BEYOND GROUP PROPORTIONS: QUALITY FILTERING AND DIALECT GAPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Quality filters decide which documents survive into a language model's training data, and their effects fall unevenly across varieties of English. The usual way to check for this is to count how much data each group keeps. We ask whether that count tells the right story for African American English (AAE), the variety spoken by many Black Americans, by comparing models trained on datasets that contain exactly the same number of AAE-aligned and white-aligned documents but different documents within each group. In a controlled continued-pretraining experiment on tweets, replacing the documents chosen by the DataComp-LM quality classifier with different documents at the same group counts removes 63% of the widening in the likelihood gap between the groups relative to uniform selection, and it improves fit to both groups relative to the filter's own documents. Restoring the original group share also narrows the gap, so the count matters, but it is not the whole account. In web pretraining from scratch at a matched token budget, an additional cut with the same classifier on an already-curated corpus raises the between-group perplexity ratio by 2.9%, whereas a second pool and budget give an estimate whose interval includes zero. Reweighting documents kept by a perplexity filter does not reproduce the benefit of changing them. These results argue for auditing which examples of a variety remain in the training data, and not only how many.

1 INTRODUCTION

Quality filtering is a standard part of preparing web data for language-model pretraining (Li et al., 2024; Penedo et al., 2024), but its effects are uneven across language varieties. Corpus audits document unequal removal and quality judgments (Dodge et al., 2021; Gururangan et al., 2022; Deas et al., 2025), while studies of trained models show consequences of detoxification and pretraining-data choices (Welbl et al., 2021; Xu et al., 2021; Longpre et al., 2024). These findings raise a question that retained group proportions alone cannot answer. When a filter changes how well a model fits a language variety, how much can we learn from the group's share of the remaining data, and what might we miss about the examples selected within that share?

A filter can change both the amount of data retained from a group and the kinds of documents that represent it. Two training sets might therefore contain exactly the same number of documents from a group while providing different examples from which a model can learn. We investigate this distinction for African American English (AAE), using the DataComp-LM (DCLM) quality classifier released for language-model pretraining. Our central finding is that **changing the retained documents can improve model fit even when group counts remain fixed**.

In controlled continued pretraining of GPT-2 124M on tweets, the filtered dataset and a fixed-share dataset, which keeps the filter's group counts but selects different documents within each group, each contain exactly $12{,}836$ AAE-aligned documents out of $29{,}491$, yet changing the documents reduces the AAE–white negative log-likelihood (NLL) gap from 0.489 to 0.416 nats. This removes 63% of the widening relative to uniform selection and, relative to the filter's documents, improves fit to both groups with a larger gain for AAE-aligned text (Section 2). In a separate web experiment, we train 150M models from scratch with matched retained-token inventories and 15B training-token presentations. An additional DataComp-LM cut on DCLM-Baseline widens the gap by 0.028 nats relative to uniform selection from the same source regions (region-matched), a 2.9% increase in the

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.

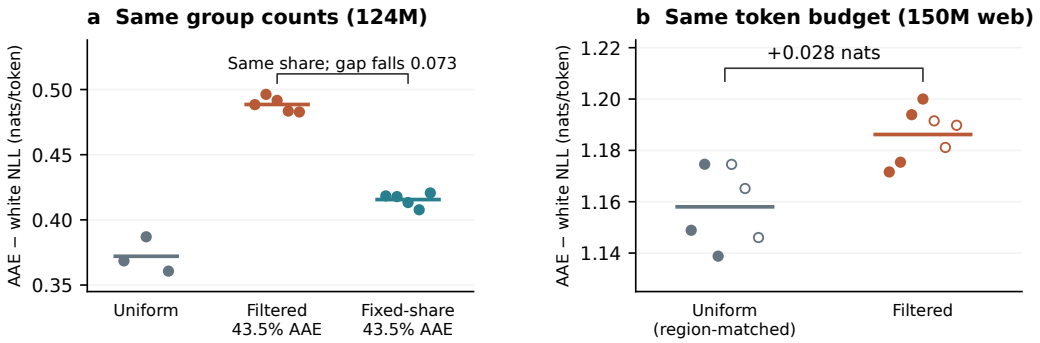

Figure 1: **Group counts and token budgets provide complementary tests of selection.** In the left panel, changing retained documents at the filter's exact group counts reduces the gap by 0.073 nats in the 124M testbed. In the right panel, filtering widens the gap by 0.028 nats in 150M web pretraining at matched token budgets. Lower gaps indicate better relative fit to AAE-aligned text. Dots show training seeds and bars show means, with open dots marking the web extension seeds. Both experiments use the released DCLM classifier but differ in training regime and in the size of the gap itself. Testbed gaps lie near 0.4 nats and web gaps near 1.2 nats, so the two panels use different vertical axes.

perplexity ratio, although a second pool gives an estimate whose interval includes zero (Section 3). Figure 1 summarizes the two comparisons.

The two experiments answer different questions. The controlled comparison shows that group proportions cannot fully explain the selection effect, and the web comparison shows that the same classifier widens the gap at a matched budget in a production pretraining pool.

These results motivate closer attention to the documents that survive within each group. We report per-group losses and retained-document characteristics, and compare changing the selected set with reweighting documents already selected. Random-selection controls reduce the tested gaps, with quality tradeoffs that depend on the outcome measured. The practical contribution is evidence for auditing which examples remain available to a model alongside how much each group contributes to its training data.

2 Same Group Counts, Different Model Fit

If quality filtering reduces a dialect's share of the training data, a wider gap in model fit might seem to follow from that loss of representation. Yet the filter also changes which documents remain within the group. We examine the importance of this second change by replacing the selected documents while holding the filter's group counts exactly fixed. In the controlled experiment below, this removes approximately 63% of the widening relative to uniform selection.

