White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/247684/

Version: Preprint

Preprint:
Niklaus, J., Yamaguchi, A., Štefánik, M. et al. (Submitted: 2026) How can we synthesize high-quality pretraining data? A systematic study of prompt design, generator model, and source data. [Preprint - arXiv] (Submitted)

https://doi.org/10.48550/arxiv.2604.13977

© 2026 The Author(s). This preprint is made available under a Creative Commons Attribution 4.0 International License. (https://creativecommons.org/licenses/by/4.0/)

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

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 Emanuel Beeching**[1]
Thibaud Frere[1] **Colin Raffel**[1] **Leandro von Werra**[1] **Thomas Wolf**[1]
[1]Hugging Face [2]University of Sheffield, United Kingdom
[3]National Institute of Informatics, Japan

arXiv:2604.13977v1 [cs.CL] 15 Apr 2026

Pipeline code: https://github.com/huggingface/finephrase

FINEPHRASE dataset: https://hf.co/datasets/HuggingFaceFW/finephrase

Abstract

Synthetic data is a standard component in training large language models, yet systematic comparisons across design dimensions, including rephrasing strategy, generator model, and source data, remain absent. We conduct extensive controlled experiments, generating over one trillion tokens, to identify critical factors in rephrasing web text into synthetic pretraining data. Our results reveal that structured output formats, such as tables, math problems, FAQs, and tutorials, consistently outperform both curated web baselines and prior synthetic methods. Notably, increasing the size of the generator model beyond 1B parameters provides no additional benefit. Our analysis also demonstrates that the selection of the original data used for mixing substantially influences performance. By applying our findings, we develop **FINEPHRASE**, a 486-billion-token open dataset of rephrased web text. We show that FINEPHRASE outperforms all existing synthetic data baselines while reducing generation costs by up to 30 times. We provide the dataset, all prompts, and the generation framework to the research community.

1 Introduction

The pretraining data landscape of large language models (LLMs) has undergone three paradigm shifts. First, models like GPT (Radford et al., 2018) and BERT (Devlin et al., 2019) relied on curated corpora including BookCorpus (Zhu et al., 2015) and Wikipedia. Scaling initiatives like C4 (Raffel et al., 2020) and The Pile (Gao et al., 2020) later expanded available datasets into hundreds of gigabytes of web text. This expansion followed scaling laws, which indicate that increased pretraining duration and data volume consistently improve downstream performance (Kaplan et al., 2020; Hoffmann et al., 2022). Efforts such as FineWeb (Penedo et al., 2024a) and DCLM (Li et al., 2024) further brought pretraining data to the trillion-token scale, covering most of the crawlable web. As the volume of available web text plateaus, the focus of the field has moved from quantity to quality: modern development pipelines now employ neural classifiers to filter for educational content (Gunasekar et al., 2023; Penedo et al., 2023) and analyze data repetition extensively (Muennighoff et al., 2023).

Synthetic data constitutes the latest step in this progression, appearing as a solution to the exhaustion of usable real-world data. Rather than merely filtering web text, recent work uses LLMs to rephrase text into variants that preserve information while optimizing the presentation (Maini et al., 2024; Kiyomaru et al., 2026, *inter alia.*). The community has adopted this approach at scale: Su et al. (2025) rephrased two trillion tokens for Nemotron-CC, and several recent LLMs utilized hundreds of billions of synthetic tokens during pretraining (Abdin et al., 2024; NVIDIA et al., 2025; Yang et al., 2025a; Singh et al., 2026).

However, the design space for synthetic data generation is expansive and currently lacks comprehensive assessment. Prior work has explored individual approaches in isolation. For

instance, WRAP (Maini et al., 2024) focuses on stylistic rewriting; Nemotron-CC (Su et al., 2025) extracts QA pairs and knowledge lists; BeyondWeb (DatologyAI et al., 2025) proposes continuation and summarization rephrasing; and REWIRE (Nguyen et al., 2025) targets the guided transformation of low-quality documents. While each study reports gains relative to a specific baseline, no systematic comparison exists across these methods. Furthermore, the interaction between key design choices, such as the choice of generator models and types of source data, remains unexplored, leaving researchers without a clear strategy to maximize the utility of synthetic data for pretraining.

To address these limitations, we conduct a controlled analysis across three axes: **the rephrasing strategy**, **the choice of generator models**, and **the selection of source and mix-in data**. First, we investigate rephrasing strategy by comparing established rephrasing prompts from prior work against four new pedagogical structured formats: tutorials, FAQs, tables, and math reformulations. These formats leverage the benefits of structured signals and educational content for the learning process of the language model (Gunasekar et al., 2023; Galke et al., 2024; Cheng et al., 2024; Yamaguchi et al., 2026). Second, to determine the influence of generator models, we test six model families with sizes ranging from 135M to 27B parameters. Finally, we evaluate how the quality of the source text and the integration of original web data (i.e., mix-in data) affect outcomes. This clarifies whether the source quality or the mix-in dataset selection serves as the primary driver of performance.

Our main contributions are as follows:

- We present a systematic comparison of synthetic pretraining data across the strategy of rephrasing, the generator model, and the source data (§4).
- We show that the choice of rephrasing prompt is the dominant factor for downstream performance; large models (10B+) are not always necessary to perform effective rephrasing; the mix-in dataset is more critical than the source data; and output diversity is more important than formatting consistency (§4 & §5).
- We create and release **FINEPHRASE**, a 486B token dataset and associated generation framework. Using SmolLM2 1.7B (Allal et al., 2025), it outperforms all existing synthetic data baselines at up to a 30-fold reduction in generation cost relative to previous methods (§6).

2 A Systematic Framework for Pretraining Data Synthesis

2.1 Problem Setting

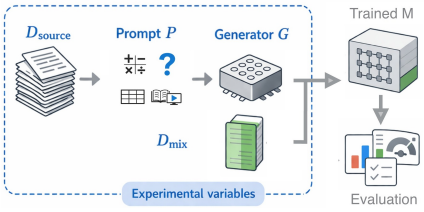


Figure 1: Overview of our experimental methodology.

To address the scaling limits of crawlable web data, we systematically evaluate methods for generating synthetic pretraining data by rephrasing web corpora. Unlike prior work evaluating isolated rephrasing methodologies (Maini et al., 2024; Su et al., 2025; Nguyen et al., 2025, *inter alia.*), we conduct a controlled ablation of the key components within the generation pipeline (Figure 1). The goal is a reproducible experimental framework for synthetic data generation that reduces reliance on intuition and replaces ad hoc practice with a rigorously designed pipeline. By isolating prompt design, model scale, and data composition, this research identifies configurations that maximize performance while minimizing compute allocation.

We construct the tested pretraining configurations from a rephrasing prompt $\mathcal{P}$ (§2.2), an instruction-tuned generator model $\mathcal{G}$ (§2.3), and both a mix-in $\mathcal{D}_{\text{mix}}$ and a source dataset $\mathcal{D}_{\text{source}}$ (§2.4). We validate the effectiveness of each configuration by training an LM $\mathcal{M}$ from scratch on a mixture of rephrased samples and $\mathcal{D}_{\text{mix}}$ for 21B tokens. Success is defined by the aggregate performance of the trained model across evaluation benchmarks spanning factual knowledge, reading comprehension, and reasoning (§3). This framework enables the identification of the optimal generation configuration, which we employ to create FINEPHRASE, a synthetic dataset consisting of 486B tokens.

Table 1: Excerpts of prompt templates for structured pedagogical rephrasing. Full templates are available in Appendix B.4.

Format	Instructional Content Example
`math`	"Rewrite the document to create a mathematical word problem... Provide a step-by-step solution that shows the calculation process clearly."
`faq`	"Rewrite the document as a comprehensive FAQ... Order questions logically from foundational to advanced... Ensure the FAQ works as a standalone document."
`table`	"Rewrite the document as a structured table that organizes the key information, then generate one question-answer pair based on the table."
`tutorial`	"Rewrite the document as a clear, step-by-step tutorial... Use numbered steps or bullet points where appropriate to enhance clarity."

2.2 Rephrasing Strategy

The first axis of investigation focuses on the strategy of rephrasing $\mathcal{P}$, which defines the structural transformation of raw web documents into synthetic pretraining samples. We investigate rephrasing strategies by comparing established prompts from prior work against four new structured pedagogical formats.

Established Prompts. We evaluate eight existing rephrasing strategies that vary by design and objective. The analysis incorporates five prompts from the Nemotron-HQ-Synth framework (Su et al., 2025): (i) *Diverse QA Pairs* generates questions across various cognitive levels with accompanying answers; (ii) *Extract Knowledge* restructures factual content into passages resembling textbook style; (iii) *Distill* produces a condensed version that retains essential information and technical terminology; (iv) *Wikipedia* paraphrases the text using formal English characteristic of Wikipedia; and (v) *Knowledge List* organizes key information into a list. Furthermore, we include the *Guided Rewrite* prompt from REWIRE (Nguyen et al., 2025). This approach employs an expert persona to perform meta-reasoning and planning before rewriting content from the voice of the original author. Finally, we examine two baseline strategies from BeyondWeb (DatologyAI et al., 2025): *Continue*, which generates a text extension matching the existing style, and *Summarize*, which creates a standalone summary of the document. We include all baseline prompt templates in Appendix B.

Pedagogical Structured Prompts. We evaluate four formats in Table 1 to determine if transforming content into pedagogical frameworks enhances the utility of the resulting tokens. We define a pedagogical format as a data structure that organizes information into discrete, logical, and queryable units. Specifically, we consider four formats: `math`, `faq`, `table`, and `tutorial`. This approach follows the observation of Gunasekar et al. (2023) that educational clarity drives model performance. By utilizing these formats, the synthesis process converts flat web text into structured signals. Prior work suggests these signals often benefit language model pretraining (Galke et al., 2024; Cheng et al., 2024; Yamaguchi et al., 2026). We keep the generator model $\mathcal{G}$ and the source dataset $\mathcal{D}_{\text{source}}$ constant when studying prompt choices to ensure that we have isolated the influence of the prompt.

2.3 Generator Model

Our second axis investigates how the generator model $\mathcal{G}$ influences synthetic data characteristics. We define this axis through two dimensions: parameter scale and architectural family.

Parameter Scale. We evaluate the influence of model size on the quality of the rephrased data. To isolate the effect of scale, we utilize Gemma 3 (Gemma Team et al., 2025) in five sizes (270M, 1B, 4B, 12B, 27B) and SmolLM2 in three sizes (135M, 360M, 1.7B). This range identifies the minimum capability threshold for synthetic data generation as well as the saturation point where additional model capacity fails to improve the utility of synthetic tokens.

Architectural Families. The underlying architecture provides a second dimension for evaluating generator influence. We examine six distinct open-source model families at a fixed scale of approximately 1B parameters: Gemma 3, Llama 3.2 (Grattafiori et al., 2024), Qwen 3 (Yang et al., 2025a), Granite 3.1 (IBM Granite Team, 2024), Falcon 3 (Falcon-LLM

Team, 2024), and SmolLM2. By evaluating models across these diverse families, we can assess the sensitivity of synthetic output to specific architectural variations.

2.4 Mix-in and Source Data

Our third axis explores the influence of data composition on pretraining success. This analysis focuses on the interaction between the source dataset $\mathcal{D}_{\text{source}}$, which provides the content for rephrasing, and the mix-in dataset $\mathcal{D}_{\text{mix}}$, which consists of the original web tokens used during pretraining.

