KNOW YOUR SOURCES: DATA SELECTION MATTERS WHEN REWRITING FOR DATA-CONSTRAINED PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality data is a cornerstone of large language model pretraining, and document rewriting with instruction-tuned models offers a promising way to expand pretraining corpora amid data scarcity. Existing pipelines rewrite documents uniformly or gate rewrite candidates with a quality measure built for raw-data filtering. However, it is not evident whether quality scoring measures, built to select from web-data, are also suitable signals for corpus rewriting. We study this in the data-constrained pretraining setting, exactly where synthetic documents may be most valuable. We formulate source selection as balancing two competing objectives: **quality** and **diversity**. We compare 5 strategies on this balance, including a DISAGREEMENT-AWARE strategy that accounts for inter-scorer disagreements. We pretrain 1.5B-parameter models from scratch for a multi-epoch, 30B token trajectory, across multiple seeds for a total of 18 runs and 54 evaluated checkpoints. DIVERSITY-ORIENTED selection ranks first on aggregated commonsense tasks at every epoch, and DISAGREEMENT-AWARE selection leads on knowledge-intensive benchmarks from the second epoch onward. These findings establish source selection for rewriting as a problem distinct from raw-data filtering[1].

1 INTRODUCTION

Scaling laws dictate that larger language models require proportionally more pretraining data (Kaplan et al., 2020; Hoffmann et al., 2022), but the stock of high-quality, naturally occurring text is finite (Villalobos et al., 2024). Prior work has explored mining high-quality data with heuristic- and model-based signals (Penedo et al., 2023; Li et al., 2024; Penedo et al., 2024), and more recent efforts synthesize training data directly (Ben Allal et al., 2024). Among synthesis approaches, *document rewriting* has emerged as a promising approach between mining and outright synthesis, where a rewriter improves a document's linguistic quality, style, and diversity while preserving its semantic content (Maini et al., 2024; Nguyen et al., 2025; Su et al., 2025).

However, existing rewriting pipelines rarely treat source selection as a first-class decision. Instead, source selection for rewriting is often based on existing signals: documents may be rewritten uniformly (Maini et al., 2024), rewritten and subsequently filtered by a quality score to retain only high-quality outputs (Nguyen et al., 2025), or partitioned into quality tiers that receive different treatments (Su et al., 2025). In each case, when rewriting is selectively applied or outputs are selectively retained, the decision is typically governed by quality alone. This formulation leaves two concerns unaddressed. First, a limited stock of high-quality text makes multi-epoch training increasingly common (Muennighoff et al., 2023; Fang et al., 2025). However, it remains unclear whether rewriting gains persist under repeated exposure to the same data. Second, a single raw-data quality score provides only one signal for allocating a limited rewriting budget. A quality score, such as the fastText classifier used to filter DCLM (Joulin et al., 2017; Li et al., 2024) or the FineWeb-Edu classifier (Penedo et al., 2024), ranks documents as they currently appear, not their value once rewritten. Ranking by such a score may therefore misallocate a limited rewriting budget. Together, these concerns converge on a single question: **which source documents should be rewritten to yield benefits that persist?**

[1]Data, checkpoints, and analysis will be released at https://huggingface.co/collections/anonymous

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

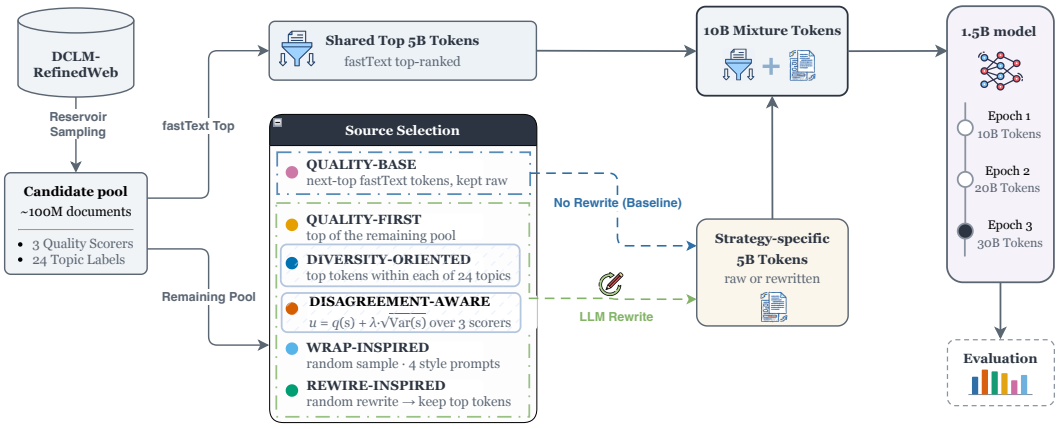


Figure 1: **Overview of the experimental pipeline.** We first reserve a 5M-document sample from DCLM-RefinedWeb for scorer and topic analysis and exclude it from training-data construction. From the remaining corpus, every setting combines a shared 5B-token base with around 5B strategy-specific tokens. Each 1.5B model is evaluated at three matched training horizons of approximately 10B, 20B, and 30B tokens.

In this work, we isolate source selection as an independent design choice in pretraining with rewritten documents. We motivate this with analysis showing that single scorers skew selection towards certain domains, reducing coverage. Furthermore, different quality metrics can disagree substantially. Source selection therefore involves a trade-off between two competing objectives: **quality** and **diversity**[2]. We test five strategies that balance this trade-off differently and compare them against a shared non-rewritten baseline: (1) QUALITY-FIRST rewriting under a single scorer, following the principle that high-quality sources benefit most from rewriting (Su et al., 2025; Almeida et al., 2026), (2) WRAP-INSPIRED, applying WRAP-style uniform rewriting across diverse format prompts (Maini et al., 2024), (3) REWIRE-INSPIRED post-hoc filtering that defers quality selection until after rewriting (Nguyen et al., 2025), (4) a DIVERSITY-ORIENTED strategy that preserves the corpus topic distribution, and (5) a DISAGREEMENT-AWARE strategy that combines quality with a bonus where scorers disagree. Among them, DIVERSITY-ORIENTED and DISAGREEMENT-AWARE have not been explored as document-level source-selection strategies for rewriting under a fixed rewriting budget.

We compare six settings in a controlled three-epoch experiment (Figure 1), pretraining 1.5B-parameter models from scratch on an approximately 10B-token mixture repeated three times, for about 30B tokens in total. This yields three key findings: (1) every rewriting strategy improves over the baseline on commonsense benchmarks in the first epoch, though these advantages largely vanish as the data repeat; (2) in our setting, the two benchmark families favor different selection strategies. Topic-preserving selection has a commonsense lead in all three epochs, while disagreement-based selection is strongest on knowledge-intensive benchmarks from the second epoch onward; (3) ranking by a quality scorer never leads by the final epoch, applied either before or after rewriting. Our results indicate that in data-constrained pretraining, *source selection should be considered a design choice, distinct from raw-data filtering.*

2 RELATED WORK

2.1 QUALITY SIGNALS FOR DATA SELECTION

Large-scale pretraining data curation has traditionally relied on heuristic filtering pipelines (Raffel et al., 2020; Penedo et al., 2023; Soldaini et al., 2024). As web-scale corpora grew, perplexity under a lightweight language model emerged as a scalable proxy for quality (Heafield, 2011; Wenzek et al., 2020; Marion et al., 2023; Ankner et al., 2025), including work that matches topic distributions to a target corpus (Feng et al., 2022). Since perplexity remains an indirect measure of quality, recent

[2]We use diversity to indicate how well a selection matches the topical distribution of the source pool.

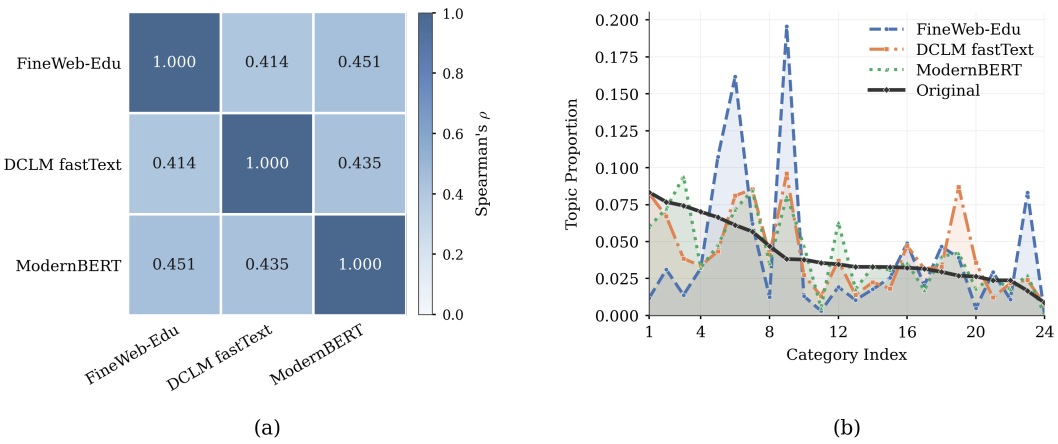


Figure 2: **Scorer behavior on a 5M-document sample from DCLM-RefinedWeb**. (a) Spearman rank correlations among the three quality scorers. (b) Topic distribution of each scorer's top-10% subset versus the full sample (24 WebOrganizer categories).

pipelines employ learned quality estimators, from fastText classifiers (Joulin et al., 2017; Li et al., 2024) to LLM-based scorers of informativeness or educational value (Sachdeva et al., 2024; Penedo et al., 2024), yielding strong gains on knowledge- and reasoning-intensive benchmarks (Li et al., 2024; Penedo et al., 2024). However, quality is multidimensional, with different criteria inducing different document rankings (Wettig et al., 2024). Individual quality filters can shift the topical composition (Wettig et al., 2025), and diversity-aware selection improves downstream performance (Tirumala et al., 2023). This tension between quality and diversity carries over to the rewriting setting, where the choice of source documents can shape both the topical coverage and the attainable quality of the resulting synthetic corpus.

2.2 Data Rewriting for Pretraining

In addition to filtering, which removes undesirable data, rewriting uses instruction-tuned language models to transform raw web documents into potentially higher-utility forms, expanding the supply of usable pretraining text. WRAP demonstrated that augmenting web data with synthetic rephrases can substantially accelerate pretraining (Maini et al., 2024). Subsequent works extend this recipe along two axes. On the data side, rewriting has been conditioned on source quality, through tier-specific strategies (Su et al., 2025), recycling of documents discarded by model-based filters (Nguyen et al., 2025), and analyses of how rewriting gains depend on source quality (Almeida et al., 2026). On the method side, prompt design, generator model, source corpus, and data mixture have been compared systematically (Niklaus et al., 2026). In these studies, source data is varied only at the corpus or quality-tier level; document-level selection under a fixed rewriting budget, and its effect on corpus-level pretraining gains, remains unexplored.

