe von Python Scripts[15] eine Datenbank von Schweizer Gerichtsurteilen erstellt. Die entsprechenden Daten konnten von der Webseite des Vereins Entscheidsuche[16] heruntergeladen werden. Die daraus resultierende Datenbank von Gerichtsurteilen wurde sodann nach dem Begriff "simap" durchsucht (case-insensitive: Gross- und Kleinschreibung wird nicht berücksichtigt). Darauf wurden diese Urteile mit regulären Ausdrücken[17] nach den Meldungsnummern und Projekt-IDs (fortan "Nummern") durchsucht.

[13] Neben der selbst erstellten Datenbank wurde die Plattform IntelliProcure[18] als Hilfsmittel zur Re-Identifikation verwendet. IntelliProcure lädt regelmässig die Daten der Simap Plattform herunter und bereitet diese auf. Auf Anfrage beim Betreiber wurde für die vorliegende Untersuchung ein Export aller dort erfassten Zuschläge (Stand 24.09.2021) zur Verfügung gestellt.[19] Dieser Export enthielt knapp 64'000 Einträge mit etwas über 14'000

---

[11] Vgl. dazu auch https://www.beschaffungswesen.ch/vergabeverfahren/rechtsschutz/beschwerdeverfahren-im-bund.
[12] Bundesgesetz über das Bundesgericht (Bundesgerichtsgesetz, BGG) vom 17. Juni 2005, SR 173.110.
[13] Vgl. dazu auch https://www.beschaffungswesen.ch/vergabeverfahren/rechtsschutz/beschwerde-ans-bundesgericht.
[14] Vgl. dazu auch https://www.beschaffungswesen.ch/vergabeverfahren/rechtsschutz/kantonales-beschwerdeverfahren.
[15] https://github.com/JoelNiklaus/SwissCourtRulingCorpus
[16] https://entscheidsuche.ch/docs
[17] Reguläre Ausdrücke: Eine reguläre Ausdrucksweise ist eine Schreibweise, die verwendet wird, um Muster in Zeichenketten zu finden und zu manipulieren. Es ist ähnlich wie ein Suchmuster, das Anwälte in Textdokumenten verwenden können, um spezifische Textpassagen zu finden oder zu ändern. Zum Beispiel können Sie eine reguläre Ausdrucksweise verwenden, um alle Telefonnummern in einem Dokument zu finden oder um alle Wörter in einem bestimmten Format zu ändern. https://de.wikipedia.org/wiki/Regul%C3%A4rer_Ausdruck
[18] https://intelliprocure.ch
[19] Sonstige Ausschreibungen waren nicht erhältlich. In der Analyse wurden daher nur Urteile untersucht, die sich auf Zuschläge beziehen.

verschiedenen Zuschlagsempfängerinnen und etwas über 2600 verschiedenen ausschreibenden Stellen. Jeder Eintrag enthält Daten wie den Titel des Projekts, die ausschreibende Stellen, die Zuschlagsempfängerin, die Nummern sowie die Preise der Zuschläge (siehe auch Art. 48 Abs. 6 i.V.m. 51 Abs. 3 BöB/IvöB2019).

[14] Im dritten Schritt der Analyse wurden nun die Daten des Exports nach den Nummern durchsucht, welche in den Gerichtsentscheiden gefunden wurden.[20] Durch dieses Vorgehen war es möglich, das jeweilige gerichtliche Beschwerdeverfahren mit einem konkreten Submissionsverfahren zu verknüpfen.

[15] Der gesamte Quellcode[21] wird veröffentlicht, damit die vorliegenden Ergebnisse einfach reproduziert werden können. Die einzelnen Daten werden aber aus Datenschutzgründen nicht veröffentlicht.[22]