Necessity of Mixing. We first evaluate whether training exclusively on synthetic data suffices to achieve optimal performance or whether original web tokens remain indispensable.

Selection of Mix-in Data. We compare several candidates for the mix-in component, including DCLM, Cosmopedia, and two tiers (high/low quality) of FineWeb (Penedo et al., 2024a). The objective is to identify which original corpus provides the most effective complementary signal to synthetic data.

Source Data Quality and Up-cycling. We apply identical rephrasing strategies to the same four source datasets: DCLM, Cosmopedia (Ben Allal et al., 2024), and the two FineWeb subsets. By fixing the mix-in dataset while varying the quality of $\mathcal{D}_{\text{source}}$, we assess whether the synthesis process can up-cycle noisy web text into high-utility training tokens.

3 Experimental Setup

Model Architecture and Tokenizer. Across all experiments, we train models with a 1.2B-parameter Qwen 2 architecture (Yang et al., 2024) featuring 28 layers, a hidden dimension of 2048, an intermediate size of 6144, and attention mechanisms with 16 query heads and 8 key-value heads through grouped-query attention (Ainslie et al., 2023). For tokenization, we employ the Llama 3.2 tokenizer with a vocabulary size of 128,256.

Data. Each experiment involves rephrasing approximately 10.5 billion tokens. If the completion count falls below the target because of the specific prompt, the process incorporates repeated samples to ensure a consistent volume across configurations. The default setup employs a 50/50 mixture of synthetic and original tokens, utilizing Gemma 3 1B as the generator $\mathcal{G}$ and a high-quality subset of FineWeb (FWHQ) for both the source $\mathcal{D}_{\text{source}}$ and the mix-in $\mathcal{D}_{\text{mix}}$. The FineWeb-Edu classifier scores samples from 0 to 5 via Llama-3-70B-Instruct based on their educational quality. We regard scores of 4 or 5 as HQ and scores of 0 or 1 as low-quality (LQ) to evaluate the influence of the quality of the seed data.

Hyperparameters. We pretrain all models using 64 NVIDIA H100 GPUs. The global batch size is 512 samples with a sequence length of 4096 tokens. Each model is trained for 10,000 steps, totaling approximately 21 billion tokens. We use the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, a weight decay of 0.1, and gradient clipping at 1.0. All runs leverage bfloat16 precision, Flash Attention 2 (Dao, 2024), and rotary position embeddings (Su et al., 2024).

Evaluation. Following Kydlíček et al. (2025), we assess performance across 12 benchmarks organized into 6 categories. We employ 3-shot cloze-format prompting and report macro-averaged scores: **General Knowledge:** ARC (Easy) (Clark et al., 2018), MMLU Redux (Gema et al., 2025); **Reading Comprehension:** SQuAD v2 (Rajpurkar et al., 2018), DROP (Dua et al., 2019); **Reasoning:** OpenBookQA (Mihaylov et al., 2018), XCSQA (Lin et al., 2021); **Natural Language Understanding:** WinoGrande (Sakaguchi et al., 2021), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019); **Math:** GSM8K (Cobbe et al., 2021); and **Table Understanding:** WikiTableQuestions (Pasupat & Liang, 2015), TriviaQA (Joshi et al., 2017). We include the full breakdown of results for every individual task in Appendix C.

Implementation Details. We built our pipeline using DataTrove (Penedo et al., 2024b, v0.6.0), with distributed inference for synthetic data generation managed through vLLM (Kwon et al., 2023, v0.8.4), training handled by Nanotron (Tazi et al., 2025, v0.4)[1], and evaluation executed via LightEval (Habib et al., 2023, v0.9.1dev0).

4 Results

4.1 Rephrasing Strategy

Preliminary Investigation: Existing Datasets. We first evaluate existing pretraining datasets to establish a performance baseline. This evaluation considers four popular curated web corpora: DCLM, Ultra-FineWeb (Wang et al., 2025), and the high/low-quality FineWeb subsets; and four synthetic datasets: Nemotron-HQ-Synth, REWIRE, Cosmopedia, and SYNTH (Langlais, 2025). As indicated in Table 2, DCLM achieves the highest macro-average performance of 13.77 and serves as the primary baseline for subsequent comparisons.

Table 2: Macro-averaged scores ($\times 100$) for different existing datasets at 10K steps. Best and second best in **bold** and underline. $\star$ indicates a synthetic dataset. Full results are in Table 10.

	Dataset	Macro Avg
Non-synthetic	DCLM	**13.77**
	Ultra-FineWeb	13.00
	FineWeb-HQ	11.82
	FineWeb-LQ	8.83
Synthetic	Nemotron-HQ-Synth$\star$	13.54
	REWIRE$\star$	13.49
	Cosmopedia$\star$	10.33
	SYNTH$\star$	10.03

The synthetic datasets Nemotron-HQ-Synth and REWIRE yield results comparable to the performance of DCLM, with a negligible delta of 0.28 points. However, because Nemotron-HQ-Synth consists of five distinct strategies (§2.2), the aggregate score may mask the high performance of specific transformations while incorporating the deficiencies of others. This lack of granularity necessitates the systematic ablation of these components, which we present in the following analysis (Table 3).

Established Prompts. We now examine the existing rephrasing prompts described in §2.2 to identify the most effective transformation. Table 3 illustrates that among established strategies, only *Diverse QA Pairs*, *Continue*, and *Guided Rewrite* rival the DCLM baseline.

Table 3: Macro-averaged scores for different existing rephrasing approaches. Full results are in Table 11.

	Prompt	Macro Avg
	DCLM	13.77
Nemotron-CC	Diverse QA Pairs	**14.58**
	Distill	13.17
	Wikipedia	13.14
	Knowledge List	13.11
	Extract Knowledge	11.81
REWIRE	Guided Rewrite	13.72
Beyond Web	Summarize	13.01
	Continue	13.73

The success of *Diverse QA Pairs* is noteworthy. It achieves a score of 14.58, representing a substantial improvement over 13.77 for DCLM. This result suggests that the conversion of raw text into a structured query-response format creates a more effective training signal than simple restatement (e.g., *Distill* and *Wikipedia*). This indeed supports the hypothesis that structured pedagogical prompts enhance the utility of the resulting tokens (§2.2). In contrast, strategies such as *Extract Knowledge* or *Knowledge List* show a substantial performance deficit, with drops ranging from 0.6 to 1.96. This divergence indicates that most rephrasing methodologies provide negligible utility beyond curated web text.

Pedagogical Structured Prompts. We evaluate four pedagogically structured prompts: `math`, `faq`, `table`, and `tutorial`. As shown in Table 4, all four formats consistently surpass the DCLM baseline: `math` (+1.54), `table` (+1.06), `faq` (+0.68), and `tutorial` (+0.53). Notably, the `math` and `table` formats also exceed the performance of the most effective existing prompt, *Diverse QA Pairs*, by margins of 0.73 and 0.25 points, respectively. These results confirm that the restructuring of source data into pedagogically rich formats provides greater utility than simple paraphrasing. We examine the specific task domains where these formats improve performance in §5.1.

Table 4: New pedagogical structured prompts vs. DCLM. Full results are in Table 12.

Prompt	Macro Avg
DCLM	13.77
`math`	15.31
`faq`	14.45
`table`	14.83
`tutorial`	14.30

[1] https://github.com/huggingface/nanotron/

4.2 Generator Model

Parameter Scale. Figure 2 compares the impact of model size on the quality of rephrased data using the Gemma 3 family across five scales with the best-performing `math` prompt. We observe a clear performance saturation point at the 1B parameter scale. The 1B model (15.31) outperforms the 27B variant (14.76), suggesting diminishing returns from increased model capacity. This trend persists within the SmolLM2 family using the `tutorial` prompt, which yields macro-average scores of 12.69 for 135M, 14.45 for 360M, and 15.88 for 1.7B.

An exception to this trend arises with the complex *Guided Rewrite* prompt from REWIRE (Appendix B.2). In this case, the 4B model demonstrates a moderate gain (14.58) over the 270M and 1B models (13.29 and 13.72).[2] This outcome likely results from the multi-step formatting requirements of the prompt, which may elevate the minimum capability threshold for the generator. These results indicate that *large models are not always necessary for effective synthesis*; this finding contrasts with the claim of REWIRE that larger models are inherently required for successful rephrasing (Nguyen et al., 2025). Furthermore, comparing Gemma 3 1B against 12B across varying levels of source data quality yields heterogeneous results. For the majority of prompts, the 12B model provides no consistent advantage, irrespective of the quality of the source data (Table 19 in Appendix C).

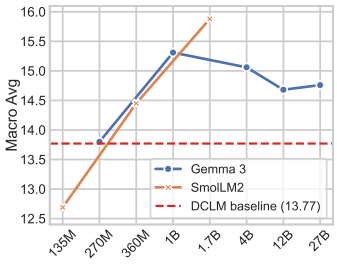


Figure 2: Macro-averaged scores for different Gemma 3 and SmolLM2 model scales. Full results are in Table 13.

Architectural Families. Table 5 compares the results of six architectural families described in §2.3 averaged across the four pedagogical prompts (§2.2). The SmolLM2 1.7B model surpasses all other architectural families by margins ranging from 1.01 to 2.06. This performance advantage is primarily attributed to superior reading comprehension metrics. Specifically, the performance of SmolLM2 exceeds the average of the other families by at least 5.12 points on SQuAD v2 (Table 14). This superior performance likely stems from the specific composition of the instruction-tuning data of SmolLM2, which includes explicit rewrite tasks (i.e., Smol-Rewrite) (Allal et al., 2025). Consequently, the model possesses a preexisting proficiency for structural transformation prior to the application of the experimental prompts.

Table 5: Comparison of model families across the four new prompts (all ∼1B scale). Full results are in Table 14.

Family	Macro Avg
DCLM	13.77
SmolLM2	16.55
Falcon 3	15.54
Qwen 3	14.49
Gemma 3	14.72
Granite 3.1	14.87
Llama 3.2	14.79

4.3 Mix-in and Source Data

Necessity of Mixing. Table 6 demonstrates that training exclusively on synthetic data yields suboptimal performance. Across all evaluated prompts, the integration of synthetic samples with original web text consistently surpasses configurations restricted to synthetic data. Consistent with the findings of Kang et al. (2025), these results suggest that while pure synthetic data may not surpass the efficacy of natural text, a mixture can accelerate convergence. This improvement stems primarily from the restoration of natural language understanding (NLU) capabilities (e.g., WinoGrande, PIQA, and HellaSwag), which pure synthetic data fails to provide in isolation (Table 15). Without the inclusion of original web tokens, the model remains susceptible to the risk of model collapse (Shumailov et al., 2024), as the generator fails to replicate the rich linguistic diversity of the crawlable web. We further discuss the performance trade-offs between synthetic pedagogical data and original web text in §5.1.

Table 6: Comparison of synthetic-only and mixed training. Full results are in Table 15.

Prompt	Synthetic Only	Mixed
DCLM	13.77	
`math`	15.20	**15.31**
`faq`	13.12	**14.45**
`table`	13.74	**14.83**
`tutorial`	12.32	**14.30**

[2]The full results are in Table 18 in Appendix C.