2.3 Repetition under Data Constraints

Neural scaling laws prescribe that training tokens scale with model size for compute-optimal pretraining (Kaplan et al., 2020; Hoffmann et al., 2022), implying that pretraining will increasingly operate in data-constrained regimes where data repetition becomes necessary. Muennighoff et al. (2023) showed that up to four epochs of repeated data match training on unique data, with a data-constrained scaling law capturing the diminishing value of repetition beyond this regime. Repetition also interacts with selection: repeating an aggressively filtered dataset can outperform single-epoch training on its larger superset (Fang et al., 2025), and repeated training on carefully selected subsets can outperform randomly selected new data (Tirumala et al., 2023). However, whether the gains from document-level selection in a rewritten corpus persist across epochs remains unexamined, motivating our multi-epoch design.

Table 1: **Quality scorers and their selection profiles.** Topic shift is the Jensen–Shannon divergence in bits between each scorer's top-10% subset and the full-sample topic distribution over 24 WebOrganizer categories. OH: OpenHermes-2.5; our scorer is detailed in Appendix A.

	Construction				Selection profile	
Scorer	**Backbone**	**Head**	**Supervision**	**Labels**	**JS**	**TV**
DCLM fastText	n-gram	classifier	OH + ELI5	binary	0.049	0.204
FineWeb-Edu	Arctic-M	linear	Llama-3-70B	0–5	0.179	0.422
Ours	ModernBERT	ridge	Claude Haiku	0–5	**0.040**	**0.181**

3 Data Properties

We first draw a 5M-document sample from DCLM-RefinedWeb for scorer and topic analysis. We exclude this sample from subsequent training-data construction. From the remaining corpus, we form the shared top-5B fastText base and the candidate pool used for strategy-specific selection. This analysis exposes **quality** and **diversity** as competing objectives in source selection. Throughout, quality refers to a document's score under these scorers, and diversity to how closely the topic distribution of the selected sources matches that of the full pool, measured over the 24 WebOrganizer categories (Wettig et al., 2025).

We compare DCLM fastText (Joulin et al., 2017; Li et al., 2024), FineWeb-Edu (Penedo et al., 2024), and a ModernBERT-based scorer (Warner et al., 2025) trained for this study. They use different supervision signals and therefore capture different notions of quality. Table 1 summarizes their construction and selection profiles.

To construct the ModernBERT scorer, we annotate roughly 50,000 documents, primarily from DCLM-RefinedWeb with a small STEM supplement, using Claude Haiku 4.5 (Anthropic, 2025) under a five-criterion rubric and fit a ridge-regression head on frozen ModernBERT embeddings. The resulting scorer reaches a Spearman correlation of 0.731 with held-out annotations; the rubric, construction, and evaluation details are available in Appendix A.

Source selection is scorer-dependent. Figure 2 reports pairwise Spearman correlations of 0.414 – 0.451 among the three scorers. The correlations are positive but moderate, showing that the same document can receive different ranks under different scorers. **This disagreement in ranking translates directly into disagreement in selection: a top-k gate over each ranking yields a different subset.** The composition of the rewrite pool therefore depends on the choice of scorer, not on document quality alone.

Quality filtering shifts topic coverage. We classify every document into one of 24 WebOrganizer categories (Wettig et al., 2025). Figure 2 compares the topic distribution of each scorer's top-10% subset with the full sample, while Table 1 summarizes the resulting distribution shifts. All three subsets depart from the original distribution, but to markedly different degrees. FineWeb-Edu shifts furthest, with a Jensen–Shannon divergence (JS) of 0.179 from the full-sample distribution, reallocating over 40% of the topic mass toward the educational domains its supervision targets. DCLM fastText and ModernBERT move far less (JS $\approx$ 0.04 – 0.05), yet each still redistributes roughly a fifth of the topic mass in systematic, scorer-specific directions (Appendix B). Source selection can therefore carry scorer-specific topic skew into the rewritten corpus.

Quality and diversity compete. Document-level **quality** asks which sources are most valuable under a given criterion; corpus-level **diversity** asks how the rewriting budget is spread across topics. The two objectives compete: quality ranking inherits the chosen scorer's domain preferences, whereas topic-aware selection preserves coverage at the cost of admitting lower-scoring sources. Therefore, we argue that source selection is a trade-off between quality and diversity. This motivates the five source-selection strategies of Section 4.2, each approaching this trade-off in different ways. We compare them in controlled multi-epoch pretraining experiments.

Table 2: **Overview of the six 10B-token training mixtures.** All settings share the same fixed 5B-token component selected from the top of the DCLM ranking and differ in the approximately 5B strategy-specific tokens that complete the mixture. Strategies are defined in Section 4.2; topics are the 24 WebOrganizer categories (Wettig et al., 2025).

Strategy	Notation	Data Composition (10B)	
		Fixed (5B)	Strategy-specific Tokens (5B)
QUALITY-BASE	–	DCLM top 5B	Next 5B fastText-ranked tokens, used verbatim.
QUALITY-FIRST	$\mathcal{S}_{\mathrm{QF}}$		Rewrites of top documents.
WRAP-INSPIRED	$\mathcal{S}_{\mathrm{WRAP}}$		Rewrites of random samples with format prompts.
REWIRE-INSPIRED	$\mathcal{S}_{\mathrm{RI}}$		Top rewrites from a $2\times$ larger random pool.
DIVERSITY-ORIENTED	$\mathcal{S}_{\mathrm{DIV}}$		Rewrites of documents within each topic.
DISAGREEMENT-AWARE	$\mathcal{S}_{\mathrm{DA}}$		Rewrites of documents ranked by $u_i = q_i + \lambda\sqrt{v_i}$.

4 SOURCE SELECTION FOR REWRITING

4.1 SETUP AND NOTATION

Candidate pool and token budgets. Let $\mathcal{D} = \{d_1, \ldots, d_N\}$ denote the candidate pool of raw web documents, from which sources for rewriting are selected. For a document d, $\ell(d)$ is the token length, and $\pi(\cdot)$ is the rewriting operator that maps a source document to its rewritten form. Two token budgets govern the pipeline, and both are held fixed across strategies. On the source side, a source selection strategy is a mapping that, given a budget of B_s tokens, returns a subset $\mathcal{S} \subseteq \mathcal{D}$ of B_s tokens. The selected sources are then transformed into a fixed target of B rewritten tokens. Since rewriting compresses its inputs, multiple prompt passes are generated, and rewritten outputs are retained in order until the target budget B is reached (Section 5). All strategies are compared at matched pretraining compute. The four non-WRAP strategies use the same two Nemotron-family rewriting prompts, while WRAP-INSPIRED serves as a recipe-level baseline with diverse format transformations.

Quality and Diversity. To characterize each candidate document for selection, we annotate it with the three quality scorers introduced in Section 3: the DCLM fastText classifier (Joulin et al., 2017; Li et al., 2024), the FineWeb-Edu scorer (Penedo et al., 2024), and our ModernBERT scorer (Warner et al., 2025) (Appendix A), yielding raw scores $x_i^{(s)}$, $s \in \{1, 2, 3\}$. Because the scores lie on incomparable scales and contain ties, each is mapped to its tie-aware global percentile

$$r_i^{(s)} \;=\; \frac{\mathrm{rank}\big(x_i^{(s)};\, \{x_j^{(s)}\}_{j=1}^{N}\big)}{N} \;\in\; (0, 1], \tag{1}$$

where tied documents receive the mean of the ranks they would occupy. From the percentiles we derive the consensus quality and scorer disagreement

$$q_i \;=\; \mathrm{Mean}_s\big(r_i^{(s)}\big), \qquad v_i \;=\; \mathrm{Var}_s\big(r_i^{(s)}\big). \tag{2}$$

Diversity is measured over topics (Section 3), so each document additionally carries a topic label $t_i \in \{1, \ldots, 24\}$ from the WebOrganizer taxonomy (Wettig et al., 2025), which the DIVERSITY-ORIENTED strategy below uses to control the topical composition of the selected sources.

4.2 SELECTION STRATEGIES

General Budget Selection Operator. We rank a candidate set by a scoring function and keep the top-scoring documents until the token budget is filled. Formally, given a candidate set $\mathcal{A}$, a scoring function f, and a token budget b, which takes the source budget B_s or the rewritten-output budget B depending on the strategy, we define

$$\mathrm{Top}_b(\mathcal{A}, f) \;=\; \big\{\, d_i \in \mathcal{A} : f(d_i) \geq \theta_b \,\big\}, \tag{3}$$

where θ_b is the largest threshold such that $\sum_{d \in \mathrm{Top}_b(\mathcal{A},f)} \ell(d) \geq b$. Length ℓ is measured on the elements of $\mathcal{A}$, i.e., source documents or rewritten outputs. The threshold θ_b is therefore not set in advance but determined by the budget. Ties at the budget boundary are broken as needed so that selection stops when the token budget is first reached.

The five strategies below instantiate equation 3 with different choices of the domain $\mathcal{A}$, the scoring function f, and the ordering of selection and rewriting.

QUALITY-FIRST. Documents are ranked over the entire candidate pool $\mathcal{D}$ by a single quality scorer and retained until the source budget B_s is filled:

$$\mathcal{S}_{\mathrm{QF}} \;=\; \mathrm{Top}_{B_s}\big(\mathcal{D},\, r^{(\mathrm{fastText})}\big), \tag{4}$$

where $r^{(\mathrm{fastText})} \in (0, 1]$ is a document's global percentile under the fastText scorer, with higher values indicating higher estimated quality. We use the DCLM fastText scorer as the primary single-scorer baseline because DCLM-RefinedWeb is the pre-fastText pool from the DCLM pipeline, and the original DCLM study found this scorer to outperform the other quality-filtering methods evaluated on this pool (Li et al., 2024). This instantiates the prevailing principle that high-quality sources benefit most from rewriting (Su et al., 2025; Almeida et al., 2026). Quality is maximized, and coverage is not controlled.

WRAP-INSPIRED. The WRAP-INSPIRED strategy $\mathcal{S}_{\mathrm{WRAP}}$ applies no source selection, and documents are sampled uniformly at random from the candidate pool $\mathcal{D}$ until the source budget B_s is reached. Quality is not scored, and coverage is not controlled. Each sampled document is rewritten under one of four style prompts adapted from WRAP (Maini et al., 2024), and is also rewritten with the shared *distill* prompt. We reimplement these prompts rather than reuse the original templates, and add to each a constraint against introducing new facts (Appendix C), so the strategy follows WRAP's principle of combining uniform source sampling with diverse format transformations, rather than reproducing its exact implementation.