4. Ergebnisse

[16] Die Tabelle 1 gibt eine Übersicht zu den genauen Zahlen pro Filterungsschritt. Es fällt auf, dass bedeutend mehr Entscheide mit dem Wort "simap" in Deutsch und Französisch gefunden werden konnten als beispielsweise im Vergleich in Italienisch. Diese Verteilung korreliert mit der Verteilung der entsprechenden Sprachgruppen in der Schweizer Wohnbevölkerung (62.6%, 22.9%, 8.2%)[23]. Französisch ist, im Vergleich zur Sprecherverteilung, allerdings übervertreten. Deutsch und Italienisch sind dagegen untervertreten.[24] Interessanterweise werden weit weniger Nummern in den französischen Urteilen im Vergleich zu den deutschsprachigen gefunden. In den kantonalen französischen Urteilen werden die Nummern mit einer Ausnahme nie genannt.
Die Verteilung der Anzahl Urteile, welche Nummern enthalten, korreliert sehr stark mit der Verteilung von re-identifizierten Urteilen (also den Urteilen, bei welchen ein Link zur IntelliProcure Datenbank hergestellt werden konnte). Die re-identifizierten Urteile stammen *fast ausschliesslich vom Bundesverwaltungsgericht* (alle französischen und italienischen Urteile und 91% der deutschen Urteile). Insgesamt stammen 36% der Urteile welche "simap" enthalten vom Bundesverwaltungsgericht, 91% der Urteile welche Nummern enthalten und 93% der re-identifizierten Urteile.

---

[20] Diese Strategie der Re-Identifizierung wurde schon früher in der einschlägigen Literatur erwähnt, vgl. beispielsweise DANIEL KETTIGER, Anonymisierung: Rechtliche Aspekte, in: Daniel Hürlimann/Daniel Kettiger (Hrsg.), Anonymisierung von Urteilen, Helbing und Lichtenhahn, Basel 2021, S. 21-30, Rz. 19 und 24.

[21] https://github.com/JoelNiklaus/SwissCourtDecisionReIdentification

[22] Die Universität Bern unterliegt dem Datenschutzrecht des Kantons Bern. Für die Re-Identifizierungsversuche im Projekt „Open Justice vs. Privacy" wurde bei der Datenschutzaufsichtsstelle des Kantons Bern ein Datenschutzkonzept hinterlegt.

[23] https://www.eda.admin.ch/aboutswitzerland/de/home/gesellschaft/sprachen/die-sprachen---fakten-und-zahlen.html

[24] Da keine Entscheide in Rätoromanisch gefunden wurden, wird diese Sprache fortan nicht mehr erwähnt.

Sprache	enthalten "simap"	enthalten Meldungsnummern oder Projekt-IDs	wurden re-identifiziert
Deutsch	557 (54.8%)	274 (80.6%)	220 (81.2%)
Französisch	414 (40.7%)	31 (9.1%)	28 (10.3%)
Italienisch	46 (4.5%)	35 (10.3%)	23 (8.5%)
Alle	1017 (100%)	340 (100%)	271 (100%)

Tabelle 1: Anzahl betroffene Urteile nach Sprache und Filterungsschritt. Die zweite Kolonne beschreibt die Anzahl Urteile, welche den Suchterm "simap" enthalten. Die dritte Kolonne beschreibt die Anzahl Urteile, welche eine Meldungsnummer oder Projekt-ID enthalten. Die letzte Kolonne beschreibt die Anzahl Urteile, bei welchen die gefundene Meldungsnummer oder Projekt-ID mit einer entsprechenden Nummer aus den Zuschlägen in Verbindung gebracht werden konnte.

[17] In Abbildung 1 ist die Preisverteilung der Zuschläge aus den re-identifizierten Urteilen ersichtlich (jeder Punkt stellt ein re-identifiziertes Urteil dar). Es gibt sehr wenige Zuschläge über 150 Millionen CHF und beim Grossteil der Zuschläge handelt es sich um Beträge unter 50 Millionen CHF. Es konnten jedoch auch einige sehr grosse Beträge im Bereich um 250 Millionen CHF ausfindig gemacht werden. Hier handelt es sich jedoch vermutlich um Rahmenverträge, wo nicht immer alle Leistungen bezogen werden.