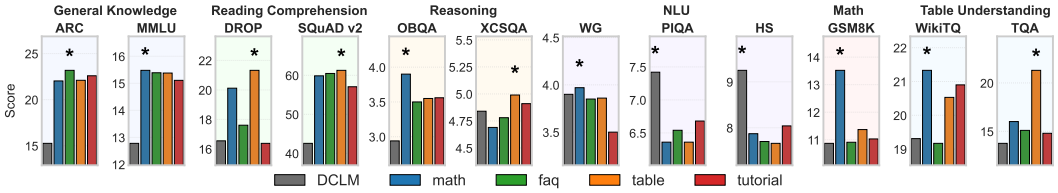


Figure 3: Fine-grained evaluation of pedagogical formats against the DCLM baseline. The best result in each task is marked with ⋆. We use the best-performing SmolLM2 1.7B as $\mathcal{G}$.

Selection of Mix-in Data. We evaluate the impact of the selection of $\mathcal{D}_{\text{mix}}$ by fixing $\mathcal{D}_{\text{source}}$ to FWHQ or FWLQ and employing the `tutorial` prompt. As shown in Table 7, all configurations incorporating a mix-in dataset surpass the source-only baseline (indicated in gray). In high-quality source settings, DCLM and FWHQ serve as the most effective mix-in components, outperforming the baseline by 2.50 and 2.48 points, respectively. Conversely, when rephrasing low-quality data, the inclusion of FWHQ as the mix-in dataset yields slightly superior utility (12.99) compared to DCLM (12.62). These results establish that the selection of the mix-in dataset is a critical determinant of final performance, with FWHQ and DCLM representing the optimal choices for the mix-in component.

Source Data Quality and Up-cycling. We further examine the influence of the quality of $\mathcal{D}_{\text{source}}$ using the `tutorial` prompt. When we set $\mathcal{D}_{\text{source}} = \mathcal{D}_{\text{mix}}$, the quality of the source material serves as a primary factor of the final performance of the model (Table 8). Specifically, large variance exists between configurations, ranging from 9.63 for FWLQ to 14.3 for FWHQ. However, a distinct trend emerges when the quality of $\mathcal{D}_{\text{mix}}$ is fixed at a high level: the inherent quality of $\mathcal{D}_{\text{source}}$ becomes less dominant. In this high-quality mix-in setting, the performance delta narrows to a maximum of 1.78 points. These results demonstrate that the rephrasing of low-quality data yields competitive results if the training process pairs the synthetic samples with a robust mix-in corpus. This observation is encouraging: it suggests that the generation pipeline can successfully up-cycle noisy web text into high-utility training tokens, thereby effectively expanding the reservoir of available source data for pretraining.

Table 7: Comparison of different mix-in dataset $\mathcal{D}_{\text{mix}}$ choices with FWHQ (HQ) or FWLQ (LQ) as $\mathcal{D}_{\text{source}}$. Full results in Table 16.

	$\mathcal{D}_{\text{source}}$	
Mix-in ($\mathcal{D}_{\text{mix}}$)	**HQ**	**LQ**
FineWeb	11.82	8.83
Cosmopedia	12.99	10.83
DCLM	14.32	12.62
FWHQ	14.30	12.99
FWLQ	13.61	9.63

Table 8: Comparison of different source dataset $\mathcal{D}_{\text{source}}$ choices. Full results in Table 17.

$\mathcal{D}_{\text{source}}$	$\mathcal{D}_{\text{source}} = \mathcal{D}_{\text{mix}}$	$\mathcal{D}_{\text{mix}} = \textbf{HQ}$
Cosmopedia	10.36	13.88
DCLM	13.69	14.77
FWHQ	14.30	14.30
FWLQ	9.63	12.99

5 Analysis

Having completed our systematic study to determine broad trends, we now dig into finer-grained details of our results to better understand specific factors and broader trends.

5.1 Performance Trade-off between Synthetic and Web Data

A fine-grained evaluation reveals a performance trade-off between synthetic pedagogical data and original web text. Figure 3 shows a tension between specialized knowledge and commonsense reasoning, particularly within Natural Language Understanding (NLU). Synthetic formats consistently fall behind the DCLM baseline on HellaSwag (HS) and PIQA. For instance, the `table` format achieves only 7.66 on HS compared to 9.25 for DCLM.

Nevertheless, these formats achieve higher aggregate scores (§4.1) because improvements in General Knowledge and Reading Comprehension outweigh these specific deficits. Specifically, the `table` format improves SQuAD v2 scores by 18.72 over the baseline, while the

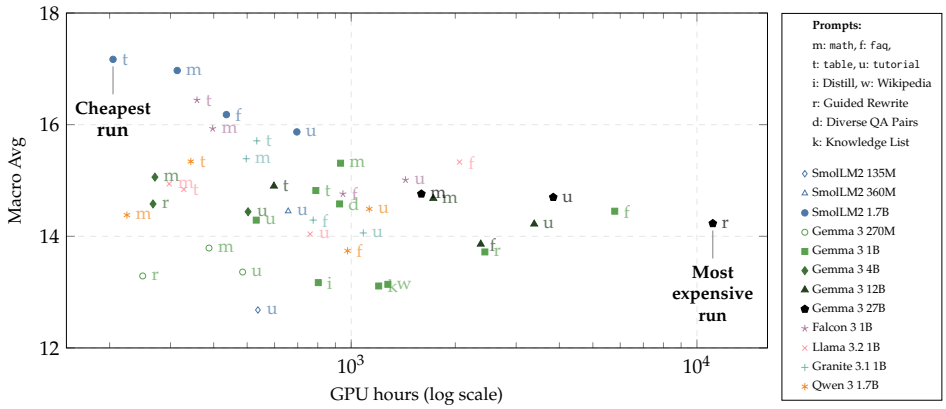


Figure 4: GPU cost vs. performance. Symbols distinguish model variants; adjacent characters represent prompt formats (e.g., **t**: table, **m**: math).

`faq` format increases ARC scores by 7.92. Structured synthesis thus yields high-density signals for specialized knowledge and logic, though it inherently reduces the linguistic diversity found in natural text. As discussed in §4.3, original web data acts as a synergistic component that reinstates the performance of the model in NLU; rephrasing cannot replicate this capability in isolation.

5.2 Qualitative Analysis: Template Collapse and Output Diversity

We evaluate 1,000 outputs from SmolLM2 1.7B and Qwen 3 1.7B using the `math` prompt to identify the qualitative factors that drive the performance gap between these architectural families. The analysis focuses on three primary criteria: solution completeness, variance in output length, and structural repetition.

The results reveal a performance paradox: while Qwen 3 achieves 100% instruction adherence, with 100% of samples containing distinct problem and solution sections, only 68% of the outputs from SmolLM2 feature complete solutions (see examples in Appendix D). Despite this inconsistency, the data of SmolLM2 yields superior downstream performance as observed in §4.2. We identify the cause as *template collapse*: a phenomenon where the extreme consistency of a generator produces repetitive and uniform training data that saturates the learning signal. For example, 115 outputs from Qwen 3 begin with identical introductory text. Conversely, the most frequent starting pattern in the data of SmolLM2 appears only three times. Furthermore, SmolLM2 shows a wider length variance (4 to 4,000 tokens) compared to the rigid formatting of Qwen 3 (100 to 2,600 tokens). These findings establish that for the objective of pretraining, the diversity of the output exerts a more substantial influence than the consistency of the format. A generator that is too well-behaved may inadvertently filter out the structural variety necessary for a robust signal, thereby exacerbating the narrow linguistic scope inherent to synthetic data (§5.1).

5.3 Cost-Benefit Analysis of Model Scale

The most efficient iteration (utilizing `table` with SmolLM2 1.7B) necessitates 8 GPU-days, whereas the most resource-intensive (utilizing *Guided Rewrite* with Gemma 3 27B) consumes over 15 GPU-months (Figure 4). The Pareto frontier is dominated by small models employing the four structured prompts described in §4.1. Specifically, models in the 1B-parameter class, such as Gemma 3 1B and SmolLM2 1.7B, often achieve the optimal cost-to-performance ratio when paired with these structured formats. Scaling the generator to 12B or 27B parameters increases GPU time by 5 to 10 times while simultaneously decreasing performance in most cases. These findings suggest that researchers should prioritize the design of the rephrasing prompt over the scale of the generator model.

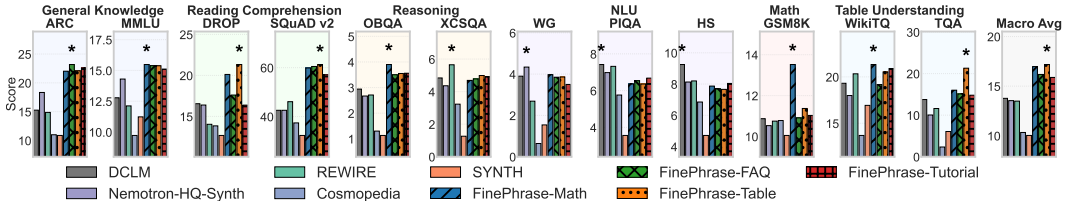


Figure 5: FINEPHRASE prompts vs. baselines. The best result in each task is marked with ⋆. All FINEPHRASE runs use SmolLM2 1.7B on FineWeb, mixed with FineWeb-HQ.

Investing in prompt architecture provides a higher return on compute than simply increasing the parameter count of the generator.

6 FinePhrase

We apply our empirical findings to build FINEPHRASE, a large-scale synthetic dataset. The construction of the dataset follows the optimal configuration identified in our analysis: we utilize SmolLM2 1.7B as the generator model (§4.2) and employ the four pedagogical structured formats: `math`, `faq`, `table`, and `tutorial` (§4.1), given the cost-benefit analysis in §5.3.

Scale and Efficiency. FINEPHRASE rephrases documents from FineWeb into four structured formats, yielding 1.35 billion samples and 486 billion completion tokens.[3] By using 100 NVIDIA H100 GPUs and SmolLM2 1.7B with suffix-32 speculative decoding, the generation pipeline achieves a throughput of approximately 9,200 tokens per second per GPU. Consequently, the full generation necessitates only 612 GPU-days (approximately 14,700 GPU-hours). As shown in Table 9, this throughput represents a 30× efficiency gain over the process used for REWIRE (which employs Llama 3.3 70B) and a 13× increase over the methodology of Cosmopedia (which utilizes Mixtral 8x7B). FINEPHRASE generates a higher volume of tokens than REWIRE while consuming 24× less total compute. This demonstrates that structured rephrasing with compact models provides a step-change in generation throughput.

Table 9: Generation cost comparison (all H100 GPU-hours).

Dataset	Generator	Tokens	GPU-hrs	Tok/GPU-hr
Cosmopedia	Mixtral 8x7B	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

Results. Figure 5 demonstrates that every FINEPHRASE variant substantially surpasses the established benchmarks in terms of macro-average scores. FINEPHRASE-Table achieves a peak macro-average of 17.18, representing an improvement of 3.41 points over DCLM and 3.63 points over Nemotron-HQ-Synth. The breakdown results validate the trade-off discussed in §5.1: logically dense FINEPHRASE formats dominate factual knowledge and reading comprehension (ARC, DROP, SQuAD v2), whereas the DCLM retain a nominal advantage in NLU (PIQA, HellaSwag). This performance profile validates the design of FINEPHRASE as a synergistic component intended for integration with original web data.