REWIRE-INSPIRED. Selection is deferred until after rewriting. Documents are sampled uniformly at random from the candidate pool $\mathcal{D}$, but the sample is enlarged by an oversampling factor $\kappa > 1$, giving a pool $\mathcal{R}$ of κB_s tokens. Every document in $\mathcal{R}$ is then rewritten, yielding the rewritten set $\pi(\mathcal{R})$. From this set, the highest-scoring outputs are retained until the target budget B is filled:

$$\mathcal{S}_{\mathrm{RI}} \;=\; \mathrm{Top}_{B}\big(\pi(\mathcal{R}),\, r^{(\mathrm{fastText})}\big), \tag{5}$$

Here, fastText is applied again to each rewritten output in $\pi(\mathcal{R})$, so documents are ranked by the quality they attain after rewriting rather than before (Nguyen et al., 2025). Unlike the other strategies, $\mathcal{S}_{\mathrm{RI}}$ is a set of rewritten outputs rather than source documents, and is therefore sized by B rather than B_s; coverage is not controlled. This arm isolates the timing of quality selection rather than reproducing REWIRE, and incurs higher generation compute because it rewrites an oversampled pool before filtering.

DIVERSITY-ORIENTED. The candidate pool is partitioned by WebOrganizer topic label, and each topic is guaranteed a fixed share of the source budget. Topic c receives $p_c B_s$ tokens, where p_c is its share of pool tokens, and its documents $\mathcal{D}_c$ are ranked by consensus quality q_i within the topic.

$$\mathcal{S}_{\mathrm{DIV}} \;=\; \bigcup_{c} \mathrm{Top}_{p_c B_s}\big(\mathcal{D}_c,\, q\big). \tag{6}$$

As no topic can be crowded out by another, the selection reproduces the corpus topic distribution and avoids the skew that single-scorer selection induces (Section 3). It is the coverage extreme of the trade-off, with quality acting only within topic.

DISAGREEMENT-AWARE. For each scorer, we rank documents by its percentile score and retain the highest-ranked documents until a 10B-token budget is reached. This focuses the candidate pool on high-quality regions of each ranking, consistent with prior work on aggressive quality filtering (Li et al., 2024; Penedo et al., 2024). We define $\mathcal{U}$ as the union of these three selections. Within $\mathcal{U}$,

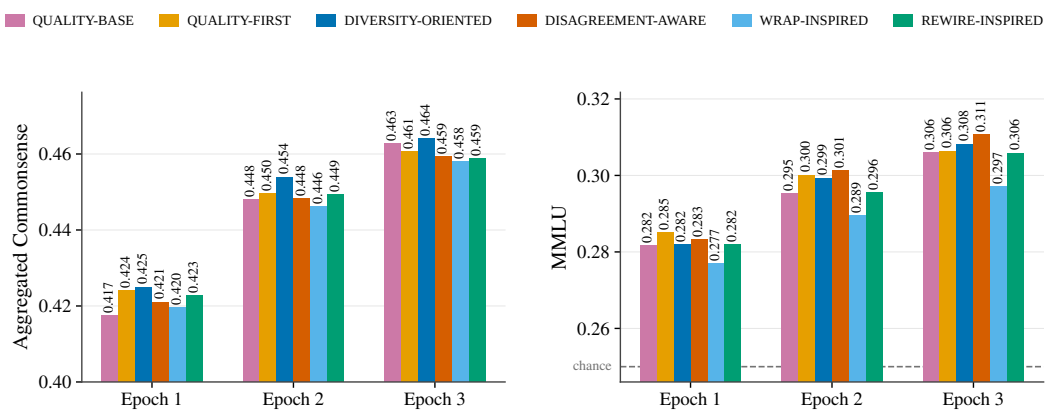


Figure 3: **Accuracy at the end of each epoch**, averaged over three seeds. (a) Mean over ARC-Easy, HellaSwag, PIQA, SIQA, OpenBookQA, and CommonsenseQA. (b) MMLU macro-average over 57 subsets; the dashed line marks 25% chance. Seed-level standard deviations and seed counts are in Appendix F.

a quality floor and a variance cap define the feasible domain $\mathcal{U}_\tau = \{\, d_i \in \mathcal{U} : q_i \geq \tau_q,\ v_i \leq \tau_v \,\}$, over which documents are ranked by the disagreement-augmented utility $u_i = q_i + \lambda\sqrt{v_i}$:

$$\mathcal{S}_{\text{DA}} \;=\; \text{Top}_{B_s}\big(\mathcal{U}_\tau,\, u\big). \tag{7}$$

Each component serves a distinct purpose. The variance bonus favors documents on which the scorers disagree; these are documents for which single-scorer rankings are less consistent (Section 3). The quality floor removes documents with low consensus quality. The variance cap serves as a safeguard against extreme scorer disagreement. The weight λ controls the balance between quality and disagreement: $\lambda = 0$ recovers quality-only ranking, and larger values favor disagreement. Its effect on document selection is analyzed in Appendix D. Unlike DIVERSITY-ORIENTED, this strategy does not explicitly constrain topic coverage; instead, it augments consensus quality with scorer disagreement, yielding a selection that lies between QUALITY-FIRST and DIVERSITY-ORIENTED in document overlap.

5 EXPERIMENTAL SETUP

5.1 DATA MIXTURES

Every setting trains on an approximately 10B-token mixture, consisting of a fixed 5B-token base of top-ranked DCLM fastText documents and around 5B strategy-specific tokens. The candidate pool $\mathcal{D}$ of Section 4 is the remainder after this base is removed, so every strategy, including QUALITY-FIRST, selects outside the base. Because rewriting compresses its inputs, producing $B = 5$B rewritten tokens requires a larger source pool, and we set the source budget to $B_s = 10$B tokens. The oversampling factor is $\kappa = 2$ and the disagreement weight is $\lambda = 0.5$; the quality floor and variance cap are set by percentile and analyzed in Appendix D. Every constant is fixed without reference to downstream performance. At $\lambda = 0.5$ the selected documents overlap 42.1% with QUALITY-FIRST and 61.2% with DIVERSITY-ORIENTED, against 11.1% expected under independence, so the strategy coincides with neither extreme. The six settings are summarized in Table 2.

5.2 REWRITING

We use *Qwen2.5-7B-Instruct* (Qwen et al., 2025) for rewriting with two prompts from Nemotron-CC (Su et al., 2025), Wikipedia-style rephrasing, and distillation, based on templates released in the FinePhrase repository (Niklaus et al., 2026). The generator and prompts were chosen based on a preliminary comparison that measured hallucination rate and compression ratio across generator

sizes (Appendix C). Neither prompt alters document format, and the same two prompts are used across the four non-WRAP rewriting strategies. The exception is WRAP-INSPIRED, which assigns each source document to one of four adapted style-specific prompts and uses the shared *distill* pass.

5.3 TRAINING

All models share a Llama 2 architecture (Touvron et al., 2023) with approximately 1.5B parameters and are trained with Nanotron (Hugging Face, 2024). We use the AdamW optimizer (Loshchilov & Hutter, 2019) and a warmup-stable-decay schedule (Hu et al., 2024), following the SmolLM2 recipe (Ben Allal et al., 2025). Full hyperparameters are in Appendix E. Each model is trained for approximately 30B tokens. This matches the compute-optimal ratio of about 20 tokens per parameter (Hoffmann et al., 2022). It also remains within the regime where repeated data retains much of the value of unique data (Muennighoff et al., 2023). Architecture, schedule, and token count are thus matched across settings; the mixtures differ only in their strategy-specific tokens.

5.4 EVALUATION

We evaluate each setting at three matched training horizons rather than at a single endpoint. Evaluation uses LightEval (Habib et al., 2023) on six commonsense benchmarks, reported as a mean: ARC-Easy (Clark et al., 2018), HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), OpenBookQA (Mihaylov et al., 2018), and CommonsenseQA (Talmor et al., 2019). We additionally report the results on MMLU (Hendrycks et al., 2021) as a knowledge-intensive benchmark. Multiple-choice options are scored by length-normalized log-likelihood, so longer options are not penalized.

6 MAIN FINDINGS

We compare five source-selection strategies against a no-rewriting baseline (Figure 3), asking how much benefit each strategy provides under multi-epoch training. While the absolute magnitude of score differences are small, we emphasize that the following results are for small models and are averaged across 3 seeds (for a total of 54 models evaluated). Prior work, e.g. Penedo et al. (2024), trains similarly sized models for both short and long durations and shows that carefully measured data interventions hold at multiple scales.

6.1 REWRITING GAINS ERODE WITH REPETITION

First epoch: every strategy beats the baseline. After a single pass, all five rewriting strategies exceed the non-rewritten QUALITY-BASE in mean accuracy over the six commonsense tasks (Figure 3a), by 0.2 to 0.8 accuracy points, consistent with the gains reported by prior recipes (Maini et al., 2024; Su et al., 2025).

Second and third epochs: the performance advantage shrinks. The advantage over the baseline, averaged across the five strategies, falls to 0.15 points at the second epoch and -0.25 at the third. By then QUALITY-FIRST, REWIRE-INSPIRED, WRAP-INSPIRED and DISAGREEMENT-AWARE all sit below the baseline, by 0.2 to 0.5 points, within seed-level noise, and each is positive in only one of three seeds (Appendix F).

Topic-preserving selection stays ahead throughout. DIVERSITY-ORIENTED ranks first on the six-task commonsense mean at all three epochs, the only setting to do so. Because suite averages can be dominated by individual tasks (Dehghani et al., 2021), we ask which tasks carry the margin. It is positive in all three seeds on HellaSwag at every epoch, and leave-one-out recomputation of the six-task average leaves the final-epoch margin between -0.14 and $+0.42$ points, turning negative when either HellaSwag or PIQA is removed.

> **Finding:** Single-pass comparisons measure how quickly a corpus is learned, not how much of its advantage survives reuse. Under repetition, topic-preserving selection decays more slowly than selection by quality rank.

6.2 SOURCE SELECTION DETERMINES WHICH BENCHMARK FAMILY REWRITING IMPROVES

We focus on MMLU after the third epoch as separation becomes clearer (Figure 3b). DISAGREEMENT-AWARE achieves the highest MMLU score at 31.1, 0.49 points above the baseline across all seeds, and gains in three of the four MMLU categories, with the smallest STEM loss of any setting (Appendix Table 15). It already led at the second epoch, 0.60 points above the baseline and positive in all three seeds. On aggregated commonsense benchmarks, however, it falls below the non-rewritten baseline, losing on every task except PIQA (Appendix F). DIVERSITY-ORIENTED shows the opposite pattern: leading on commonsense, with the largest gains on PIQA and HellaSwag. Although it also improves on MMLU, the gains come from Humanities, Social Sciences, and Other, while STEM is negative.

> **Finding:** DISAGREEMENT-AWARE leads on knowledge from the second epoch onward, DIVERSITY-ORIENTED on commonsense average at every epoch; in our setting, neither leads on both.