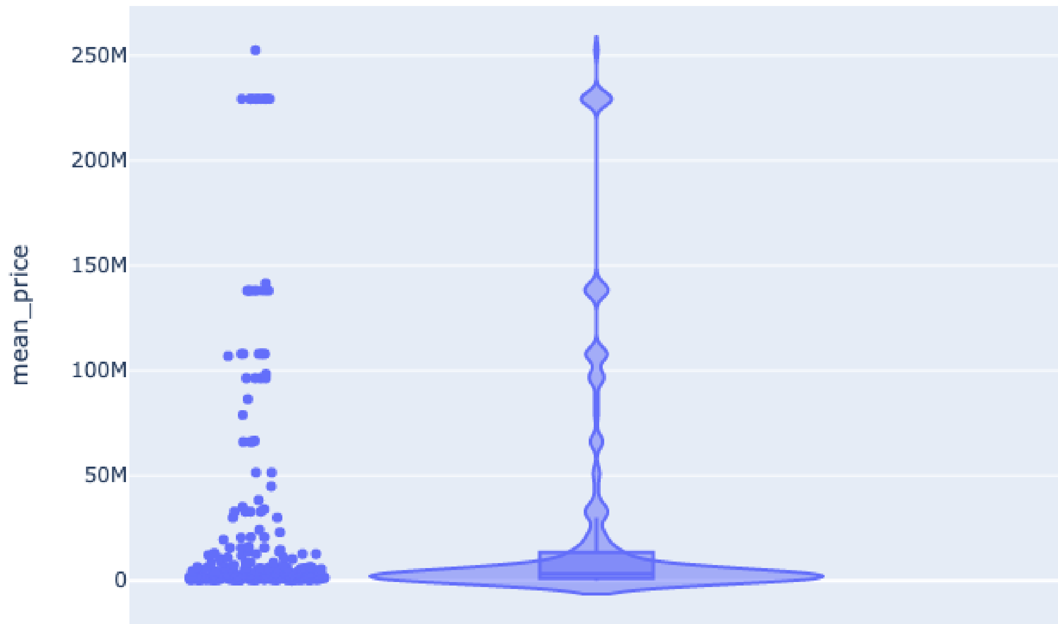


Abbildung 1: Preisverteilung (in CHF) der Zuschläge aus den re-identifizierten Urteilen.

5. Würdigung

[18] Die Untersuchung hat zweierlei gezeigt: Zum einen war es durch das angewandte Verfahren des String-Matchings nicht möglich, die Beschwerdeführer zu re-identifizieren. Einzig die Beschwerdegegner sind mit hoher Wahrscheinlichkeit identifizierbar, da es sich hierbei mutmasslich in allen Fällen um die Vergabestelle handeln wird. Aus datenschutzrechtlicher Sicht ist dies unproblematisch, da die Vergabestelle – als öffentliche

Körperschaft – notorisch keinen Schutz «persönlicher» Daten beanspruchen kann. Die Vergabestelle steht auch nicht im Wettbewerb und bedarf somit keines wettbewerbsrechtlichen Schutzes.

[19] Zum anderen lässt sich mittels der angewandten Methode aber in einer Vielzahl der Fälle zuordnen, bei welchen Submissionen der Rechtsweg beschritten worden ist. Die Veröffentlichung dieser Information ist gesetzlich nicht vorgesehen. So ist vergaberechtlich nur geregelt, dass die Zuschlagsempfängerin mit der Publikation des Zuschlags offenbart wird (vgl. Art. 48 Abs. 6 i.V.m. 51 Abs. 3 BöB/IvöB2019), nicht aber, wenn sie sich beispielsweise als Zuschlagsempfänger/In im Beschwerdeverfahren beteiligt. Ohne hinreichende gesetzliche Rechtfertigung erscheint fraglich, ob eine Re-Identifikation in einem Beschwerdeurteil den Absichten des Gesetzgebers entspricht.

[20] Zusammengefasst konnten in der vorgenommenen Untersuchung zwar keine gravierenden Datenschutzlücken aufgefunden werden, die mittels modernen Re-Identifikationsstrategien ausgenutzt werden könnten; dennoch war und ist es möglich, mittels den publizierten Meldungsnummern resp. Projekt-ID aus publizierten Urteilen zusätzliche Erkenntnisse zu gewinnen, die vom Gesetzgeber so nicht vorgesehen sind. Die Publikation von nicht anonymisierten Meldungsnummer resp. Projekt-ID in Gerichtsurteilen hat im Grunde keinen juristischen oder anderweitigen Mehrwert und es wäre daher wünschenswert, wenn die Gerichte ihre diesbezügliche Praxis überdenken würden. So könnte auch diese – wenn auch kleine – Lücke der Urteilsanonymisierung mit verhältnismässig geringem Aufwand geschlossen werden.