7 Conclusions

We evaluated synthetic pretraining across rephrasing strategies, generator models, and source data. Accordingly, we found that: the design of rephrasing prompts primarily determines performance; structured pedagogical formats exceed the utility of simple paraphrasing; generator capacities beyond 1B parameters yield negligible gains for most rephrasing prompts; source quality has minimal impact when paired with robust mix-in data; and generative diversity outweighs formatting consistency. While synthetic tokens provide logical depth, original web data remains essential to preserve commonsense reasoning and

[3]We include some actual samples in Appendix E.

linguistic variety. These principles culminate in FINEPHRASE, a 486-billion-token dataset that achieves up to a 30× cost reduction while outperforming established baselines.

Limitations. As the scale of the model was not a primary covariate, experiments use a 1.2B model trained on 21B tokens. While this setup follows FineWeb (Penedo et al., 2024a), results may vary at larger scales. Evaluation is restricted to standard English benchmarks and our results may therefore be benchmark- and language-specific. Finally, cost comparisons with prior work rely on reported figures that can fluctuate across hardware environments.

Reproducibility Statement

All training hyperparameters, prompt templates, and evaluation details are provided in §3 and Appendix B. The FINEPHRASE dataset (486B tokens), the complete ablation data, the underlying generation framework, the code for generating the FINEPHRASE dataset, and the full suite of benchmark results are available at https://hf.co/datasets/HuggingFaceFW/finephrase, https://hf.co/datasets/HuggingFaceFW/finephrase-rephrased, https://github.com/huggingface/finephrase, https://github.com/huggingface/datatrove/blob/main/examples/inference/finephrase.py, and https://huggingface.co/spaces/HuggingFaceFW/finephrase/blob/main/app/src/content/assets/data/benchmark-results.csv, respectively.

Ethics Statement

Synthetic data generation raises concerns regarding the practice of training on model outputs. We address the model collapse concern directly through a pipeline that integrates synthetic data with original web data, which ensures that training does not occur exclusively on the outputs of a model. We do not filter for or against any specific topics, and the content distribution of the source data remains consistent with the distribution of FineWeb. Furthermore, the synthetic data is always obtained by rephrasing seed data, and not by simply sampling many times from a model with a fixed prompt. The dataset and the code are released openly to facilitate scrutiny and replication by the research community.

Acknowledgment

The authors acknowledge the use of generative AI tools in the preparation of this manuscript. Specifically, Gemini 3 Thinking was employed to generate Figure 4 from raw samples, followed by manual verification by the authors.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report. *arXiv preprint*, arXiv:2412.08905, 2024. URL https://arxiv.org/abs/2412.08905.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebron, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4895–4901, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.298. URL https://aclanthology.org/2023.emnlp-main.298/.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martin Blazquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Agustín Piqueres Lajarín, Hynek Kydlíček, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan Son NGUYEN, Ben Burtenshaw,

Clémentine Fourrier, Haojun Zhao, Hugo Larcher, Mathieu Morlon, Cyril Zakka, Colin Raffel, Leandro Von Werra, and Thomas Wolf. SmolLM2: When smol goes big — data-centric training of a fully open small language model. In *Proceedings of the Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=3JiCl2A14H.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. Cosmopedia, 2024. URL https://huggingface.co/datasets/HuggingFaceTB/cosmopedia. Data repository.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, Apr. 2020. doi: 10.1609/aaai.v34i05.6239. URL https://ojs.aaai.org/index.php/AAAI/article/view/6239.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. Instruction pre-training: Language models are supervised multitask learners. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2529–2550, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.148. URL https://aclanthology.org/2024.emnlp-main.148/.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint*, arXiv:1803.05457, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint*, arXiv:2110.14168, 2021. URL https://arxiv.org/abs/2110.14168.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=mZn2Xyh9Ec.

DatologyAI, Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Urbanek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Charvi Bannur, Christina Baek, Darren Teh, David Schwab, Haakon Mongstad, Haoli Yin, Josh Wills, Kaleigh Mentzer, Luke Merrick, Ricardo Monti, Rishabh Adiga, Siddharth Joshi, Spandan Das, Zhengping Wang, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. BeyondWeb: Lessons from scaling synthetic data for trillion-scale pretraining. *arXiv preprint*, arXiv:2508.10975, 2025. URL https://arxiv.org/abs/2508.10975.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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/.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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. 2368–2378, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1246. URL https://aclanthology.org/N19-1246/.

Falcon-LLM Team. The Falcon 3 family of open models, 2024. URL https://huggingface.co/blog/falcon3. Blog post.

Lukas Galke, Yoav Ram, and Limor Raviv. Deep neural networks and humans both benefit from compositional language structure. *Nature Communications*, 15(10816), 2024. ISSN 2041-1723.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint*, arXiv:2101.00027, 2020. URL https://arxiv.org/abs/2101.00027.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. Are we done with MMLU? In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5069–5096, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.262. URL https://aclanthology.org/2025.naacl-long.262/.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report. *arXiv preprint*, arXiv:2503.19786, 2025. URL https://arxiv.org/abs/2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide,

Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models. *arXiv preprint*, arXiv:2407.21783, 2024. URL https://arxiv.org/abs/2407.21783.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need. *arXiv preprint*, arXiv:2306.11644, 2023. URL https://arxiv.org/abs/2306.11644.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. LightEval: A lightweight framework for LLM evaluation, 2023. URL https://github.com/huggingface/lighteval. GitHub repository.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran

Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

IBM Granite Team. Granite 3.0 language models, 2024. URL https://github.com/ibm-granite/granite-3.0-language-models. Technical Report.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Feiyang Kang, Newsha Ardalani, Michael Kuchnik, Youssef Emad, Mostafa Elhoushi, Shubhabrata Sengupta, Shang-Wen Li, Ramya Raghavendra, Ruoxi Jia, and Carole-Jean Wu. Demystifying synthetic data in LLM pre-training: A systematic study of scaling laws, benefits, and pitfalls. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10739–10758, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.544. URL https://aclanthology.org/2025.emnlp-main.544/.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint*, arXiv:2001.08361, 2020. URL https://arxiv.org/abs/2001.08361.

Hirokazu Kiyomaru, Yusuke Oda, Takashi Kodama, Chaoran Liu, and Daisuke Kawahara. Scaling data-constrained language models with synthetic data. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1002–1016, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-386-9. URL https://aclanthology.org/2026.findings-eacl.52/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Hynek Kydlíček, Guilherme Penedo, and Leandro von Werra. FinePDFs: Liberating 3T of the finest tokens from PDFs, 2025. URL https://huggingface.co/spaces/HuggingFaceFW/FinePDFsBlog. Blog post.

Pierre-Carl Langlais. SYNTH: the new data frontier, 2025. URL https://huggingface.co/blog/Pclanglais/synth-data-frontier. Blog post.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. DataComp-LM: In search of the next generation of training sets for language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 14200–14282. Curran Associates, Inc., 2024. doi: 10.52202/

079017-0455. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/19e4ea30dded58259665db375885e412-Paper-Datasets_and_Benchmarks_Track.pdf.

Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. Common sense beyond English: Evaluating and improving multilingual language models for commonsense reasoning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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)*, pp. 1274–1287, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.102. URL https://aclanthology.org/2021.acl-long.102/.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the Seventh International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14044–14072, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.757. URL https://aclanthology.org/2024.acl-long.757/.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 50358–50376. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf.

Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. Recycling the web: A method to enhance pre-training data quality and quantity for language models. In *Proceedings of the Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=lkjhBdz3rn.

NVIDIA, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, Abhinav Khattar, Adi Renduchintala, Aditya Vavre, Akanksha Shukla, Akhiad Bercovich, Aleksander Ficek, Aleksandr Shaposhnikov, Alex Kondratenko, Alexander Bukharin, Alexandre Milesi, Ali Taghibakhshi, Alisa Liu, Amelia Barton, Ameya Sunil Mahabaleshwarkar, Amir Klein, Amit Zuker, Amnon Geifman, Amy Shen, Anahita Bhiwandiwalla, Andrew Tao, Anjulie Agrusa, Ankur Verma, Ann Guan, Anubhav Mandarwal, Arham Mehta, Ashwath Aithal, Ashwin Poojary, Asif Ahamed, Asit Mishra, Asma Kuriparambil Thekkumpate, Ayush Dattagupta, Banghua Zhu, Bardiya Sadeghi, Barnaby Simkin, Ben Lanir, Benedikt Schifferer, Besmira Nushi, Bilal Kartal, Bita Darvish Rouhani, Boris Ginsburg, Brandon Norick, Brandon Soubasis, Branislav Kisacanin, Brian Yu, Bryan Catanzaro, Carlo del Mundo, Chantal Hwang, Charles Wang, Cheng-Ping Hsieh, Chenghao Zhang, Chenhan Yu, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Christopher Parisien, Collin Neale, Cyril Meurillon, Damon Mosk-Aoyama, Dan Su, Dane Corneil, Daniel Afrimi, Daniel Lo, Daniel Rohrer, Daniel Serebrenik, Daria Gitman, Daria Levy, Darko Stosic, David Mosallanezhad, Deepak Narayanan, Dhruv Nathawani, Dima Rekesh, Dina Yared, Divyanshu Kakwani, Dong Ahn, Duncan Riach, Dusan Stosic, Edgar Minasyan, Edward Lin, Eileen Long, Eileen Peters Long, Elad Segal, Elena Lantz, Ellie Evans, Elliott Ning, Eric Chung, Eric Harper, Eric Tramel, Erick Galinkin, Erik Pounds, Evan Briones, Evelina Bakhturina, Evgeny Tsykunov, Faisal Ladhak, Fay Wang, Fei Jia, Felipe Soares,