6.3 REWRITING SOURCE SELECTION IS DISTINCT FROM RAW-DATA FILTERING

Quality ranking alone is not the best use of rewriting budget. QUALITY-FIRST ranks documents by fastText, the scorer used to filter DCLM web data (Joulin et al., 2017; Li et al., 2024), and its only aggregate lead is on MMLU after a single epoch. REWIRE-INSPIRED applies the same scorer after rewriting, and ends indistinguishable from the baseline on MMLU and 0.4 points below it on commonsense. In our setting, quality ranking does not lead at the final epoch, whether applied before or after rewriting. Rewriting pipelines guided by a quality scorer have been reported to help at larger scale (Su et al., 2025; Nguyen et al., 2025; Almeida et al., 2026). At the scale we studied here, both strategies beat the non-rewritten baseline on commonsense tasks in the first epoch, but neither matches DIVERSITY-ORIENTED there at any epoch.

Neither leading strategy relies on a single quality ranking. DIVERSITY-ORIENTED explicitly addresses coverage, giving every topic a quota so that none are crowded out. DISAGREEMENT-AWARE instead favors the documents on which the three scorers disagree. These are documents for which the three scorer rankings differ.

> **Finding:** Choosing sources for rewriting is a different problem from filtering raw data. A quality score describes a document as it stands, not whether it should be prioritized for rewriting.

7 CONCLUSION

We investigate source selection in data-constrained rewriting. Through a controlled three-epoch comparison of 18 1.5B-parameter models pretrained from scratch on 30B tokens, we have three main findings: (1) every rewriting strategy initially improves over the baseline on commonsense benchmarks, but the advantage nearly vanishes under repetition, most slowly under topic-preserving selection; (2) DIVERSITY-ORIENTED selection ranks first on commonsense tasks at every epoch, while DISAGREEMENT-AWARE selection leads on MMLU from the second epoch onward, so selection strategy depends on target family; (3) ranking by quality alone never leads by the final epoch, both before and after rewriting. These findings establish source selection for rewriting as a problem distinct from raw-data filtering. Our results indicate that practitioners must carefully balance quality and coverage for data selection.

LIMITATIONS

Due to compute constraints, we train only 1.5B-parameter models on 30B tokens. At this scale, absolute scores are modest and differences between methods are small, so we draw all conclusions from relative comparisons under a fixed evaluation protocol: three seeds, one checkpoint per epoch, and the mean over six tasks. Our results suggest that factors beyond data quality, such as diversity, affect the value of rewriting. Verifying this at larger scale is an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and polishing. All research ideas, experimental design, analyses, and conclusions are the authors' own.

ETHICS STATEMENT

This work studies source selection for synthetic data rewriting in small-scale language-model pretraining. Our experiments build on publicly available web corpora, primarily DCLM-RefinedWeb, and inherit the limitations and potential biases of these source datasets. We do not make new claims regarding the licensing or ownership of the underlying data.

Synthetic rewriting may introduce unsupported or distorted information into training data. We therefore evaluate hallucination behavior during prompt selection and use grounding instructions in the production rewriting prompts. These checks reduce, but do not eliminate, the possibility of factual distortion. Our experiments train base language models for controlled evaluation rather than models intended for direct user interaction.

Source selection determines which parts of a web corpus are amplified through rewriting and can therefore affect the resulting data distribution. This motivates us to treat source selection as a distinct design choice rather than relying on quality filtering alone. Our experiments compare quality-oriented selection with strategies that preserve the original topic distribution or exploit disagreement across quality scorers.

REFERENCES

Thales Sales Almeida, Rodrigo Nogueira, and Hélio Pedrini. Synthetic rewriting as a quality multiplier: Evidence from Portuguese continued pretraining, 2026. URL https://arxiv.org/abs/2603.24826.

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, and Mansheej Paul. Perplexed by perplexity: Perplexity-based data pruning with small reference models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=1GTARJhxtq.

Anthropic. Claude haiku 4.5 system card. https://www.anthropic.com/claude/haiku, 2025. Accessed: 2026-08-03.

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

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM2: When Smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

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, 2020. doi: 10.1609/aaai.v34i05.6239.

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, 2018. URL https://arxiv.org/abs/1803.05457.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The benchmark lottery. *arXiv preprint arXiv:2107.07002*, 2021.

Alex Fang, Hadi Pouransari, Matt Jordan, Alexander Toshev, Vaishaal Shankar, Ludwig Schmidt, and Tom Gunter. Datasets, documents, and repetitions: The practicalities of unequal data quality, 2025. URL https://arxiv.org/abs/2503.07879.

Yukun Feng, Patrick Xia, Benjamin Van Durme, and João Sedoc. Automatic document selection for efficient encoder pretraining. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9522–9530, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.647. URL https://aclanthology.org/2022.emnlp-main.647/.

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

Kenneth Heafield. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pp. 187–197, Edinburgh, Scotland, 2011. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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. *Advances in Neural Information Processing Systems*, 35:30016–30030, 2022.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zhen Leng Thai, Kai Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. MiniCPM: Unveiling the potential of small language models with scalable training strategies, 2024. URL https://arxiv.org/abs/2404.06395.

Hugging Face. Nanotron: Minimalistic large language model 3D-parallelism training. https://github.com/huggingface/nanotron, 2024.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 427–431, Valencia, Spain, 2017. Association for Computational Linguistics.

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, 2020. URL https://arxiv.org/abs/2001.08361.

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 W. 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. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024. doi: 10.52202/079017-0455.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14044–14072, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.757.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining LLMs at scale, 2023. URL https://arxiv.org/abs/2309.04564.

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

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, 2025. URL https://arxiv.org/abs/2506.04689.

Joel Niklaus, Atsuki Yamaguchi, Michal Štefánik, Guilherme Penedo, Hynek Kydlíček, Elie Bakouch, Lewis Tunstall, Edward Emanuel Beeching, Thibaud Frere, Colin Raffel, Leandro von Werra, and Thomas Wolf. How can we synthesize high-quality pretraining data? A systematic study of prompt design, generator model, and source data, 2026. URL https://arxiv.org/abs/2604.13977.

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. *Advances in Neural Information Processing Systems*, 36:79155–79172, 2023.

Guilherme Penedo, Hynek Kydlíček, Loubna B. 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. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024. doi: 10.52202/079017-0970.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. How to train data-efficient LLMs, 2024. URL https://arxiv.org/abs/2402.09668.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454/.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.840.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2459–2475, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.123.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. D4: Improving LLM pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36:53983–53995, 2023.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49523–49544. PMLR, 2024. URL https://proceedings.mlr.press/v235/villalobos24a.html.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.127.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4003–4012, Marseille, France, 2020. European Language Resources Association.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. QuRating: Selecting high-quality data for training language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 52915–52971. PMLR, 2024.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the Web: Constructing domains enhances pre-training data curation, 2025. URL https://arxiv.org/abs/2502.10341.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472.

A MODERNBERT QUALITY SCORER DETAILS

We train a lightweight quality scorer that distills the judgments of a strong instruction-tuned model into a cheap, frozen-embedding regressor. The scorer is produced in three stages: (i) assembling a $\sim$50k-document training mixture, (ii) annotating each document with Claude Haiku 4.5 under a five-criterion quality rubric, and (iii) fitting a ridge-regression head on frozen ModernBERT embeddings. The resulting model is the one applied at scale to assign a continuous distilled quality score to every document drawn from DCLM-RefinedWeb.

A.1 TRAINING MIXTURE AND SAMPLING

The training pool contains $50{,}427$ documents drawn from three sources (Table 3). The bulk is $49{,}998$ web documents sampled from DCLM-RefinedWeb. We add a small, targeted set of mathematical and technical documents, 218 from OpenWebMath and 211 from algebraic-stack, so that the scorer is exposed to STEM-style text that is rare in raw web data.

The web documents are selected with a three-pass streaming selector as two equal halves: a uniform *random* sample of $24{,}999$ documents drawn by reservoir sampling over all documents longer than 200 characters, and a *length-diversified* sample of $24{,}999$ documents that partitions documents into five length buckets by character-count quantiles and caps any single URL domain at 5% within each bucket. For annotation, each document's text is truncated to $4{,}096$ characters.

A.2 CLAUDE HAIKU ANNOTATION

We annotate every document with Claude Haiku 4.5 via the batch API at temperature 0. Unlike the single-axis educational-value scale of FineWeb-Edu, our rubric is intentionally *broader*: it scores five independent binary criteria that together target general pretraining utility, not just educational content. The model awards one point per satisfied criterion, yielding an additive integer score in $\{0, \ldots, 5\}$. The exact annotation prompt is reproduced in Figure 4.

A.3 ENCODER AND HEAD

For scorer training, the same 4,096-character text shown to the annotator is encoded with a *frozen* nomic-ai/modernbert-embed-base model. We mean-pool the final hidden states over non-padding tokens, with a maximum length of $2{,}048$ tokens and fp16 inference, to obtain a 768-dimensional embedding, then L_2-normalize it. On top of these fixed embeddings we fit a closed-form ridge-regression head that predicts a continuous score from the discrete 0–5 annotations. Because the backbone is frozen and the head is a linear regressor, scoring is cheap at corpus scale.

A.4 EVALUATION

We evaluate with five-fold stratified cross-validation over ten score bins, repeated across the five random seeds $\{42, 123, 777, 2024, 2025\}$, giving 25 runs per configuration. We report the Spearman rank correlation ρ between predicted and held-out annotated scores, the mean-squared error, and Top-10% precision, computed as the overlap between the top-k predicted and annotated examples within each held-out fold, where $k = \lfloor 0.1n \rfloor$; ties at the annotation cutoff are resolved deterministically by the sorting routine. Table 4 compares two embedding backbones, each with and without L_2 normalization.

Table 3: **Training mixture for the distilled quality scorer.** A small set of mathematical and technical documents is added so that the scorer sees STEM-style text, which is rare in raw web data.