[21] Allgemein werden sich Gerichte in Zukunft vermehrt Gedanken betreffend die in den Urteilen angewandten Anonymisierungsregeln machen müssen. Auch vermeintlich unbedeutende Identifikationsmerkmale[25], wie im vorliegenden Beispiel die Meldungsnummern resp. Projekt-ID, können mit neuen Identifikationsstrategien zu skalierbaren Re-Identifizierungen führen, indem Zweitquellen zu Hilfe gezogen werden.

MSc Joel Niklaus ist Doktorand am Institut für Informatik der Universität Bern.
MLaw Magda Chodup ist Doktorandin am Kompetenzzentrum für Public Management (KPM) der Universität Bern.
MLaw Thomas Lüthi ist praktizierender Rechtsanwalt und war wissenschaftlicher Mitarbeiter an der Forschungsstelle für Digitale Nachhaltigkeit an der Universität Bern.
Mag. rer. publ. Daniel Kettiger ist praktizierender Rechtsanwalt sowie Projektleiter und Justizforscher am Kompetenzzentrum für Public Management (KPM) der Universität Bern.

Dieser Beitrag entstand im Rahmen des Nationalen Forschungsprogramms NFP 77 «Digitale Transformation» mit Unterstützung des Schweizerischen Nationalfonds (SNF).

---

[25] Eine Übersicht über gängige Identifikationsmerkmale findet sich bei KETTIGER (Fn. 20), Rz. 19-24.

Declaration of consent

on the basis of Article 30 of the RSL Phil.-nat. 18

Name/First Name:

Registration Number:

Study program:

Bachelor ☐ Master ☐ Dissertation ☐

Title of the thesis:

Supervisor:

I declare herewith that this thesis is my own work and that I have not used any sources other than those stated. I have indicated the adoption of quotations as well as thoughts taken from other authors as such in the thesis. I am aware that the Senate pursuant to Article 36 paragraph 1 litera r of the University Act of 5 September, 1996 is authorized to revoke the title awarded on the basis of this thesis.

For the purposes of evaluation and verification of compliance with the declaration of originality and the regulations governing plagiarism, I hereby grant the University of Bern the right to process my personal data and to perform the acts of use this requires, in particular, to reproduce the written thesis and to store it permanently in a database, and to use said database, or to make said database available, to enable comparison with future theses submitted by others.

Place/Date

Signature

Joel Niklaus

PhD Student · Natural Language Processing

Schwarztorstrasse 108, 3008 Bern, Switzerland | Swiss | 05/09/1995
(+41) 78 664 46 59 | joel@niklaus.ai | niklaus.ai | joelniklaus | joelniklaus |
@joelniklaus | joelniklaus

"The secret of getting ahead is getting started." – Mark Twain

Education

PhD Student in Computer Science — *Bern, Switzerland*
University of Bern — *Jul. 2020 - Jan. 2024*
NRP 77 "Open Justice vs. Privacy"
Supervisors: PD Dr. Matthias Stürmer, Prof. Dr. Paolo Favaro, Prof. Dr. Thomas Myrach, Prof. Dr. Daniel E. Ho, Dr. Ilias Chalkidis
Supervision of Bachelor and Master theses
Research Visit at RegLab, **Stanford University**, CA, US (Oct. 2022 –– Apr. 2023) advised by Prof. Dr. Daniel Ho and collaborating with the Center for Research on Foundation Models (CRFM)
Thesis: Decoding Legalese Without Borders: Multilingual Evaluation of Language Models on Long Legal Texts

Master of Science in Computer Science – insigni cum laude — *Bern, Switzerland*
Universities of Bern, Neuchâtel and Fribourg — *Sep. 2018 - Dec. 2019*
Major 90: Computer Science with a specialization in Data Science (Grade average: **5.50/6**)
Minor 30: Sports Science (Grade average: 4.96/6) and various additional courses
Exchange Semester at **Nanyang Technological University**, Singapore (Spring 2019) (Grade average: 6/6)
Thesis: JassTheRipper – A High-Human Artificial Intelligence for the Swiss Card Game Jass (Grade: **6/6**)

Bachelor of Science in Computer Science – magna cum laude — *Bern, Switzerland*
University of Bern — *Sep. 2013 - Jun. 2018*
Major 90: Computer Science (Grade average: **5.41/6**)
Minor 60: Sports Science (Grade average: 5.09/6)
Minor 30: Mathematics
Extracurricular Minor 15: Business Administration
Exchange Semester at **University of Exeter**, England (Spring 2017) (Grade average: 5/6)
Thesis: Machine Learning for Indoor Positioning (Grade **6/6**)