Feng Chen, Ferenc Galko, Frank Sun, Frankie Siino, Gal Hubara Agam, Ganesh Ajjanagadde, Gantavya Bhatt, Gargi Prasad, George Armstrong, Gerald Shen, Gorkem Batmaz, Grigor Nalbandyan, Haifeng Qian, Harsh Sharma, Hayley Ross, Helen Ngo, Herbert Hum, Herman Sahota, Hexin Wang, Himanshu Soni, Hiren Upadhyay, Huizi Mao, Huy C Nguyen, Huy Q Nguyen, Iain Cunningham, Ido Galil, Ido Shahaf, Igor Gitman, Ilya Loshchilov, Itamar Schen, Itay Levy, Ivan Moshkov, Izik Golan, Izzy Putterman, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jatin Mitra, Jeffrey Glick, Jenny Chen, Jesse Oliver, Jian Zhang, Jiaqi Zeng, Jie Lou, Jimmy Zhang, Jinhang Choi, Jining Huang, Joey Conway, Joey Guman, John Kamalu, Johnny Greco, Jonathan Cohen, Joseph Jennings, Joyjit Daw, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kai Xu, Kan Zhu, Kari Briski, Katherine Cheung, Katherine Luna, Keith Wyss, Keshav Santhanam, Kevin Shih, Kezhi Kong, Khushi Bhardwaj, Kirthi Shankar, Krishna C. Puvvada, Krzysztof Pawelec, Kumar Anik, Lawrence McAfee, Laya Sleiman, Leon Derczynski, Li Ding, Lizzie Wei, Lucas Liebenwein, Luis Vega, Maanu Grover, Maarten Van Segbroeck, Maer Rodrigues de Melo, Mahdi Nazemi, Makesh Narsimhan Sreedhar, Manoj Kilaru, Maor Ashkenazi, Marc Romeijn, Marcin Chochowski, Mark Cai, Markus Kliegl, Maryam Moosaei, Matt Kulka, Matvei Novikov, Mehrzad Samadi, Melissa Corpuz, Mengru Wang, Meredith Price, Michael Andersch, Michael Boone, Michael Evans, Miguel Martinez, Mikail Khona, Mike Chrzanowski, Minseok Lee, Mohammad Dabbah, Mohammad Shoeybi, Mostofa Patwary, Nabin Mulepati, Najeeb Nabwani, Natalie Hereth, Nave Assaf, Negar Habibi, Neta Zmora, Netanel Haber, Nicola Sessions, Nidhi Bhatia, Nikhil Jukar, Nikki Pope, Nikolai Ludwig, Nima Tajbakhsh, Nir Ailon, Nirmal Juluru, Nishant Sharma, Oleksii Hrinchuk, Oleksii Kuchaiev, Olivier Delalleau, Oluwatobi Olabiyi, Omer Ullman Argov, Omri Puny, Oren Tropp, Ouye Xie, Parth Chadha, Pasha Shamis, Paul Gibbons, Pavlo Molchanov, Pawel Morkisz, Peter Dykas, Peter Jin, Pinky Xu, Piotr Januszewski, Pranav Prashant Thombre, Prasoon Varshney, Pritam Gundecha, Przemek Tredak, Qing Miao, Qiyu Wan, Rabeeh Karimi Mahabadi, Rachit Garg, Ran El-Yaniv, Ran Zilberstein, Rasoul Shafipour, Rich Harang, Rick Izzo, Rima Shahbazyan, Rishabh Garg, Ritika Borkar, Ritu Gala, Riyad Islam, Robert Hesse, Roger Waleffe, Rohit Watve, Roi Koren, Ruoxi Zhang, Russell Hewett, Russell J. Hewett, Ryan Prenger, Ryan Timbrook, Sadegh Mahdavi, Sahil Modi, Samuel Kriman, Sangkug Lim, Sanjay Kariyappa, Sanjeev Satheesh, Saori Kaji, Satish Pasumarthi, Saurav Muralidharan, Sean Narentharen, Sean Narenthiran, Seonmyeong Bak, Sergey Kashirsky, Seth Poulos, Shahar Mor, Shanmugam Ramasamy, Shantanu Acharya, Shaona Ghosh, Sharath Turuvekere Sreenivas, Shelby Thomas, Shiqing Fan, Shreya Gopal, Shrimai Prabhumoye, Shubham Pachori, Shubham Toshniwal, Shuoyang Ding, Siddharth Singh, Simeng Sun, Smita Ithape, Somshubra Majumdar, Soumye Singhal, Stas Sergienko, Stefania Alborghetti, Stephen Ge, Sugam Dipak Devare, Sumeet Kumar Barua, Suseella Panguluri, Suyog Gupta, Sweta Priyadarshi, Syeda Nahida Akter, Tan Bui, Teodor-Dumitru Ene, Terry Kong, Thanh Do, Tijmen Blankevoort, Tim Moon, Tom Balough, Tomer Asida, Tomer Bar Natan, Tomer Ronen, Tugrul Konuk, Twinkle Vashishth, Udi Karpas, Ushnish De, Vahid Noorozi, Vahid Noroozi, Venkat Srinivasan, Venmugil Elango, Victor Cui, Vijay Korthikanti, Vinay Rao, Vitaly Kurin, Vitaly Lavrukhin, Vladimir Anisimov, Wanli Jiang, Wasi Uddin Ahmad, Wei Du, Wei Ping, Wenfei Zhou, Will Jennings, William Zhang, Wojciech Prazuch, Xiaowei Ren, Yashaswi Karnati, Yejin Choi, Yev Meyer, Yi-Fu Wu, Yian Zhang, Yigong Qin, Ying Lin, Yonatan Geifman, Yonggan Fu, Yoshi Subara, Yoshi Suhara, Yubo Gao, Zach Moshe, Zhen Dong, Zhongbo Zhu, Zihan Liu, Zijia Chen, and Zijie Yan. NVIDIA Nemotron 3: Efficient and open intelligence. *arXiv preprint*, arXiv:2512.20856, 2025. URL https://arxiv.org/abs/2512.20856.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In Chengqing Zong and Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1142. URL https://aclanthology.org/P15-1142/.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.),

Advances in Neural Information Processing Systems, volume 36, pp. 79155–79172. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fa3ed726cc5073b9c31e3e49a807789c-Paper-Datasets_and_Benchmarks.pdf.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/370df50ccfdf8bde18f8f9c2d9151bda-Paper-Datasets_and_Benchmarks_Track.pdf.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. DataTrove: Large scale data processing, 2024b. URL https://github.com/huggingface/datatrove. GitHub repository.

Zeyu Qin, Qingxiu Dong, Xingxing Zhang, Li Dong, Xiaolong Huang, Ziyi Yang, MAHMOUD KHADEMI, Dongdong Zhang, Hany Hassan Awadalla, Yi R. Fung, Weizhu Chen, Minhao Cheng, and Furu Wei. Scaling laws of synthetic data for language model. In *Proceedings of the Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=UmUXPXHtdl.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Technical report.

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. URL http://jmlr.org/papers/v21/20-074.html.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: an adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106, August 2021. ISSN 0001-0782. doi: 10.1145/3474381. URL https://doi.org/10.1145/3474381.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi: 10.1038/s41586-024-07566-y. URL https://doi.org/10.1038/s41586-024-07566-y.

Varun Singh, Lucas Krauss, Sami Jaghouar, Matej Sirovatka, Charles Goddard, Fares Obied, Jack Min Ong, Jannik Straube, Fern, Aria Harley, Conner Stewart, Colin Kealty, Maziyar Panahi, Simon Kirsten, Anushka Deshpande, Anneketh Vij, Arthur Bresnu, Pranav Veldurthi, Raghav Ravishankar, Hardik Bishnoi, DatologyAI Team, Arcee AI Team, Prime Intellect Team, Mark McQuade, Johannes Hagemann, and Lucas Atkins. Arcee trinity large technical report. *arXiv preprint*, arXiv:2602.17004, 2026. URL https://arxiv.org/abs/2602.17004.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2459–2475, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.123. URL https://aclanthology.org/2025.acl-long.123/.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063. URL https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Nouamane Tazi, Ferdinand Mom, Haojun Zhao, Phuc Nguyen, Mohamed Mekkouri, Leandro Werra, and Thomas Wolf. The ultra-scale playbook: Training LLMs on GPU clusters, 2025. URL https://huggingface.co/spaces/nanotron/ultrascale-playbook. Blog post.

Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, Xu Han, and Zhiyuan Liu. Ultra-FineWeb: Efficient data filtering and verification for high-quality LLM training data. *arXiv preprint*, arXiv:2505.05427, 2025. URL https://arxiv.org/abs/2505.05427.

Atsuki Yamaguchi, Maggie Mi, and Nikolaos Aletras. Enhancing linguistic competence of language models through pre-training with language learning tasks. *arXiv preprint*, arXiv:2601.03448, 2026. URL https://arxiv.org/abs/2601.03448.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *arXiv preprint*, arXiv:2407.10671, 2024. URL https://arxiv.org/abs/2407.10671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint*, arXiv:2505.09388, 2025a. URL https://arxiv.org/abs/2505.09388.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. Synthetic continued pretraining. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=07yvxWDSla.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 19–27, 2015. doi: 10.1109/ICCV.2015.11.

A Related work

Synthetic Data for Pretraining. Foundational research such as WRAP (Maini et al., 2024) demonstrated that rephrasing web text into diverse styles improves pretraining at a 1.3B

scale. Nemotron-CC (Su et al., 2025) scaled this approach to 6.3T tokens using multiple prompts; we isolate the individual prompts of that study and reveals that only Diverse QA Pairs matches the performance of the DCLM baseline. REWIRE (Nguyen et al., 2025) employs Llama 3.3 70B for guided rewriting and suggests that reasoning capabilities of large models are necessary to process data of low quality. The comparison in §4.2 reveals no consistent advantage for larger models. Indeed, FINEPHRASE achieves superior results with a 1.7B model at a cost up to 30× lower than the guided rewrite strategy of REWIRE. BeyondWeb (DatologyAI et al., 2025) reports strong performance with continuation and summarization rephrasing but did not provide a public data release. We test those prompts and show that neither matches the performance of DCLM. Cosmopedia (Ben Allal et al., 2024) generates content from scratch rather than rephrasing. This approach underperforms every rephrasing approach in our analysis. EntiGraph (Yang et al., 2025b) targets continued pretraining with entity-centric augmentation; their diversity scaling complements our finding that prompt-level diversity saturates at approximately 20B tokens.

Understanding Synthetic Data. Kang et al. (2025) trained over 1,000 LLMs and determined that a mixture of ∼30% synthetic data substantially accelerates convergence, while generator models exceeding 8B parameters do not yield superior pretraining data. This study extends this analysis to prompt design, the choice of mix-in datasets, and the phenomenon of template collapse. While Qin et al. (2025) investigate the scaling laws of synthetic data, we focus on design choices at fixed scale. The literature on model collapse (Shumailov et al., 2024) documents the degradation resulting from training exclusively on model outputs. Our pipeline circumvents this issue by consistently mixing synthetic content with the original data.

B Prompt Templates

B.1 Nemotron-CC Prompts

Diverse QA Pairs.

```
Task: Read the text, ask questions and answer them.
Follow these instructions:
1. Ask diverse questions that require different cognitive skills or cover
↪  different aspects of the text.
1. Ask questions in various forms such as:
    - Yes/No questions that require determining whether a statement is true or
    ↪  false.
    - Open-ended questions that begin with words like what, how, when, where,
    ↪  why and who.
    - Multi-choice questions that offers two or more options to choose from.
    ↪  Include the options in the question.
    - Comparison questions that compare two quantities or objects and determine
    ↪  the relationship between them.
    - Reading comprehension questions that test the ability to understand and
    ↪  analyze the text.
    - Problem-solving questions that test the ability to solve mathematical,
    ↪  physical, or logical problems.

1. Focus on asking questions about factual information, important knowledge, or
↪  concrete details in the text.
1. Write questions and answers using clear and concise language.
1. Use plain text. Do not use Markdown.
1. Each question and answer pair should be on a separate line. Tag the question
↪  with "Question:" and the answer with "Answer:".

Text:
[TEXT]
Task:
```

```
After reading the above text, ask up to 8 questions and provide the correct
↪  answers following the instructions. Give your response in this format:
Here are the questions and answers based on the provided text:
- Question: [first question] Answer: [first answer]
- Question: [second question] Answer: [second answer]

....
```

Extract Knowledge.

```
Your task is to rewrite knowledge from the provided text following these
↪  instructions:
- Rewrite the text as a passage or passages using easy-to-understand and
↪  high-quality English like sentences in textbooks and Wikipedia.
- Focus on content in disciplines such as humanities, social sciences, natural
↪  sciences, technology, engineering, math, law and legal, business,
↪  management, art, education, agricultural sciences, politics, and history.
- Disregard content that does not contain useful facts or knowledge.
- Retain examples, explanations of reasoning processes, and supporting evidence
↪  to maintain the text's depth and context.
- Do not add or alter details. Only restate what is already in the text.
- Write in plain text.
- Do not add titles, subtitles, note, or comment.

Text:
[TEXT]
Task:
Rewrite facts and knowledge from the above text as a passage or passages
↪  following the instructions.
```