Source	**Documents**
DCLM-RefinedWeb (web)	49,998
OpenWebMath (math)	218
algebraic-stack (math)	211
Total	**50,427**

```
You are evaluating a text sample for the quality of language model pretraining.

Score the sample on five binary criteria. Give 1 point for each satisfied criterion.

Criteria:
1. coherence: The text is readable, well-formed, and not badly fragmented or repetitive.
2. informativeness: The text contains meaningful, non-trivial content rather than
   boilerplate, spam, or filler.
3. internal_reliability: The text is internally consistent and not obviously fabricated.
   Do NOT use outside knowledge.
4. lm_usefulness: The text is useful for learning language or knowledge patterns.
5. exceptional_value: The text shows especially high value (clear explanation, dense
   information, strong structure, or high-quality writing).

Scoring
Use the following additive 5-point scoring system. Add one point for each satisfied
criterion.
+1 point: The text is coherent and readable
+1 point: The text contains meaningful, non-trivial information
+1 point: The text appears reasonably reliable or consistent
+1 point: The text is useful for learning language or knowledge patterns
+1 point: The text shows strong reasoning, explanation, or especially high value

Guidelines:
- Focus only on the text itself
- Do NOT reward length
- Do NOT penalize technical content if clear
- Strongly penalize: navigation text, cookie banners, spam, link lists, SEO content,
  forum fragments, image metadata, duplicated patterns

Return ONLY valid JSON:
{
"coherence": 0 or 1,
"informativeness": 0 or 1,
"internal_reliability": 0 or 1,
"lm_usefulness": 0 or 1,
"exceptional_value": 0 or 1,
"total_score": 0-5,
"justification": "max 100 words"
}
```

Figure 4: **Annotation prompt.** The system prompt used to label every document in the training mixture. Line breaks are adjusted for display; the wording is otherwise unchanged. The model returns the five binary criteria, their additive total in $\{0, \ldots, 5\}$, and a short justification, as JSON.

Table 4: **Backbone and normalization ablation for the ridge quality scorer.** Values are mean and standard deviation over five-fold cross-validation repeated across five seeds. All heads use ridge regression with $\alpha = 1.0$. The configuration adopted in this work is shown in bold.

Backbone	L_2 **norm**	**Spearman** ρ	**MSE**	**Top-10%**
nomic-embed-text-v1.5	no	0.6717 ± 0.0055	1.0342 ± 0.0128	0.2874 ± 0.0095
nomic-embed-text-v1.5	yes	0.6723 ± 0.0060	1.0353 ± 0.0132	0.2950 ± 0.0099
modernbert-embed-base	no	0.7300 ± 0.0038	0.8548 ± 0.0090	0.3253 ± 0.0101
modernbert-embed-base	**yes**	$\mathbf{0.7314 \pm 0.0037}$	$\mathbf{0.8532 \pm 0.0086}$	$\mathbf{0.3288 \pm 0.0122}$

B Category-Level Topic Breakdown

This appendix expands on the topic-shift analysis of Figure 2. For each scorer we take the top-10% of documents from the $5{,}042{,}921$-document study corpus, roughly 505k documents, label each one with the WebOrganizer topic classifier, and compare the resulting topic distribution against the full

Table 5: **Deviation of each scorer's top-10% selection from the original topic distribution.** JS is the Jensen–Shannon divergence in bits and TV the total variation distance.

Scorer	Selected	JS	TV
FineWeb-Edu	505,674	0.179	0.422
DCLM fastText	504,294	0.049	0.204
ModernBERT	504,293	0.040	0.181

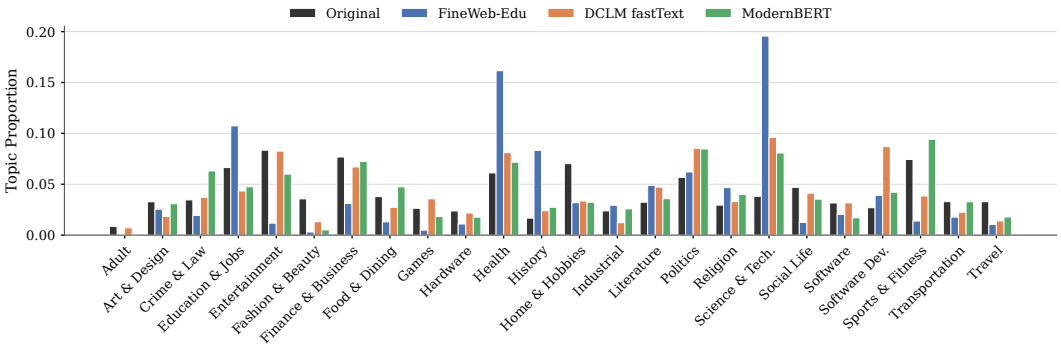


Figure 5: **Topic proportions over the 24 WebOrganizer categories** for the original corpus and for the top-10% selection under each scorer.

corpus. Table 5 reports the Jensen–Shannon divergence and the total variation distance, the latter being the fraction of probability mass reallocated.

Figure 5 shows the per-category distributions, and Figure 6 presents the same information as a log-ratio heatmap, where red marks over-represented categories and blue under-represented ones.

No scorer reproduces the original distribution. FineWeb-Edu deviates furthest, concentrating mass on Education & Jobs, Health, Science & Tech., and History, the academic categories its supervision targets, while suppressing Entertainment, Games, Sports & Fitness, and Fashion & Beauty by up to several orders of magnitude in log-space. DCLM fastText and ModernBERT shift coverage less aggressively but pull in systematic directions of their own, both over-representing Science & Tech. and Software Dev. and under-representing Fashion & Beauty and Travel. Single-scorer quality filtering therefore reshapes topic composition under every scorer we study, not only under FineWeb-Edu.

C Rewriting Model and Prompt Selection

The rewriter and the two Nemotron-family prompts are held fixed across the four non-WRAP rewriting strategies. WRAP-Inspired instead assigns each source document to one of four adapted style prompts and also uses the shared *distill* prompt. This appendix reports the comparison used to fix the rewriter and prompts.

We compare 14 prompt templates released in the FinePhrase repository (Niklaus et al., 2026), grouped into three families. The *rewire* family contains a single guided-rewrite prompt, the *format* family contains eight prompts that transform a document into a target genre such as article, FAQ or tutorial, and the *nemotron* family contains five prompts adapted from Nemotron-CC (Su et al., 2025). Each template is run over the same 100 documents sampled from the Disagreement-Aware selection, under both Qwen2.5-1.5B-Instruct and Qwen2.5-7B-Instruct (Qwen et al., 2025) with greedy decoding and a maximum generation length of 4,096 tokens, giving 2,800 outputs. Every output is judged against its source document by Claude Haiku 4.5. We report both whether any hallucination is detected and whether it is *major*. Major hallucinations contain fabricated factual content, such as an invented entity, a reversed figure, or a misattributed quotation.

Rewriter scale. Across all 14 templates, Qwen2.5-1.5B-Instruct truncated 11.3% of generations at the 4,096-token limit, compared with 0.2% for Qwen2.5-7B-Instruct. A direct repetition scan

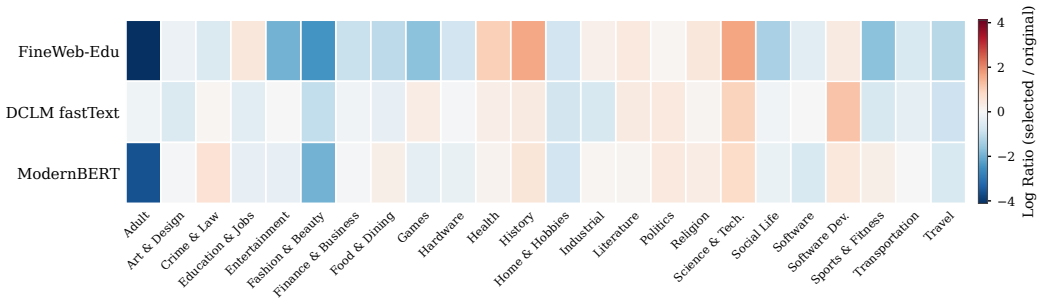


Figure 6: **Per-category log-ratio** $\log(\textbf{selected}/\textbf{original})$ **for each scorer.** No row is flat. Every scorer systematically over- and under-represents specific categories.

Table 6: **Rewriter and prompt selection.** Comp. is the mean ratio of output to source tokens on this prompt-selection sample and is used only to compare rewriters and prompts. Trunc. is the fraction of generations reaching the 4,096-token limit. Halluc. and Major denote the fractions with any and major hallucination, respectively. Family rows average over their constituent prompts.

Scope	Model	Comp.	Trunc.	Halluc.	Major
All prompts	1.5B	0.938	0.11	0.64	0.30
All prompts	7B	0.640	0.00	0.56	0.16
rewire	1.5B	0.922	0.07	0.39	0.20
rewire	7B	0.781	0.01	0.60	0.14
format	1.5B	1.094	0.14	0.71	0.37
format	7B	0.717	0.00	0.65	0.24
nemotron	1.5B	0.691	0.08	0.57	0.21
nemotron	7B	0.488	0.00	0.41	0.05
wikipedia_style	7B	0.637	0.00	0.54	0.05
distill	7B	0.432	0.00	0.43	0.04

shows that 5.8% of 1.5B outputs repeat the same sentence at least 40 times, compared with 0.07% for 7B. We therefore use Qwen2.5-7B-Instruct for corpus rewriting.

At the rewriter scale studied here, the format family shows the highest major-hallucination rates: 37% with 1.5B and 24% with 7B (30.6% pooled across both models). Under the 7B rewriter that we deploy, the nemotron family has the lowest hallucination rates, with 41% any hallucination and 5% major hallucination, compared with $60\%/14\%$ for rewire and $65\%/24\%$ for format. We therefore focus prompt selection within the nemotron family.

We therefore adopt the nemotron prompts *wikipedia_style_rephrasing* and *distill* under Qwen2.5-7B-Instruct. We do not select prompts by hallucination rate alone: several lower-hallucination templates either compress the source aggressively or alter its document format. In contrast, *wikipedia_style_rephrasing* and *distill* keep major hallucination at or below 5%, show no truncation in our sample, and retain complementary amounts of source content, with compression ratios of 0.637 and 0.432, respectively.

Because hallucination remains non-negligible at corpus scale, we add an explicit grounding instruction to the FinePhrase *wikipedia_style_rephrasing* template. In a separate 20-document pilot with Qwen2.5-7B-Instruct, this reduced major hallucination from 15% to 5% while leaving overall hallucination unchanged at 45%. The *distill* template already instructs the model to include only information present in the source and is used as released. The final production prompts are reproduced in Figure 7.

For WRAP-INSPIRED, each source document is assigned to one of four style prompts adapted from Maini et al. (2024). All selected sources are also rewritten with *distill*; although both passes are generated over the full source set, only enough *distill* output is retained to fill the remaining rewritten-token budget. We do not reuse the original WRAP prompt wording: the four production

Wikipedia-style rephrasing.

```
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:".

Important: Do not add any information, claims, or details that are not explicitly stated in or directly inferable from the original document. Your task is only to rephrase and restructure the existing content into clean, encyclopedic prose. If the source document is vague or incomplete on a topic, leave it vague or incomplete -- do not fill gaps with outside knowledge. Text: [TEXT]
```

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.
```

Figure 7: **Production prompts used for the non-WRAP rewriting strategies.** The Wikipedia-style template adds an explicit grounding instruction to the FinePhrase version, whereas *distill* is used as released.

Easy. Content a young child could follow.