Swiss Matura — *Winterthur, Switzerland*
Kantonsschule Rychenberg Winterthur — *Aug. 2007 - Jul. 2013*
Bilingual (German and English) with a focus on ancient languages (Greek & Latin) (Grade average: 4.96/6)
Thesis: scholarcoach.ch – A Tutoring Portal (Grade **6/6**)

Experience

Research

AI Resident — *Mountain View, United States*
(Google) X — *Jul. 2023 - Nov. 2023*
Pretraining and Instruction Tuning LLMs in the Legal Domain on large TPU clusters.

Data Science Intern — *Zug, Switzerland*
Thomson Reuters Labs — *Mar. 2022 - Aug. 2022*
Cheaply Pretraining Longformers for the Legal Domain

AI Consultant — *Bern, Switzerland*
Niklaus.ai — *Apr. 2022 - present*
I consult companies on topics in AI, and especially in NLP.
Selected customers: Darrow AI Ltd. (Legal Judgment Prediction in the US)

Project Lead
Bern, Switzerland

BERN UNIVERSITY OF APPLIED SCIENCES *Jan. 2022 - Nov. 2022*

I managed a team of 6 people and lead the project Swiss Long Legal BERT for the Swiss Federal Supreme Court. We pretrained a host of domain specific legal models that improved recall of the anonymization system from 83% to 93%

Data Scientist
Bern, Switzerland

ARTORG CENTER FOR BIOMEDICAL ENGINEERING RESEARCH *Mar. 2020 - Sep. 2020*

Maintenance and extension of tele-health applications
Development of cognitive models from existing data
Development of algorithms for dynamic difficulty adjustments for tele-apps
Analysis and presentation of complex datasets

SOFTWARE DEVELOPMENT

Independent Software Developer
Bern, Switzerland

JOELNIKLAUS.CH *Aug. 2012 - present*

Successful implementation of various orders for customers (`https://joelniklaus.ch`)
Selected customers: University of Bern, University for senior citizens Bern, Claire & George, Freies Gymnasium Bern, KS Rychenberg

Web Developer
Bern, Switzerland

FASTFORWARD WEBSOLUTIONS GMBH *Sep. 2017 - Jun. 2018*

User interface development with Aurelia for a calculator of supplementary services (old age insurance) for Pro-Senectute Switzerland (`https://prosenectute.ch/de/dienstleistungen/beratung/finanzen/el-rechner.html`)

TEACHING

Lecturer for Transformers Module
Bern, Switzerland

UNIVERSITY OF BERN *Nov. 2023 - present*

I teach transformers to students in the CAS Natural Language Processing.

Lecturer for Natural Language Processing and AI
Bern, Switzerland

BERN UNIVERSITY OF APPLIED SCIENCES *Nov. 2023 - present*

I co-lecture about LLMs in three applied modules on AI.

Lecturer for Software Engineering
Bern, Switzerland

BERN UNIVERSITY OF APPLIED SCIENCES *Aug. 2021 - Jul. 2023*

I planned and conducted the technical part of the module Software Engineering (Code Management, DevOps, Web Technologies, etc.).

Lecturer for Natural Language Processing Seminar
Bern, Switzerland

UNIVERSITY OF BERN *Aug. 2021 - Jul. 2022*

I lead, planned and executed the entire Seminar (20 BSc/MSc students in total). I supervised various individual and group projects.

High School Teacher for Computer Science & Robotics
Bern, Switzerland

FREIES GYMNASIUM BERN *Aug. 2019 - Jul. 2020*

I assembled the coursework material for the brand new course computer science as the first teacher; I conducted difficult conversations with parents and students.