Distill.

```
Your task is to read and paraphrase the provided text following these
↪  instructions:
- Aim to create a condensed but accurate and informative version of the original
↪  text, not a simplistic summary.
- Capture and preserve the crucial information, key concepts, important values,
↪  and factual details in the original text, while making it more readable and
↪  accessible.
- Retain technical terms, specialized vocabulary, and complex concepts.
- Retain examples, explanations of reasoning processes, and supporting evidence
↪  to maintain the text's depth and context.
- Only include information that is present in the original text. Do not adding
↪  new or unsubstantiated claims.
- Write in plain text.

Here is the text:
[TEXT]
Task:
After thoroughly reading the above text, paraphrase it in high-quality and clear
↪  English following the instructions.
```

Wikipedia.

```
For the following paragraph give me a diverse paraphrase of the same in high
↪  quality English language as in sentences on Wikipedia. Begin your answer on
↪  a separate line with "Here is a paraphrased version:".
Text:
[TEXT]
```

Knowledge List.

```
Review the text and extract the key information. Follow these instructions:
- Carefully read the above text and provide a concise and organized list of
 ↪  factual information, concrete details, key concepts, and important numbers
 ↪  and statistics extracted from the text.
- Ensure each point is clear, specific, and supported by the original text.
- Ensure the extract text is information-dense and easier to learn from.
- Do not add titles or headings.

Text:
[TEXT]
Task:
Extract the factual information, concrete details, and key concepts from the
 ↪  above text following the instructions.
```

B.2 REWIRE

```
Below is a draft from an AI Assistant when trying to accomplish task or solving
 ↪  a problem. Analyze and understand the task and problem(s) to be solved. Then
 ↪  pretend to be the expert who is most skillful to acomplish this task, write
 ↪  down the detailed thinking process and internal monologue that went into
 ↪  identifying a strategy and lay out a plan about how to solve this problem.
 ↪  Experts usually apply meta-reasoning and planning to reason about how to
 ↪  best accomplish the task before jumping to solution.
Deliberate meta-reasoning also involves reflection which can help identify
 ↪  issues and take a step back to explore other paths. Below are some generic
 ↪  examples of starting questions experts could ask themselves during
 ↪  meta-reasoning process. The expert will come up with the most relevant
 ↪  questions that can help with their thinking process, which are also very
 ↪  specific to the task.
Let's first try to understand the task and exactly what problem(s) to be solved.
 ↪  What is the core issue or problem that needs to be addressed? What are the
 ↪  key assumptions underlying this problem?
How can I break down this problem into smaller, more manageable parts? How can I
 ↪  simplify the problem so that it is easier to solve?
What kinds of solution typically are produced for this kind of problem
 ↪  specification? Given the problem specification and the current best
 ↪  solution, have a guess about other possible solutions. Let's imagine the
 ↪  current best solution is totally wrong, what other ways are there to think
 ↪  about the problem specific
What is the best way to modify this current best solution, given what you know
 ↪  about these kinds of problem specification?
Am I on the right track? Let's check our progress so far.
Let's make a step by step plan and implement it with good notion and explanation.
Finally, write an improved response after thinking about how to accomplish the
 ↪  task. Take information and details from the original draft whenever they are
 ↪  useful. Therefore, the improved response should not be shorter than the
 ↪  original response. The improved response should have better formatting and
 ↪  readability, with more coherent and in-depth reasoning, while removing any
 ↪  noise or digression. Note that the best experts chosen to answer each prompt
 ↪  may be different, so please make sure the you do not sound like the same
 ↪  expert for all tasks.
IMPORTANT: Start your analysis and thinking right away. DO NOT add any filler
 ↪  text, explanations or notes about your response. Put the thinking and
 ↪  planning between {'<'}thinking starts{'>'} and {'<'}thinking ends{'>'}, and
 ↪  the improved response between {'<'}improved response starts{'>'} and
 ↪ {'<'}improved response ends{'>'}.
Original Draft: [TEXT]
```

B.3 BeyondWeb

Continue.

```
Continue the following text in the same style as the original. Start with the
↪  continuation directly.
Text:
[TEXT]
```

Summarize.

```
Summarize the following text. Write a standalone summary without referencing the
↪  text. Directly start with the summary. Do not say anything else.
Text:
[TEXT]
Summary:
```

B.4 Structured Prompts

math.

```
Rewrite the document to create a mathematical word problem based on the
↪  numerical data or relationships in the text. Provide a step-by-step solution
↪  that shows the calculation process clearly. Create a problem that requires
↪  multi-step reasoning and basic arithmetic operations. It should include the
↪  question followed by a detailed solution showing each calculation step.
↪  Output only the problem and solution, nothing else.
Document:
[TEXT]
```

faq.

```
Rewrite the document as a comprehensive FAQ (Frequently Asked Questions).
↪  Extract or infer the key questions a reader would have about this topic,
↪  then provide clear, direct answers. Order questions logically—from
↪  foundational to advanced, or by topic area. Each answer should be
↪  self-contained and understandable without reference to other answers. Ensure
↪  the FAQ works as a standalone document. Output only the FAQ, nothing else.
Document:
[TEXT]
```

table.

```
Rewrite the document as a structured table that organizes the key information,
↪  then generate one question-answer pair based on the table. First extract the
↪  main data points and organize them into a clear table format with
↪  appropriate headers using markdown table syntax with proper alignment. After
↪  the table, generate one insightful question that can be answered using the
↪  table data. Provide a clear, concise answer to the question based on the
↪  information in the table. Output only the table followed by the
↪  question-answer pair, nothing else.
Document:
[TEXT]
```

tutorial.

```
Rewrite the document as a clear, step-by-step tutorial or instructional guide.
↪  Use numbered steps or bullet points where appropriate to enhance clarity.
↪  Preserve all essential information while ensuring the style feels didactic
↪  and easy to follow. Output only the tutorial, nothing else.
```

```
Document:
[TEXT]
```

C Supplementary Results

- Table 10: Compares existing web-crawled and synthetic datasets (e.g., DCLM, FineWeb, Cosmopedia).
- Table 11: Evaluates existing rephrasing methods such as *Diverse QA Pairs*, *Distill*, *Knowledge List*, and *Guided Rewriting*.
- Table 12: Provides a detailed breakdown of performance across the four new pedagogical structured prompts (`math`, `faq`, `table`, `tutorial`).
- Table 13: Examines the impact of model scaling on performance across 270M to 27B parameter regimes using Gemma 3 models.
- Table 14: Reports mean scores and standard deviations across major model families.
- Table 15: Compares purely synthetic training against mixed training ($\mathcal{D}_{\text{mix}}$) for each specialized prompt.
- Table 16: Analyzes the performance sensitivity relative to the quality of mix-in data ($\mathcal{D}_{\text{mix}}$).
- Table 17: Investigates the performance sensitivity relative to the choice of the source dataset ($\mathcal{D}_{\text{source}}$).
- Table 18: Summarizes macro-averaged scores for model scales using the guided rewrite baseline.
- Table 19: Provides a multi-dimensional macro-average comparison across data quality, model scale, and prompt type.

Table 10: Breakdown scores ($\times 100$) for different existing datasets at 10K steps. $\star$ indicates a synthetic dataset.

	General Knowledge		Reading Comp.		Reasoning		NLU			Math	Table Understanding	
Approach	**ARC**	**MMLU**	**DROP**	**SQuAD v2**	**OBQA**	**XCSQA**	**WG**	**PIQA**	**HS**	**GSM8K**	**WikiTQ**	**TQA**
DCLM	15.26	12.78	16.55	42.56	2.94	4.84	3.90	7.42	9.25	10.87	19.31	13.76
Ultra-FineWeb	15.04	12.09	14.67	40.44	2.09	2.83	3.63	6.96	9.03	10.32	17.91	14.62
FineWeb-HQ	17.31	12.15	12.33	34.54	2.37	2.41	2.63	5.73	7.51	10.85	15.01	11.62
FineWeb-LQ	3.87	6.97	12.17	23.86	0.78	0.69	2.45	6.30	8.95	9.48	12.75	10.83
Nemotron-HQ-Synth$\star$	18.33	14.30	16.39	42.56	2.67	4.36	4.34	6.98	8.13	10.53	17.99	10.04
REWIRE$\star$	14.87	12.12	14.04	46.08	2.71	5.65	2.69	7.32	8.20	10.75	20.33	11.60
Cosmopedia$\star$	10.99	9.72	13.85	37.40	1.30	3.23	0.64	5.76	6.85	10.78	13.69	2.34
SYNTH$\star$	10.87	11.23	12.68	32.27	1.13	1.26	1.54	3.56	4.71	10.05	16.93	6.07

Table 11: Breakdown scores for different existing rephrasing approaches.

	General Knowledge		Reading Comp.		Reasoning		NLU			Math	Table Understanding	
Approach	**ARC**	**MMLU**	**DROP**	**SQuAD v2**	**OBQA**	**XCSQA**	**WG**	**PIQA**	**HS**	**GSM8K**	**WikiTQ**	**TQA**
DCLM	15.26	12.78	16.55	42.56	2.94	4.84	3.90	7.42	9.25	10.87	19.31	13.76
Diverse QA Pairs	21.56	13.58	17.62	51.65	3.03	3.06	3.38	5.86	7.51	10.91	17.93	12.03
Distill	20.28	13.58	14.88	39.65	2.64	3.19	3.27	6.20	7.53	10.94	16.38	12.61
Wikipedia	19.46	13.52	14.24	42.90	2.61	2.91	3.60	5.90	7.40	10.41	16.38	11.77
Knowledge List	19.96	14.06	17.83	39.91	3.06	2.94	2.46	6.03	7.32	11.16	13.96	11.52
Extract Knowledge	17.52	12.69	13.56	34.45	2.33	2.68	3.57	5.60	6.97	10.49	14.50	10.45
Guided Rewrite	19.63	13.39	16.83	44.95	2.80	3.34	3.16	5.69	7.44	10.97	17.60	11.93
Summarize	19.51	13.45	16.43	39.55	2.78	2.92	2.88	5.57	7.18	11.20	16.43	10.83
Continue	20.49	13.82	15.18	43.27	3.09	3.88	4.07	6.08	7.52	10.88	17.94	12.03

Table 12: Breakdown scores for new pedagogical structured prompts vs. DCLM.

	General Knowledge		Reading Comp.		Reasoning		NLU			Math	Table Understanding	
Approach	**ARC**	**MMLU**	**DROP**	**SQuAD v2**	**OBQA**	**XCSQA**	**WG**	**PIQA**	**HS**	**GSM8K**	**WikiTQ**	**TQA**
DCLM	15.26	12.78	16.55	42.56	2.94	4.84	3.90	7.42	9.25	10.87	19.31	13.76
`math`	20.73	14.37	16.04	53.81	2.89	3.76	3.12	6.11	7.68	12.36	20.30	14.50
`faq`	21.57	14.22	16.50	50.20	3.47	3.27	2.74	6.09	7.70	10.85	16.80	13.46
`table`	22.76	14.66	14.64	47.37	3.23	3.91	2.88	5.94	7.38	11.24	21.61	15.45
`tutorial`	21.34	14.11	17.98	45.49	3.06	3.51	3.71	6.07	7.71	10.54	17.29	14.42

Table 13: Breakdown scores for different Gemma 3 model scales.