```
Rewrite the following passage in simple, plain language that a young child could understand. Keep the key information; do not add new facts. Output ONLY the rewritten passage, with no preamble or commentary.

Passage:
```

Hard. Formal register aimed at a specialist reader.

```
Rewrite the following passage using precise, formal, scholarly language. Preserve all information and add no new facts. Output ONLY the rewritten passage, with no preamble or commentary.

Passage:
```

Wiki. Neutral encyclopedic prose.

```
Rewrite the following passage in the clear, neutral, factual style of a high-quality encyclopedia article. Preserve all information and add no new facts. Output ONLY the rewritten passage, with no preamble or commentary.

Passage:
```

QA. Narrative converted to question-answer form.

```
Convert the following passage into a series of clear question-and-answer pairs that cover its key information. Add no new facts. Format each as 'Q: ...' and 'A: ...'. Output ONLY the question-answer pairs, with no preamble or commentary.

Passage:
```

Figure 8: **Style prompts used by WRAP-INSPIRED.** Each prompt is followed by the source document.

prompts are reimplemented as direct rewriting instructions, each with an explicit constraint against introducing new facts. The prompts are given in Figure 8.

Table 7: **Retained rewritten tokens by prompt pass.** Token counts are in billions. For REWIRE-INSPIRED, counts are after post-rewrite fastText filtering.

Strategy	Pass 1	Distill	Distill share
QUALITY-FIRST	3.26	1.74	34.9%
WRAP-INSPIRED	4.16	0.84	16.9%
REWIRE-INSPIRED	2.35	2.65	53.1%
DIVERSITY-ORIENTED	3.70	1.19	24.4%
DISAGREEMENT-AWARE	3.54	1.46	29.1%

Table 8: **Quality floor against variance cap over $\mathcal{U}$.** The bold cell is the entire independent contribution of the variance cap. Every other document it rejects already falls below the quality floor.

	$v \leq \tau_v$	$v > \tau_v$	**Total**
$q \geq \tau_q$	10,435,667	**51,783**	10,487,450
$q < \tau_q$	3,048,194	1,446,424	4,494,618
Total	13,483,861	1,498,207	14,982,068

Production prompt composition. Both rewriting passes are generated over the full selected source set. The first pass uses the grounded Wikipedia-style prompt for the four non-WRAP strategies and one of the four style prompts for WRAP-INSPIRED; the second pass uses *distill*. For the four strategies that select sources before rewriting, the final corpus is constructed from first-pass outputs and supplemented with second-pass outputs to approach the target rewritten-token budget. REWIRE-INSPIRED instead applies its post-rewrite fastText filtering before the final budget is filled. Production token yields can vary with the source documents selected by each strategy. Table 7 reports the retained token composition.

D DOCUMENT OVERLAP BETWEEN STRATEGIES

This appendix reports pairwise document overlap between the selection strategies of Section 4.2, and characterizes the feasible domain that DISAGREEMENT-AWARE ranks over. To show how the weight λ moves that ranking, four further values are included alongside the $\lambda = 0.5$ used in the main text. No model is trained on the four.

D.1 FEASIBLE DOMAIN

The candidate domain $\mathcal{U}$ is the union of the highest-ranked documents under each scorer, with each selection capped at a 10B-token budget. The resulting union contains 15.0M documents and 21.96B tokens. A document need satisfy only one scorer to enter $\mathcal{U}$, and fewer than 2% satisfy all three. Applying the quality floor and variance cap, set at the 30th percentile of q and the 90th percentile of v over $\mathcal{U}$, leaves a feasible domain $\mathcal{U}_\tau$ of 10.4M documents and 16.70B tokens, or 70% of $\mathcal{U}$ by document count.

The two constraints are far from independent. Applied separately they would remove 30% and 10% of $\mathcal{U}$, but their combined removal is 30.4%, since the documents the cap targets are almost all already below the floor. Only $51{,}783$ documents, 0.35% of $\mathcal{U}$, are removed by the cap alone (Table 8). The cap is therefore a narrow safeguard against pathological disagreement rather than a second filtering stage, and the size of the feasible domain is set by the quality floor.

D.2 THE ROLE OF λ

Within $\mathcal{U}_\tau$ documents are ranked by $u_\lambda(d) = q(d) + \lambda\sqrt{v(d)}$ and admitted until the budget is filled. The parameter enters only through this ranking. Both the candidate domain and the two thresholds are fixed before λ is applied, so all five settings rank the identical $10{,}435{,}667$-document domain and differ only in the order.

Table 9: **The five values of λ.** All rank the same feasible domain and fill the same 10B-token budget. The document count rises with λ because higher values favor shorter documents. Only $\lambda = 0.5$, in bold, is used for pretraining.

λ	Documents	% of $\mathcal{U}_\tau$	u_λ at cutoff
0	5,542,662	53.1	0.815
0.25	5,558,959	53.3	0.849
0.5	**5,602,476**	**53.7**	**0.883**
0.75	5,702,556	54.6	0.916
1	5,876,483	56.3	0.944

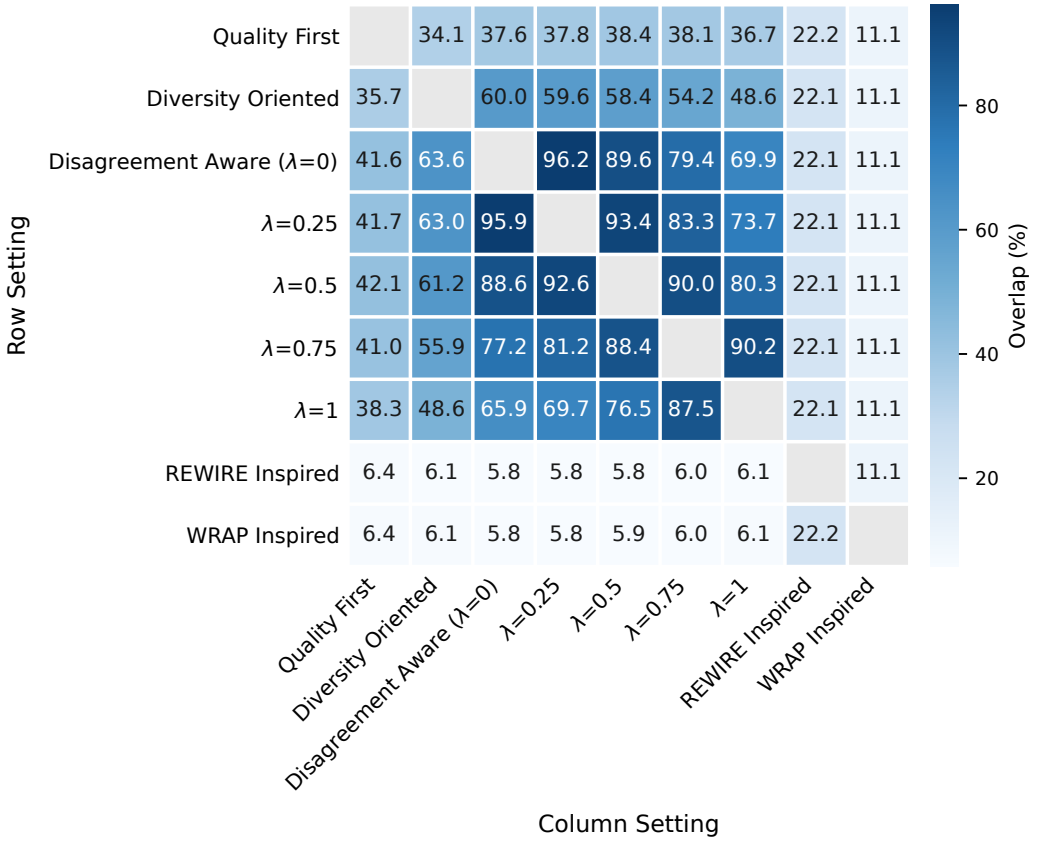


Figure 9: **Pairwise document overlap across nine selection settings.** Each cell is the percentage of the row setting's documents that also appear in the column setting. The five DISAGREEMENT-AWARE rows form a graded band around the diagonal, so λ moves the selection continuously rather than switching between regimes.

Across $\mathcal{U}_\tau$, 36.8% of documents are selected at every λ and 27.7% at none, so λ contests the remaining 35.5%, roughly 3.7M documents.

D.3 OVERLAP BETWEEN SETTINGS

Figure 9 reports $|A \cap B|/|A|$ for every ordered pair of settings, the fraction of the row setting's documents that also appear in the column setting. The measure is asymmetric because the settings select different numbers of documents.

The two single-objective strategies largely disagree. Only 34.1% of QUALITY-FIRST's documents appear in DIVERSITY-ORIENTED, and 35.7% in the other direction. Two reasonable strate-

Table 10: **Three documents from $\mathcal{U}_\tau$ across the λ grid.** Bold marks the settings that admit the document. At each λ, documents are admitted in descending order of u_λ until the token budget is filled; the cutoff in Table 9 is the utility of the last document admitted. Values are rounded to three decimals, which can hide a narrow miss: B is not admitted at $\lambda = 0.75$, nor C at $\lambda = 0.25$; they fall 1,596 and 2,523 places below their cutoffs.

	A	**B**	**C**
DCLM fastText r_1	0.903	0.826	0.957
FineWeb-Edu r_2	0.799	0.972	0.588
ModernBERT r_3	0.807	0.289	0.880
Consensus q	0.836	0.696	0.809
Disagreement v	0.0022	0.0865	0.0253
$\sqrt{v}$	0.047	0.294	0.159
u at $\lambda = 0$	**0.836**	0.696	0.809
u at $\lambda = 0.25$	0.848	0.769	0.848
u at $\lambda = 0.5$	0.860	0.843	**0.888**
u at $\lambda = 0.75$	0.872	0.916	**0.928**
u at $\lambda = 1$	0.884	**0.990**	**0.968**

gies applied to the same pool under the same budget therefore agree on barely a third of what to keep, which is what makes the choice of strategy consequential.

λ interpolates smoothly. Adjacent settings on the grid overlap by 96.2%, 93.4%, 90.0% and 90.2%, and overlap decays monotonically with distance, down to 69.9% between the endpoints. Intermediate values give intermediate selections.

The selection sits between the two strategies without being contained in them. At $\lambda = 0.5$, 42.1% of the selection is shared with QUALITY-FIRST and 61.2% with DIVERSITY-ORIENTED, more with each than the two share with each other, which is 34.1%. By that measure the criterion is intermediate between them. It is nevertheless not contained in their union. 26.7% of its documents, 1,495,841 in total, appear in neither, and that exclusive fraction grows from 25.7% at $\lambda = 0$ to 36.2% at $\lambda = 1$. Ranking by disagreement therefore lands between the two single-objective strategies while also reaching documents that neither selects.

The rewriting baselines are independent of quality-based selection. Every setting overlaps REWIRE-INSPIRED at 22.1% and WRAP-INSPIRED at 11.1%, measured on their pre-rewrite source pools. These match those pools' share of the candidate corpus, 22.15% and 11.07%, which is the overlap expected under independence. Both baselines select without conditioning on the quality signals, so the pale final columns of Figure 9 reflect a base rate rather than a substantive similarity.

D.4 EXAMPLES

Table 10 follows three documents through the ranking, and Figures 10–12 reproduce them.