High School Teacher for Information and Communication Technology
Winterthur, Switzerland

KANTONSSCHULE RYCHENBERG WINTERTHUR *Feb. 2016 - Aug. 2020*

Already during bachelors I started teaching students in applied computer skills; I supervised various Matura theses

Teaching Assistant for "Programming Languages" (MSc), "Software Engineering" (BSc) and "Computer Networks" (BSc)
Bern, Switzerland

UNIVERSITY OF BERN *Sep. 2015 - Jun. 2021*

Supervision of students
Organisation and execution of practice hours
Composition and Correction of lab assignments and exams

Awards

2021	**Research Grant (60'000 CHF)**, Pretraining Swiss Long Legal BERT, Swiss Federal Supreme Court	*Bern, Switzerland*
2019	**Nominee**, Best Master Thesis, University of Fribourg, Department of Computer Science	*Fribourg, Switzerland*
2017	**Winner**, Excellent Work in Computer Science (Bachelor Thesis), Joint Alumni Association of the Universities Bern, Neuchâtel & Fribourg	*Bern, Switzerland*
2017	**2nd Place**, JassChallenge, Zühlke (Large Swiss Software Consulting Company)	*Bern, Switzerland*
2016	**2nd Place**, RoboChallenge, Zühlke	*Bern, Switzerland*
2013	**Nominee**, Best Matura Thesis, Kantonsschule Rychenberg	*Winterthur, Switzerland*

Projects

2019 – 2020	**Identifai**, Building an admission system with face recognition
2017 – 2018	**DeinLohn**, Building a platform for disclosing and comparing salaries across jobs and industries (`https://deinlohn.ch`)
2014 – 2017	**ProjectEuler**, Solving challenging mathematical problems with computer programming (`https://projecteuler.net/profile/JoelNiklaus.png`)

Presentations

NLP Circle @ Bern University of Applied Sciences *Bern, online*
NEW POSSIBILITIES IN LEGAL NLP: PRETRAINING AND EVALUATING LLMS WITH MULTILEGALPILE, LEXTREME, AND SCALE *Jul. 2023*
Invited Talk

NLP Group Lunch @ Stanford University *Stanford, online*
NEW POSSIBILITIES IN LEGAL NLP: PRETRAINING AND EVALUATING LLMS WITH MULTILEGALPILE, LEXTREME, AND SCALE *Jun. 2023*
Invited Talk

E2 by Ari Massoudi *online*
MULTILEGALPILE AND LEXTREME: PRETRAINING AND EVALUATING MULTILINGUAL LEGAL LANGUAGE MODELS *Jun. 2023*
Interview

Alumni Talk @ Thomson Reuters Labs *Zug, online*
MULTILEGALPILE AND LEXTREME: PRETRAINING AND EVALUATING MULTILINGUAL LEGAL LANGUAGE MODELS *Jun. 2023*
Invited Talk

Pizza Seminar @ Bern University of Applied Sciences *Bern, online*
LARGE LANGUAGE MODELS *May 2023*
Invited Talk

CAS Digitale Transformation @ University of Bern *Bern, online*
LARGE LANGUAGE MODELS *Apr. 2023*
Invited Talk

Digital Snack @ Department of Business, Bern University of Applied Sciences *Bern, online*
MÖGLICHKEITEN UND GRENZEN MODERNER NATURAL LANGUAGE PROCESSING (NLP) TECHNOLOGIEN *Sep. 2022*
Invited Talk

TechTalk @ ICT Warrior Academy, Armed Forces Command Support Organisation, Swiss Army *Bern, online*
KÜNSTLICHE INTELLIGENZ MIT DER MENSCHLICHEN SPRACHE *Apr. 2022*
Invited Talk

CAS Advanced Machine Learning: Introduction to NLP *Bern, online*
PRETRAINED LANGUAGE MODELS *Feb. 2022*
Invited Talk

AI and Law Conference by recode.law *online*

SWISS-JUDGMENT-PREDICTION: A MULTILINGUAL LEGAL JUDGMENT PREDICTION BENCHMARK *Jan. 2022*

Invited Talk

International Conference of Artificial Intelligence and Law (ICAIL) 2021 *online*

ESRA: AN END-TO-END SYSTEM FOR RE-IDENTIFICATION AND ANONYMIZATION FOR SWISS COURT RULINGS *Jun. 2021*

Doctoral Consortium presentation

Advanced Language Processing Winter School (ALPS) 2021 *online*

ESRA: AN END-TO-END SYSTEM FOR RE-IDENTIFICATION AND ANONYMIZATION FOR SWISS COURT RULINGS *Jan. 2021*

Poster presentation

PyData Zurich *Zurich, Switzerland*

PEDIATRIC BONEAGE PREDICTION *Nov. 2018*

Predicting the age of children based on hand xray images with deep neural networks (`https://meetup.com/en-AU/PyData-Zurich/events/255193573`)