How Can We Synthesize High-Quality Pretraining Data?

A systematic study of prompt design, generator model, and source data

Joel Niklaus[1] · Atsuki Yamaguchi[2] · Michal Štefánik[3] · Guilherme Penedo[1] · Hynek Kydlíček[1] · Elie Bakouch[1] · Lewis Tunstall[1] · Edward Beeching[1] · Thibaud Frere[1] · Colin Raffel[1] · Leandro von Werra[1] · Thomas Wolf[1]

[1]Hugging Face · [2]University of Sheffield · [3]National Institute of Informatics, Japan | **COLM 2026**, San Francisco · Poster F-4-126, Franciscan, Wednesday afternoon

Paper
arXiv 2604.13977

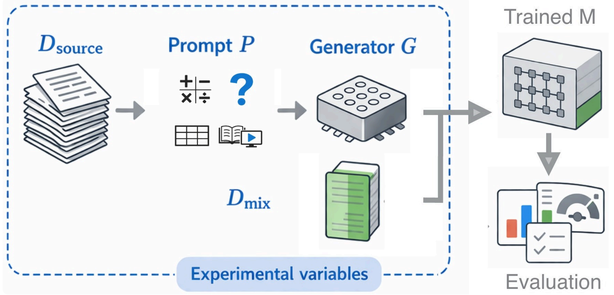
Dataset
HuggingFaceFW/finephrase

Prompts + code
huggingface/finephrase

The question

Crawlable web text is running out, so labs now rephrase the web with LLMs at trillion-token scale. But every prior method (WRAP, Nemotron-CC, BeyondWeb, REWIRE) reports gains against its own baseline. Nobody had compared **which prompt**, **which generator**, and **which source data** matter, under one controlled setup.

What we did



Rephrase ~10.5B source tokens with a prompt *P* and generator *G*, mix 50/50 with original web data, pretrain a 1.2B model from scratch on 21B tokens, evaluate on 12 benchmarks (3-shot cloze, macro average).

90
rephrasing configurations

333
train-and-evaluate runs

>1T
synthetic tokens generated

12.7
GPU-years of generation

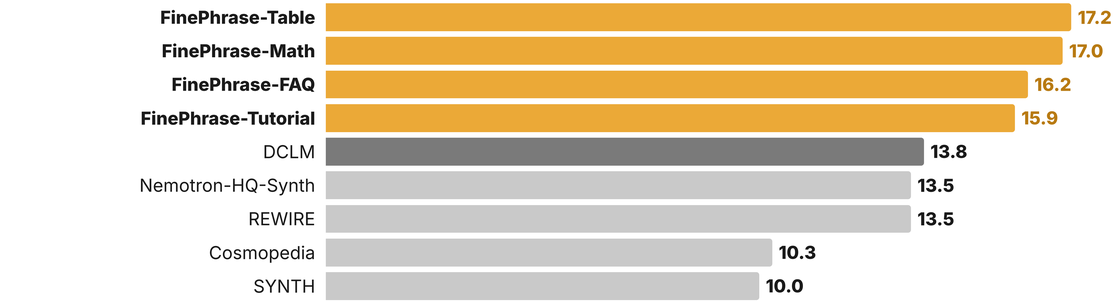
The FinePhrase recipe

- **Generator:** `SmolLM2-1.7B-Instruct`
- **Prompts:** Table, Math, FAQ, Tutorial
- **Source:** FineWeb (educational subset)
- **Training:** mix with strong original web data (DCLM or FineWeb-HQ)

The prompt matters more than the model.

Rephrasing the web into **tables, math problems, FAQs and tutorials** with a **1.7B** model beats rewrites from 70B models and curated web data, at up to **30× lower** generation cost.

FinePhrase beats every existing synthetic dataset



Macro average over 12 benchmarks (×100) of a 1.2B model trained on 21B tokens. FinePhrase-Table: +3.4 over DCLM, +3.6 over Nemotron-HQ-Synth. FinePhrase wins on knowledge and reading comprehension (ARC, SQuAD, DROP); DCLM keeps a small edge on PIQA and HellaSwag, which is why we always mix.

486B
tokens, 1.35B samples, openly released

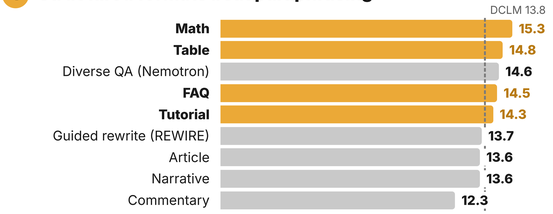
30×
more tokens per GPU-hour than REWIRE

14.7K
H100 GPU-hours total (612 GPU-days)

9.2K
tokens / s / GPU with speculative decoding

DATASET	GENERATOR	TOKENS	GPU-HRS	TOK / GPU-HR
Cosmopedia	Mixtral 8×7B	25B	>10K	<2.5M
SYNTH	fine-tuned	80B	4K	20M
REWIRE	Llama 3.3 70B	400B	~352K	~1.1M
FinePhrase	**SmolLM2 1.7B**	**486B**	**~14.7K**	**~33.1M**

1 Structured formats beat paraphrasing

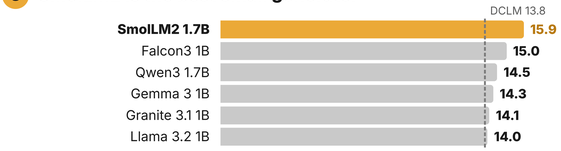


Same generator (Gemma 3 1B), same source. Pedagogical restructuring wins. Free-form rewrites (article, narrative, commentary) do not even beat DCLM.

2 Generators above 1B do not help

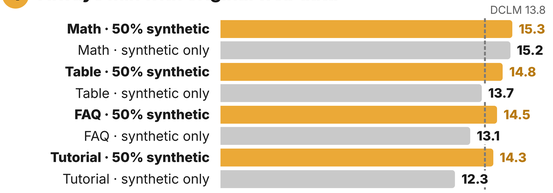
Gemma 3 family, Math prompt. 270M is too small; from 1B up, quality is flat while cost explodes (27B guided rewrite: 11K GPU-hours vs 205 for FinePhrase-Table).

3 SmolLM2 is the best small generator



Tutorial prompt, ~1B instruct models. Messier, more diverse outputs beat polished ones: Qwen3 follows the format 100% of the time but collapses onto a few templates; SmolLM2 only 68%, yet trains better.

4 Always mix with original web data



Synthetic-only training loses commonsense and NLU. Which data you mix in matters more than the source quality: DCLM or FineWeb-HQ 14.3 vs Cosmopedia 13.0.