Document A, admitted by consensus. All three scorers place it between the 80th and 90th percentile. It is clean first-person prose with no boilerplate, so there is little for the scorers to disagree about, and $v = 0.0022$ sits near the bottom of the feasible domain. The bonus is worth 0.023 at $\lambda = 0.5$, not enough to keep pace with the rising cutoff, so the document is admitted at $\lambda = 0$ and displaced by $\lambda = 0.25$.

Document B, admitted by disagreement. The page is a student-made photo book retelling the Perseus myth in fifteen numbered panels, followed by the hosting site's own page material, including a size picker, uploader metadata, and a promotional banner repeated twice. FineWeb-Edu reads the narrative and places the document in the top 3%. ModernBERT reads the tail and places it in the bottom 30%, which follows from its annotation rubric, where navigation text and duplicated patterns are penalized explicitly (Figure 4). The two scorers respond to different halves of the same

document. Among the three case-study documents, it has the lowest consensus quality, and only the full bonus at $\lambda = 1$ admits it.

Document C, at the trained setting. Document C is a Sitecore developer question with two answers, and it sits between the other two, with v an order of magnitude above A and under a third of B's. DCLM fastText and ModernBERT both read clean technical problem-solving, a well-posed question followed by working code and numbered implementation steps, and place it at the 96th and 88th percentiles. FineWeb-Edu dissents at the 59th, since it rewards academic educational value and vendor-specific developer documentation is not schoolwork. The dissent runs opposite to Document B, where a scorer was pulled down by boilerplate the others tolerated. The document falls just outside the budget at $\lambda = 0$ and enters at $\lambda = 0.5$ with roughly 400,000 ranks to spare.

The bonus decides a minority of the selection. A document excluded at $\lambda = 0$ has q below that setting's cutoff utility of 0.815, and admission at $\lambda = 0.5$ requires $q + 0.5\sqrt{v}$ to reach its cutoff of 0.883 (Table 9), so the bonus can only decide the outcome when $v > 0.019$. That threshold sits at the 78th percentile of v within the $\lambda = 0.5$ selection, against a median of 0.007. A worked example of the bonus is therefore atypical by construction. Document C's $v = 0.025$ is about $3.5\times$ the median, which is close to the least atypical such example available, and still only $0.29\times$ Document B's. Consistent with this, 88.6% of the documents selected at $\lambda = 0.5$ are also selected at $\lambda = 0$, where the bonus is switched off. The criterion reorders the selection at its margin rather than replacing it.

Memories from Vailsburg

At the time of the Newark riots I was 10 years old and living in the Vailsburg section, two blocks from Vailsburg Park. I still to this day remember the 10 p.m. curfew and the National Guard stationed at the corner of the block along with tanks positioned in the park. I was born and raised in Newark and spent my first 20 years in the city.

Living in the West Ward of Newark was interesting in the sense of diversity due to the "given" boundary lines. . . . . east of the Garden State Parkway was where the black population was continuing to grow and west of the Parkway was where the white population was continuing to diminish. I have great memories of playing baseball at Vailsburg Park and growing up with some great guys from the neighborhood. My family also worked as employees for the city of grandmother worked in City Hall and my mom worked in the Housing Authority.

Several years ago after residing in other parts of the country, I returned to New Jersey and now reside in Somerset County with my family. Today's article brought back a lot of memories both good and bad. I took my sons to the old neighborhood to let them see where I had grown up. My home was an empty had burned down to the ground and I learned that after we sold the home in the mid 70's the neighborhood became a high crime area.

After 40 years and a lot of lessons learned, I sense some opportunities for revitalizing the city;however, I remember an analogy of another city that experienced the same type of difficulties that Newark had and that city took a different course. The city I reference is Baltimore. Baltimore had pretty much the same type of issues Newark had some forty years ago and somehow found a way to rebound and by no means are they perfect but I believe they are ahead of Newark in many ways.

I would not want to be Cory Booker because he is faced with a huge mountain to climb and maybe he should take a hard look at other major cities that climbed the mountain. He needs to be careful in his approach and I wish him all the best. I would love to bring my kids back and show them a revitalized city.

- Ed Fagan, Branchburg

Figure 10: **Document A**, a memoir of the 1967 Newark riots, admitted at $\lambda = 0$ only. Text is verbatim, including two spans dropped by the corpus extractor.

Blank Canvas
Blank Canvas - Page Text Content FC: The Story of Perseus
1: There once was a king named Acrisisus of Argo. He always dreamed of having a son. Instead he had a beautiful daughter, Danae. 2: As Danae grew older. Her father was told by the God, Apollo, that if she was to ever have a son, the child would kill him. 3: In order for Acrisius to keep Danae from having a child he had to lock her in a cellar to keep her away from any guys. 4: One night, while Danae was sound asleep, the god of all gods, Zues, created a golden rain that sent down a little baby boy named Perseus. | Danae was afraid to tell her father, so she tried her hardest to keep the baby a secret. 5: When Acrisius came down to check on his daughter he discovered the boy. In fear for his own life, Acrisius banished the two sending them away in a boat to nowhere. 6: Days after aimless floating, Danae and Perseus finally landed on an island. Here they met Dictys and his very loving wife. 7: Years later Dictys's brother Polydectes became known of Danae and fell in love with her. By now Perseus was full grown. | Polydectes was not fond of Perseus so he needed a plan on how to get rid of him. 8: In order to get rid of Perseus, Polydectes sent him on an impossible mission to receive the head of one of the 3 gorgons, Medusa. 9: This was impossible for whenever anyone looks at a gorgon they instantly turn into stone. 10: As brave as Perseus was, he accepted the challenge. Perseus didn't tell his mother what he was off to do because he knew she wouldn't approve. Perseus packed his things and went on his way. 11: After days and days of traveling and fighting Perseus finally returned with the head of Medusa. | He revenged Polydectes and turned him into stone with Medusa's head. 12: Time later Perseus and Danae decided that they would go back to Argo to see if Acrisius had changed. They soon found out that the King was driven away from the city. 13: When in Greece, Perseus found out about a athletic competition. So he competed in a discus throwing contest. 14: Ironically Acrisius was at the same event. He came as a spectator. When Perseus threw his discus it curved into the stands and hit Acrisius, killing him. 15: No matter what Acrisius did, Apollos oracle came true. | THE END

Sizes: mini|medium|large|gargantuan
Default User • By: kiara k. • Joined: over 4 years ago • Published Mixbooks: 1
No contributors

About This Mixbook
• Title: Blank Canvas • Theme for Mixbook Scrapbookers • Tags: None • Published: over 4 years ago

Get up to 50% off Your first order
Get up to 50% off Your first order

Figure 11: **Document B**, a retelling of the Perseus myth followed by the hosting site's page material, admitted at $\lambda = 1$ only. Text is verbatim, including the original spelling.

I am trying to figure out if i can have multiple columns and multiple rows for one particular field under an item in Content Editor.

Field Name:
[column 1][column 2][column 3][column 4][column 5]

This serves as a drawing. So in Content Editor, where you can fill out Details of the item, I would like to have multiple columns and multiple rows. So the data will be served under one particular "Type". Hopefully this makes sense

• Can the rows be predefined (for example each item will have 5 rows) or do they need to be different for each item?
– Koen Heye
Jan 26, 2017 at 20:29

2 Answers 2

I think you are talking about custom field type.

Looks like you are talking about something that cannot be done with standard field types thus you will have to create your own custom field type for that

How to create custom field type

1. Implement your custom field