	General Knowledge		Reading Comp.		Reasoning		NLU			Math	Table Understanding	
Approach	ARC	MMLU	DROP	SQuAD v2	OBQA	XCSQA	WG	PIQA	HS	GSM8K	WikiTQ	TQA
DCLM	15.26	12.78	16.55	42.56	2.94	4.84	3.90	7.42	9.25	10.87	19.31	13.76
270M	19.85	13.61	18.57	44.42	2.87	2.70	2.90	5.90	7.40	10.60	17.31	12.70
1B	20.73	14.37	16.04	53.81	2.89	3.76	3.12	6.11	7.68	12.36	20.30	14.50
4B	19.32	14.35	19.24	46.03	3.15	3.43	3.32	5.60	7.44	13.51	20.88	15.03
12B	19.77	13.97	17.30	43.39	3.15	3.01	3.83	5.67	7.44	13.96	19.86	14.84
27B	20.81	14.82	17.11	42.08	3.36	2.84	4.25	5.80	7.19	14.30	20.51	13.79

Table 14: Breakdown scores for different model families across the four new prompts (all ∼1B scale) with standard deviations in parentheses.

	General Knowledge		Reading Comp.		Reasoning		NLU			Math	Table Understanding	
Approach	ARC	MMLU	DROP	SQuAD v2	OBQA	XCSQA	WG	PIQA	HS	GSM8K	WikiTQ	TQA
DCLM	15.26 (0.00)	12.78 (0.00)	16.55 (0.00)	42.56 (0.00)	2.94 (0.00)	4.84 (0.00)	3.90 (0.00)	7.42 (0.00)	9.25 (0.00)	10.87 (0.00)	19.31 (0.00)	13.76 (0.00)
SmolLM2	22.47 (0.46)	15.34 (0.14)	18.86 (1.96)	59.67 (1.58)	3.63 (0.16)	4.84 (0.11)	3.79 (0.18)	6.48 (0.13)	7.82 (0.15)	11.71 (1.06)	20.48 (0.81)	16.80 (2.64)
Falcon 3	22.20 (0.44)	15.37 (0.23)	17.29 (2.45)	51.55 (2.39)	3.30 (0.13)	4.59 (0.13)	3.73 (0.23)	6.33 (0.24)	7.75 (0.12)	11.33 (0.54)	20.45 (1.49)	16.00 (1.72)
Qwen 3	22.29 (2.07)	15.02 (0.79)	15.47 (1.07)	46.24 (3.00)	3.28 (0.15)	3.84 (0.56)	3.51 (0.21)	6.11 (0.22)	7.54 (0.16)	11.47 (0.68)	18.76 (0.72)	13.30 (1.56)
Gemma 3 1B	21.60 (0.74)	14.34 (0.21)	16.29 (1.19)	49.22 (3.14)	3.16 (0.21)	3.61 (0.24)	3.11 (0.37)	6.05 (0.06)	7.62 (0.14)	11.25 (0.69)	19.00 (2.02)	14.46 (0.70)
Granite 3	20.49 (0.39)	14.66 (0.19)	17.20 (2.04)	50.90 (2.99)	3.27 (0.17)	4.01 (0.25)	3.02 (0.42)	6.28 (0.14)	7.71 (0.09)	11.20 (0.79)	18.43 (1.54)	14.56 (2.23)
LLaMA 3.2	21.32 (0.85)	14.40 (0.30)	16.79 (1.29)	49.26 (3.37)	3.23 (0.18)	4.03 (0.15)	3.29 (0.26)	5.92 (0.20)	7.49 (0.19)	11.35 (0.79)	19.52 (1.33)	13.93 (1.57)

Table 15: Breakdown scores between synthetic-only and mixed training.

	General Knowledge		Reading Comp.		Reasoning		NLU			Math	Table Understanding	
Approach	ARC	MMLU	DROP	SQuAD v2	OBQA	XCSQA	WG	PIQA	HS	GSM8K	WikiTQ	TQA
DCLM	15.26	12.78	16.55	42.56	2.94	4.84	3.90	7.42	9.25	10.87	19.31	13.76
`math` w/o $\mathcal{D}_{\text{mix}}$	20.24	14.02	17.23	54.07	2.64	2.96	3.02	5.76	6.99	12.36	20.34	14.26
`math`	20.73	14.37	16.04	53.81	2.89	3.76	3.12	6.11	7.68	12.36	20.30	14.50
`faq` w/o $\mathcal{D}_{\text{mix}}$	21.34	13.02	16.28	40.95	2.47	3.24	2.70	5.55	6.63	10.83	16.24	11.11
`faq`	21.57	14.22	16.50	50.20	3.47	3.27	2.74	6.09	7.70	10.85	16.80	13.46
`table` w/o $\mathcal{D}_{\text{mix}}$	21.73	13.68	15.54	46.61	2.63	3.62	2.91	5.02	6.05	10.91	17.03	12.00
`table`	22.76	14.66	14.64	47.37	3.23	3.91	2.88	5.94	7.38	11.24	21.61	15.45
`tutorial` w/o $\mathcal{D}_{\text{mix}}$	20.62	12.65	15.67	33.88	2.38	2.93	2.64	5.66	6.39	10.82	14.98	12.03
`tutorial`	21.34	14.11	17.98	45.49	3.06	3.51	3.71	6.07	7.71	10.54	17.29	14.42

Table 16: Breakdown scores for different mix-in dataset $\mathcal{D}_{\text{mix}}$ choices with FWHQ (HQ) or FWLQ (LQ) as $\mathcal{D}_{\text{source}}$. Parentheses indicate the mix-in dataset quality (HQ vs. LQ).

	General Knowledge		Reading Comp.		Reasoning		NLU			Math	Table Understanding	
Approach	ARC	MMLU	DROP	SQuAD v2	OBQA	XCSQA	WG	PIQA	HS	GSM8K	WikiTQ	TQA
FineWeb-LQ	3.87	6.97	12.17	23.86	0.78	0.69	2.45	6.30	8.95	9.48	12.75	10.83
FineWeb-HQ	17.31	12.15	12.33	34.54	2.37	2.41	2.63	5.73	7.51	10.85	15.01	11.62
Cosmopedia (LQ)	11.20	9.28	14.26	38.40	1.31	1.66	1.13	5.99	6.93	10.73	14.48	6.85
Cosmopedia (HQ)	19.53	12.74	13.24	43.39	2.38	2.85	2.20	6.33	7.37	11.26	16.26	10.97
DCLM (LQ)	12.37	11.20	16.54	37.59	2.70	3.76	3.99	7.28	8.97	10.36	17.12	13.83
DCLM (HQ)	20.34	13.91	15.67	43.10	3.08	4.95	4.00	7.28	8.87	10.59	19.97	14.56
FineWeb-HQ (LQ)	18.34	12.84	14.60	38.16	2.32	2.92	3.69	6.78	8.14	10.31	18.69	12.81
FineWeb-HQ (HQ)	21.34	14.11	17.98	45.49	3.06	3.51	3.71	6.07	7.71	10.54	17.29	14.42
FineWeb-LQ (LQ)	5.73	7.56	10.98	29.96	1.35	1.84	2.75	6.94	8.39	9.36	14.56	10.19
FineWeb-LQ (HQ)	18.30	12.34	16.15	44.62	2.35	3.59	3.27	7.22	8.75	10.18	18.17	12.54

Table 17: Breakdown scores for different source dataset $\mathcal{D}_{\text{source}}$ choices. Parentheses indicate the mix-in dataset used (Cosmopedia, DCLM, or FWHQ/FWLQ).

	General Knowledge		Reading Comp.		Reasoning		NLU			Math	Table Understanding	
Approach	ARC	MMLU	DROP	SQuAD v2	OBQA	XCSQA	WG	PIQA	HS	GSM8K	WikiTQ	TQA
Cosmopedia (FWHQ)	20.33	14.17	15.01	45.23	2.84	3.69	3.65	6.95	8.26	10.79	16.44	12.91
Cosmopedia (Cosmopedia)	11.85	9.48	11.25	35.98	1.36	2.95	1.25	5.56	6.64	10.76	14.06	5.76
DCLM (FWHQ)	20.48	13.78	18.16	46.93	2.83	4.23	4.36	7.08	8.57	10.49	18.09	12.80
DCLM (DCLM)	16.26	12.45	14.17	42.69	2.27	5.37	4.39	7.50	9.16	10.28	20.11	14.36
FineWeb-HQ (FWHQ)	21.34	14.11	17.98	45.49	3.06	3.51	3.71	6.07	7.71	10.54	17.29	14.42
FineWeb-LQ (FWHQ)	18.34	12.84	14.60	38.16	2.32	2.92	3.69	6.78	8.14	10.31	18.69	12.81
FineWeb-LQ (FWLQ)	5.73	7.56	10.98	29.96	1.35	1.84	2.75	6.94	8.39	9.36	14.56	10.19

Table 18: Macro-averaged scores for different Gemma 3 model scales at 10K steps when using the guided rewrite prompt.

Scale	Macro Avg
DCLM	13.77
270M	13.29
1B	13.73
4B	14.58
12B	14.33
27B	14.24

Table 19: Macro-averaged scores for different Gemma 3 model scales at 10K steps.

Source Quality	Scale	Prompt	Macro Avg
HQ	1B	Continue	13.73
HQ	12B	Continue	14.79
LQ	1B	Continue	13.52
LQ	12B	Continue	13.40
HQ	1B	`faq`	14.45
HQ	12B	`faq`	13.87
LQ	1B	`faq`	14.39
LQ	12B	`faq`	13.28
HQ	1B	Summarize	13.02
HQ	12B	Summarize	11.54
LQ	1B	Summarize	12.49
LQ	12B	Summarize	11.52
HQ	1B	`tutorial`	14.30
HQ	12B	`tutorial`	14.23
LQ	1B	`tutorial`	12.99
LQ	12B	`tutorial`	13.43

D Supplementary Analysis

Qwen 3 Representative Output.

```
**Problem:**
A disc rotates at 120 rpm. How many revolutions in 5 minutes?

**Solution:**
1. Revolutions per minute = 120
2. Number of minutes = 5
3. Total revolutions = 120 × 5

$$120 \\times 5 = 600$$

The disc makes 600 revolutions in 5 minutes.
```

SmolLM2 Representative Output.

```
1. Continuous Wave Doppler Monitor (USCOM): The monitor uses a handheld probe
 ↪  that can measure stroke volume and fluid input to avoid overloading patients
 ↪  and diagnose problems earlier. It provides real-time data in seconds and can
 ↪  be done painlessly on the patient.
2. Noninvasive Cardiac Monitoring: This includes monitoring cardiac output,
 ↪  stroke volume, and fluid input using tools like the USCOM Noninvasive
 ↪  Cardiac Output Monitor.
3. Noninvasive Monitoring Techniques: Other alternatives include the LiDCOplus
 ↪  System for pulsed waveforms for monitoring and Arterial Pulse Contour
 ↪  Analysis for heart function and blood flow measurement.
Answer: The NICOM, NICO2, and Respironic systems are noninvasive methods for
 ↪  monitoring cardiac output. The NICOM uses BIOREACTANCE technology, the NICO2
 ↪  measures vital signs in real-time, and the Respironic monitors oxygen flow
 ↪  and provides fast patient feedback in critical settings.
```