```
namespace Sitecore.Playground.Fields.FieldTypes
   public class CustomField : Edit
      protected override void DoRender(HtmlTextWriter output)
         Attributes["placeholder"] = Translate.Text(Placeholder);
         string str = " type=\"hidden\"";
         output.Write("<input" + ControlAttributes + str + ">");
         // render your custom control here
```

Notice output.Write usage. Now you have to build your own HTML with rows, columns and then store a raw value of the field in <input>.

Perhaps you will have to override Value property from the base class to properly serialise/deserialize stored value.

As you probably know each field value can be expressed as Raw Value. Idea is that your field raw value could be for example: ...

then in C# code, you need to properly render those values as rows/columns. This is an example, I don't know what you want to store there.

1. Register your custom field with config

```
<source mode="on" namespace="Sitecore.Playground.Fields.FieldTypes" assembly="Sitecore.Playground" prefix="myCustomPrefix" />
```

1. Create field type item in core database

   /sitecore/system/Field types/Developer Types/My Custom field

Fill Control field with this value: myCustomPrefix:CustomField

This blog post might be useful for you: ...

I can't give you a complete solution but I can guide you a bit to along the right path.

You'll need a custom field type for this. Credit here goes to jammykam ...

Who wrote a nice blog article about creating a key/value data field with lookup: `https://jammykam.wordpress.com/2015/03/08/custom-sitecore-field-for-storing-keyvalue-data-with-lookups/`

I've already implemented it in a solution and it works perfectly.
You'll need to define your column headers somewhere. Cfr. Chemical Types in the example.

And then, depending on how flexible you want your rows to be, you either:

1. Create a template that has a fixed number of your custom field type.
2. Or create a template that has only 1 field of your custom field type. You then use this template to create children under the item. So each child = 1 row with data.

Here's a link to the gist of the code: `https://gist.github.com/jammykam/d935b2c5b6054ca10282`
You'll probably need to make some tweaks to it. Like if you don't want to show the 'Chemical unit'.

Hope this helps! If you have questions, just ask!

Your Answer

Figure 12: **Document C**, a Sitecore developer question with two answers, admitted from $\lambda = 0.5$ onward. Text is verbatim. Ellipses mark an example raw value and two links that the corpus extractor dropped. The C# block appears without braces, three consecutive steps are numbered 1, and the sentence beginning "Who wrote" has lost its subject to a dropped link. These are artifacts of the same extraction and appear in the corpus as printed.

Table 11: **Training configuration**, identical across all 18 runs.

Architecture	
Layers	28
Hidden size	2048
FFN size	5632, SwiGLU
Attention heads	16, multi-head
Position encoding	RoPE, $\theta = 10^4$
Normalization	RMSNorm, $\epsilon = 10^{-5}$
Vocabulary	32000, Llama 2 tokenizer
Tied embeddings	yes
Parameters	1.504B, 1.439B non-embedding
Optimization	
Optimizer	AdamW, $\beta = (0.9, 0.95)$
Weight decay	0.1
Gradient clipping	1.0
Peak learning rate	5×10^{-4}
Warmup	500 steps, linear
Schedule	constant-rate trunk, one decay branch per epoch
Decay	final 10% of each branch, linear to 0
Batching	
Sequence length	2048
Global batch	1024 sequences
Tokens per step	2.10M
Parameter precision	bfloat16
Gradients and optimizer state	fp32

E Training Hyperparameters

Each of the six settings is trained with seeds 42, 43 and 44, and all 18 runs share the configuration in Table 11.

Each run uses a shared constant-learning-rate trunk with a 500-step linear warmup. The three reported epoch checkpoints branch from this trunk and linearly decay to zero over the final 10% of their respective horizons, ending at steps 4,768, 9,537, and 14,305.

Weights are initialized once per seed and loaded by all six settings, so the six arms at a given seed start from identical parameters. Row-parallel projections are scaled by $1/\sqrt{2L}$ with $L = 28$ layers, and all other weights are drawn from $\mathcal{N}(0, 0.02)$. The model uses no dropout and the data order is fixed, so the seed controls initialization alone.

Documents are concatenated and chunked into fixed 2048-token windows separated by an end-of-sequence token, so sequences may cross document boundaries. Each corpus is shuffled once at the document level before tokenization and is then read in a fixed order, which repeats exactly in the second and third epochs.

All six settings train for 14,305 steps, corresponding to approximately 30.0B training tokens. The training horizon is therefore identical across all settings.

Each run uses four H100 GPUs on a single node with data-parallel degree four. Based on the measured throughput of this configuration, the trunk-and-branch schedule is estimated to require approximately 100 hours and 400 H100-hours per run, or about 7,200 H100-hours for the full study.

F Per-Task Results

All numbers are `acc_norm` from LightEval, six settings by three seeds (42, 43, 44) by three epoch boundaries. We report the seed-level standard deviation and the number of seeds in which a setting exceeds QUALITY-BASE, rather than significance tests. With three seeds a paired t-test rests on a two-degree-of-freedom variance estimate and a sign test floors at $p = 0.125$, so a p-value would overstate what the design supports. Differences are in accuracy points throughout.

Table 12 supports the erosion result of Section 6.1. Every rewriting setting exceeds the baseline on the seed mean after one epoch, by 0.22 to 0.76 points, although only 10 of the 15 strategy–seed pairs are individually positive. By the third epoch, DIVERSITY-ORIENTED is the only strategy with

Table 12: **Aggregated commonsense accuracy at each epoch boundary**, the mean over the six tasks of Section 5.4. The second column is absolute accuracy in points, the third the difference from QUALITY-BASE at the same epoch, both with the standard deviation over seeds. The last column counts the seeds in which the setting exceeds QUALITY-BASE.

Epoch	Accuracy	Δ	Seeds > 0
QUALITY-BASE			
1	41.75 ± 0.67	—	—
2	44.80 ± 0.07	—	—
3	46.27 ± 0.67	—	—
QUALITY-FIRST			
1	42.41 ± 0.28	$+0.67 \pm 0.80$	$2/3$
2	44.96 ± 0.41	$+0.16 \pm 0.47$	$1/3$
3	46.06 ± 0.27	-0.21 ± 0.89	$1/3$
WRAP-INSPIRED			
1	41.97 ± 0.39	$+0.22 \pm 1.05$	$1/3$
2	44.62 ± 0.44	-0.18 ± 0.39	$1/3$
3	45.81 ± 0.24	-0.46 ± 0.91	$1/3$
REWIRE-INSPIRED			
1	42.28 ± 0.14	$+0.54 \pm 0.54$	$\mathbf{3/3}$
2	44.94 ± 0.31	$+0.14 \pm 0.27$	$2/3$
3	45.89 ± 0.23	-0.38 ± 0.70	$1/3$
DIVERSITY-ORIENTED			
1	42.50 ± 0.43	$+0.76 \pm 0.82$	$\mathbf{3/3}$
2	45.40 ± 0.29	$+0.60 \pm 0.31$	$\mathbf{3/3}$
3	46.41 ± 0.21	$+0.14 \pm 0.86$	$2/3$
DISAGREEMENT-AWARE			
1	42.11 ± 0.29	$+0.36 \pm 0.93$	$1/3$
2	44.82 ± 0.34	$+0.02 \pm 0.40$	$1/3$
3	45.93 ± 0.11	-0.34 ± 0.77	$1/3$

Table 13: **Per-task accuracy after the third epoch.** The QUALITY-BASE row is absolute accuracy in points, every other row the difference from it, with the standard deviation over seeds taken on the per-seed differences. Bold marks the largest positive difference in a column; ARC-Easy and SIQA have no positive difference at this epoch, so nothing is bolded in those two columns.

Setting	ARC-e	HellaSwag	PIQA	SIQA	OBQA	CSQA
QUALITY-BASE	52.20 ± 1.67	48.37 ± 0.54	70.35 ± 0.42	40.52 ± 0.38	34.13 ± 1.53	32.05 ± 0.90
QUALITY-FIRST	-0.87 ± 2.42	$+0.18 \pm 0.48$	$+0.98 \pm 0.39$	-0.65 ± 0.63	-0.80 ± 1.51	-0.11 ± 1.53
WRAP-INSPIRED	-1.21 ± 2.38	-0.35 ± 0.21	$+1.09 \pm 0.43$	-0.65 ± 0.39	-1.60 ± 2.27	-0.03 ± 0.33
REWIRE-INSPIRED	-1.21 ± 2.42	$+0.49 \pm 0.27$	$+0.80 \pm 0.53$	-0.49 ± 0.65	-2.13 ± 1.14	$\mathbf{+0.25 \pm 0.46}$
DIVERSITY-ORIENTED	-0.17 ± 2.21	$\mathbf{+1.34 \pm 0.56}$	$\mathbf{+1.54 \pm 0.55}$	-0.70 ± 1.28	$\mathbf{+0.07 \pm 2.52}$	-1.23 ± 1.42
DISAGREEMENT-AWARE	-0.29 ± 1.59	-0.88 ± 0.42	$+0.69 \pm 0.17$	-0.51 ± 0.18	-0.53 ± 2.20	-0.52 ± 0.98

a positive mean margin ($+0.14$), and it is positive in two of three seeds; the other four strategies are 0.21–0.46 points below the baseline and positive in only one seed each. Recomputing the aggregate while leaving out each task in turn keeps the DIVERSITY-ORIENTED margin positive in 19 of 21 variants overall; at the third epoch it ranges from -0.14 to $+0.42$ points and turns negative when either HellaSwag or PIQA is removed.

Table 14 shows that DISAGREEMENT-AWARE leads the MMLU macro-average from the second epoch onward, by $+0.60$ and $+0.49$ points, and is positive in all three seeds at both checkpoints. Table 15 further shows that, at the final epoch, no setting improves all four MMLU categories.

Table 14: **MMLU accuracy at each epoch boundary.** MMLU is the macro-average over 57 subjects. The second column is absolute accuracy in points, the third the difference from QUALITY-BASE at the same epoch, both with the standard deviation over seeds. The last column counts the seeds in which the setting exceeds QUALITY-BASE.

Epoch	MMLU	Δ	Seeds > 0
QUALITY-BASE			
1	28.16 ± 0.27	—	—
2	29.53 ± 0.09	—	—
3	30.59 ± 0.03	—	—
QUALITY-FIRST			
1	28.51 ± 0.27	$+0.35 \pm 0.51$	$2/3$
2	30.01 ± 0.35	$+0.48 \pm 0.38$	$\mathbf{3/3}$
3	30.64 ± 0.34	$+0.05 \pm 0.37$	$2/3$
WRAP-INSPIRED			
1	27.70 ± 0.33	-0.46 ± 0.59	$1/3$
2	28.95 ± 0.19	-0.59 ± 0.25	$0/3$
3	29.72 ± 0.39	-0.87 ± 0.42	$0/3$
REWIRE-INSPIRED			
1	28.19 ± 0.24	$+0.03 \pm 0.10$	$2/3$
2	29.55 ± 0.38	$+0.02 \pm 0.30$	$1/3$
3	30.59 ± 0.50	-0.01 ± 0.47	$1/3$
DIVERSITY-ORIENTED			
1	28.19 ± 0.26	$+0.03 \pm 0.31$	$2/3$
2	29.93 ± 0.12	$+0.40 \pm 0.16$	$\mathbf{3/3}$
3	30.82 ± 0.13	$+0.23 \pm 0.15$	$\mathbf{3/3}$
DISAGREEMENT-AWARE			
1	28.32 ± 0.07	$+0.16 \pm 0.34$	$1/3$
2	30.13 ± 0.20	$+0.60 \pm 0.28$	$\mathbf{3/3}$
3	31.08 ± 0.40	$+0.49 \pm 0.42$	$\mathbf{3/3}$

Table 15: **MMLU by category at the second and third epochs.** The QUALITY-BASE row reports absolute accuracy in percentage points. The remaining rows report differences from it, averaged over three seeds, with the best setting in each column in bold. MMLU macro-averages with seed-level variability are reported in Table 14.

	Epoch 2				Epoch 3			
Setting	**STEM**	**Hum.**	**Soc.**	**Other**	**STEM**	**Hum.**	**Soc.**	**Other**
QUALITY-BASE	26.9	29.2	30.7	32.3	27.5	30.6	31.9	33.4
QUALITY-FIRST	$\mathbf{-0.22}$	$+0.56$	$\mathbf{+1.43}$	$+0.48$	-0.96	-0.05	$+0.33$	$\mathbf{+1.19}$
WRAP-INSPIRED	-1.11	-0.24	-0.53	-0.29	-1.45	-0.97	$+0.18$	-0.94
REWIRE-INSPIRED	-0.29	-0.00	-0.01	$+0.45$	-0.39	-0.32	$+0.31$	$+0.51$
DIVERSITY-ORIENTED	-0.52	$\mathbf{+1.17}$	$+1.10$	$+0.26$	-0.65	$+0.57$	$\mathbf{+1.13}$	$+0.27$
DISAGREEMENT-AWARE	-0.40	$\mathbf{+1.17}$	$+0.79$	$\mathbf{+1.18}$	$\mathbf{-0.21}$	$\mathbf{+0.99}$	$+0.54$	$+0.87$

Every rewriting strategy loses on STEM, while DISAGREEMENT-AWARE has the smallest loss there (-0.21) and gains on Humanities, Social Sciences, and Other. Its MMLU advantage coexists with a negative commonsense mean (-0.34, one seed of three): it falls below the baseline on every commonsense task except PIQA, while PIQA improves by $+0.69$ points in all three seeds.