E FinePhrase Samples

E.1 Example 1

We use s3://commoncrawl/crawl-data/CC-MAIN-2013-20/segments/1368701459211/warc/CC-MAIN-20130516105059-00001-ip-10-60-113-184.ec2.internal.warc.gz to generate samples using the four pedagogical prompts.

math.

```
Question: Who was Paul Revere Williams, and what was his significance as an
↪  architect?

Solution: Paul Revere Williams was a pioneering African-American architect who
↪  was the first black person to become a member of the American Institute of
↪  Architects in 1923. He is best known for designing homes and commercial
↪  buildings for celebrity clients in Los Angeles and around the world. By the
↪  time he died in 1980, he had created more than 2,500 buildings, including
↪  some 2,000 in Los Angeles, making him a significant figure in American
↪  architectural history.
```

faq.

```
A Trailblazing Black Architect Who Helped Shape L.A.

By early 1923, Paul Revere Williams had completed two important milestones:
↪  becoming the American Institute of Architects' first African-American member
↪  and becoming one of the first black architects licensed to practice. These
↪  actions led to the induction of the African-American Architect into the
↪  AIA's first Black Fellowship in 1957. Born in downtown Los Angeles in 1894,
↪  Williams grew up orphaned and was raised by a family friend, who advised him
↪  that he could achieve his goals by going to college and becoming an
↪  architect.
- The Early Years: Williams' career began in the early 1920s, and as the first
↪  black architect in Los Angeles, he built over 2,500 buildings. His designs
↪  often featured innovative features, such as patios as an extension of the
↪  house and hidden retractable screens.
- Celebrity Clients: Williams' clientele included famous people, such as Frank
↪  Sinatra, Lucille Ball, and Danny Thomas. Their homes reflected a luxurious,
↪  elegant style.
- Signature Staircase: Williams is renowned for his elaborate, curved
↪  staircases, which have become iconic symbols of classic California style.
- Impact on Community: Williams' homes highlighted his unique ability to make
↪  grand spaces feel welcoming and cozy, and many of his homes were located in
↪  predominantly white neighborhoods that were closed to people of color,
↪  highlighting issues of racism and discriminatory practices.
- Granddaughter's Legacy: Karen E. Hudson is chronicling Williams' life and work
↪  in her latest book, "Paul R. Williams: Classic Hollywood Style," which
↪  celebrates his legacy and the influence of his designs on the California
↪  aesthetic.
```

table.

````
```
Table: Notable Clients of Paul Williams

Architect       Clients
Frank Sinatra
Lucille Ball
````
````

```
Desi Arnaz
Cary Grant
Bill Cosby        Ellen DeGeneres        Andy Garcia


Question: Which celebrities and personalities have used and appreciated Paul
↪  Williams' designs in their homes and properties?

Answer: Frank Sinatra, Lucille Ball, Desi Arnaz, Cary Grant, Bill Cosby, Ellen
↪  DeGeneres, Andy Garcia, and Danny Thomas have used and appreciated Paul
↪  Williams' designs in their homes and properties.
```

tutorial.

```
Title: Paul Revere Williams: Pioneer Architect of Southern California

Abstract: This step-by-step tutorial illustrates the contributions of esteemed
↪  black architect Paul Revere Williams to the field of art and design,
↪  particularly in Los Angeles. Through his remarkable success, Williams was
↪  the first black architect to become an AIA member in 1923 and inducted as
↪  the first black fellow in 1957. His clients included celebrities like Frank
↪  Sinatra, Lucille Ball, and Ellen Degeneres, and his work epitomizes notable
↪  characteristics such as luxurious curving staircases and patios that can be
↪  seen in his homes. This tutorial also highlights notable homes and clients
↪  of his, including the Beverly Hills Polo Lounge, and explains the challenges
↪  he faced because of racism during his career.
```

E.2 Example 2

We use s3://commoncrawl/crawl-data/CC-MAIN-2013-20/segments/1368711005985/warc/CC-MAIN-20130516133005-00005-ip-10-60-113-184.ec2.internal.warc.gz to generate samples using the four pedagogical prompts.

math.

```
1. Continuous Wave Doppler Monitor (USCOM): The monitor uses a handheld probe
↪  that can measure stroke volume and fluid input to avoid overloading patients
↪  and diagnose problems earlier. It provides real-time data in seconds and can
↪  be done painlessly on the patient.
2. Noninvasive Cardiac Monitoring: This includes monitoring cardiac output,
↪  stroke volume, and fluid input using tools like the USCOM Noninvasive
↪  Cardiac Output Monitor.
3. Noninvasive Monitoring Techniques: Other alternatives include the LiDCOplus
↪  System for pulsed waveforms for monitoring and Arterial Pulse Contour
↪  Analysis for heart function and blood flow measurement.
Answer: The NICOM, NICO2, and Respironic systems are noninvasive methods for
↪  monitoring cardiac output. The NICOM uses BIOREACTANCE technology, the NICO2
↪  measures vital signs in real-time, and the Respironic monitors oxygen flow
↪  and provides fast patient feedback in critical settings.
```

faq.

```
Cardiac output is not typically utilized in pediatric patients because of
↪  technical restraints. However, continuous monitoring by noninvasive
↪  monitors can save lives. Here's the FAQ:

Question 1: What is cardiac output?
```

```
* Cardiac output measures the rate at which blood is pumped through the body, or
↪  cardiac stroke volume and cardiac output combined.

Question 2: Why is cardiac output crucial for pediatric patients?

* Pediatric patients are more fragile and can die quickly from hemodynamic
↪  instability, making accurate measurement of cardiac output vital.

Question 3: What noninvasive devices are available for pediatric cardiac output
↪  monitoring?

* USCOM has USCOM doppler monitor; LiDCO+ system; Respironics's NICO2 Monitor;
↪  and Cheetah's NICOM and Deltex's CardioQ.

Question 4: How do these devices work?**

* The USCOM device uses a probe on the neck for comfortable and easy placement.
* The LiDCO+ uses arterial pulse waveforms and power analysis to track cardiac
↪  output.
* The NICO2 System monitors both ventilator and respiratory variables.
* The CardioQ uses an Oesophageal Doppler probe.

Question 5: Which device is suitable for pediatric patients?**

* NICO2 System, which has proven accurate in pediatric studies.

Question 6: Are the devices easy to use?**

* Yes, all monitors are user-friendly, minimizing patient and healthcare worker
↪  discomfort.

Question 7: How do these devices help in diagnosing conditions?**

* They provide rapid, accurate readings for diagnosing conditions and adjusting
↪  treatment.

Question 8: Can these devices be used during various surgeries?**

* Yes, some devices are portable and can be used during surgeries.

Question 9: How do these devices improve patient outcomes?**

* They enable early, rapid and safe intervention, helping to correct hemodynamic
↪  instability and reducing hospital stays.

Question 10: Are these devices covered by insurance?**

* Some insurance plans cover these devices.

These FAQs answer the key questions about noninvasive monitoring for cardiac
↪  output and its advantages for critically ill pediatric patients. Each
↪  device's features and benefits are clearly explained to ensure easy
↪  understanding by readers.
```

table.

```
The document outlines several noninvasive methods for measuring cardiac output
↪  and fluid balance in patients. Each method is described as being easy to
↪  use, presenting a low risk for infection compared to PACs, and able to
↪  continuously monitor hemodynamic function.
```

```
The table summarizes the key information and provides a table for easy
↪  reference.

One question could be: How do the noninvasive monitoring systems, like the USCOM
↪  and LiDCO, measure cardiac output directly?
```

tutorial.

```
Alternatives to Invasive Cardiac Output Monitors Summary:

The company offers a monitor that uses a continuous wave doppler that can look
↪  right down into the aortic valve to check the size of the valve, and it can
↪  also look at the pulmonary valve. The monitor can measure stroke volume and
↪  evaluate fluid input, so as to avoid overloading the patient.

Bioreactance presented at ACC - The bio-reactance technology used in the NICOM
↪  system is demonstrated to provide accurate non-invasive cardiac output
↪  monitoring.

Deltex Medical - A minimally invasive alternative to PAC is proposed, providing
↪  real-time determination of cardiac function and optimized ventilation.

Deltex Medical CardioQ - A flow-based system providing non-invasive and safe
↪  monitoring of cardiac output and fluid optimization.

USCOM - Continuous doppler monitoring of aortic and pulmonary valves to measure
↪  cardiac output and detect fluid load. The monitor is easy to use and
↪  painless.

LiDCOplus - Noninvasive continuous cardiac output monitoring. Its application in
↪  pediatric ICUs is particularly beneficial.

Respironics - NICO2 Respiratory Profile Monitor for noninvasive monitoring of
↪  cardiovascular and respiratory parameters in critical care.

USCOM Summary: The USCOM doppler monitor is a continuous wave doppler that
↪  measures cardiac output and fluid load. The noninvasive nature and ease of
↪  use make it a valuable tool for pediatric ICUs.

LiDCOplus - The LiDCOplus system measures cardiac output by analyzing arterial
↪  pulse waveforms. Offers rapid and accurate monitoring of changes in cardiac
↪  output.

Respironics - The NICO2 Monitor provides real-time monitoring of multiple
↪  variables for cardiac and respiratory assessment. Reduces respiratory risks
↪  and allows for rapid treatment.

Deltex Medical - The CardioQ Oesophageal Doppler Monitor allows continuous,
↪  flow-based cardiac output monitoring with minimal risk of infection.

Cheetah Reliant - Noninvasive non-invasive continuous cardiac monitoring that
↪  provides insights into cardiac function. Allows for easy, safer and more
↪  effective therapeutic interventions.

Cheetah Reliant Summary: The NICOM is a non-invasive, continuous cardiac output
↪  monitoring system that measures stroke volume and oxygen saturation on a
↪  sub-second basis.
```

```
Deltex Medical Summary: The CardioQ system allows safe, minimally invasive
↪  monitoring of cardiovascular and respiratory function. Highly sensitive to
↪  cardiac output changes.

USCOM Summary: USCOM's doppler monitor uses continuous wave doppler waves to
↪  monitor aortic and pulmonary valves. Safe and easy to use.

LiDCOplus Summary: The LIIDCOplus system measures cardiac output by analyzing
↪  arterial pulse waveform. Provides early intervention and assessment for
↪  cardiac and respiratory conditions.

Respironics Summary: The NICO2 Monitor tracks 40 key variables, providing
↪  precise feedback. Allows for rapid intervention and optimizes ventilation
↪  parameters.

Deltex Medical Summary: The CardioQ Monitor, an economical and effective
↪  non-invasive cardiac output measurement system. Allows for personalized
↪  treatment of fluid resuscitation and ventilation optimization.

Cheetah Reliant Summary: This system measures cardiac output with minimal risk
↪  and can be easily placed in the oral cavity.

USCOM Key Features: Provides an accurate measurement of cardiac output. Can be
↪  placed in virtually any clinical setting. Offers easy and quick deployment.
↪  Safe and non-invasive. Offers rapid results. Allows personalized treatment
↪  of patients. Can be integrated with existing clinical systems. Allows early
↪  detection during admission, and rapid interventions upon discharge. Allows
↪  for better outcomes and quality of care. Allows for easier and less costly
↪  monitoring of patient health. Allows for more accurate diagnosis of cardiac
↪  and respiratory conditions.
```