The experiment uses a pool of 90,000 TwitterAAE documents, balanced between AAE-aligned and white-aligned text (Blodgett et al., 2016). Each policy retains 29,491 documents, approximately 32.8% of the pool, which we use to continue training GPT-2 124M (Radford et al., 2019) from the same released checkpoint for 4,000 updates of 32 documents each, sampled with replacement, so each document is presented about 4.3 times on average; sequences are padded and truncated at 64 tokens, and the loss is a token average over unpadded positions. We evaluate on 5,000 held-out natural tweets from each group, measuring the difference in mean per-text negative log-likelihood (NLL), AAE-aligned minus white-aligned, in nats per token. A larger gap indicates worse relative fit to AAE-aligned text. The setting allows us to study selection within continued pretraining, while Section 3 examines web models trained from scratch.

The released DataComp-LM classifier (Li et al., 2024) retains the highest-scoring documents across the pool, reducing the AAE-aligned share from 50% to 43.5%. The resulting gap is 0.489, compared with 0.372 under uniform selection at the same retained-document budget (Table 1). Our fixed-share

Policy	AAE share	Seeds	AAE NLL	White NLL	NLL gap
Uniform	50.3%	3	–	–	0.372 ± 0.014
Filtered	**43.5%**	5	5.093 ± 0.014	4.605 ± 0.010	0.489 ± 0.006
Fixed-share	**43.5%**	5	4.959 ± 0.006	4.544 ± 0.007	0.416 ± 0.005
Restored-share	50.0%	5	5.035 ± 0.008	4.642 ± 0.012	0.393 ± 0.008
CLUSTERRANDOM	49.6%	5	4.949 ± 0.009	4.577 ± 0.011	0.372 ± 0.007

Table 1: **Changing documents at fixed group counts improves fit to both groups.** Each arm retains 29,491 documents and uses the same GPT-2 124M continued-pretraining recipe. Losses are nats per token, means $\pm$ training-seed standard deviation (SD); the gap is AAE-aligned minus white-aligned NLL. Bold shares identify the exact-count comparison. Percentages of the widening are computed from unrounded means. Fixed-share and restored-share use alignment labels. The pool is balanced at 50%; the uniform row's 50.3% is the realized share of its draw. Dashes indicate that the original uniform evaluation saved only the gap, not its component losses.

alternative starts from the same full pool and chooses different documents subject to the filter's final group counts, including 12,836 AAE-aligned documents.

We first group the pool into 500 clusters of DistilBERT embeddings (Sanh et al., 2019). Selection then proceeds through five rounds, with each round's retained set becoming the next round's input. Within a round, we sample randomly in proportion to the remaining cluster sizes, then use dialect labels to swap selected and unselected documents within clusters to match the filter's group counts at that round's cumulative retention. After the fifth round, both datasets contain 43.5% AAE-aligned text but differ in which documents remain. Appendix A.1 gives the per-round retention rule.

Replacing the filtered documents lowers the gap from 0.489 to 0.416, a reduction of 0.073 nats with a 95% Welch confidence interval of [0.065, 0.081] (per-seed values in Table 9). The reduction is positive at each of the five training seeds. Expressed as a share of the filtered-minus-uniform widening, it removes approximately 63% (the uniform reference has three seeds; Table 1). Relative to filtering, it also improves fit to both groups. Fixed-share lowers AAE-aligned NLL by 0.134 nats and white-aligned NLL by 0.061 nats (Table 1), with the larger improvement on AAE-aligned text accounting for the smaller gap. Retained group proportions therefore miss a substantial part of the selection effect in this testbed.

Restoring the original group share offers another way to reduce the gap. The restored-share intervention retains the highest-scoring documents within each group while using dialect labels to restore the 50% AAE-aligned share, yielding a gap of 0.393. That reduction, 0.095 nats with Welch interval [0.085, 0.105], is positive at every seed and removes about 82% of the widening, so both levers are large in this testbed and the paper's emphasis on the second rests on what it reveals about the selected documents, not on its size. Random sampling within proportional cluster allocations, without a dialect quota (CLUSTERRANDOM, Section 4), yields 0.372, equal to the uniform-selection mean within rounding. Because restoring the share also changes which documents remain, the two reductions are alternative interventions and cannot be added to partition the filtering effect. Moreover, restored-share combines improved AAE-aligned fit with worse white-aligned fit relative to filtering, whereas fixed-share improves both.

Characteristics of the retained documents. Saved construction diagnostics and an exploratory surface-feature audit help describe what changes within the group (Table 2). The fixed-share AAE-aligned documents are shorter and have more GPT-2 tokens per byte, with a higher rare-word token share but lower coverage of distinct word types. Marker matches per word are similar across the two selections, although fewer fixed-share documents contain a marker. The pattern gives little support to a simple explanation in which the intervention uniformly increases lexical diversity or dialect-marker density. Full metrics and definitions appear in Appendix B.2.

The comparison suggests a concrete addition to representation audits. Where group labels are available, an audit can compare models trained on filtered data and on an alternative set with the same group counts, using matched training conditions and examining each group's loss. This tests whether the counts conceal consequential differences in the selected examples. The web experiment that

Within the retained AAE-aligned group	Filtered	Fixed-share
Mean characters	45.4	41.1
Mean GPT-2 tokens per byte	0.351	0.369
Pool unigram types covered (%)	39.16	36.88
Rare-token share / pool share	0.957	1.036
Marker matches per 100 words	2.92	2.95
Documents with a marker (%)	17.86	16.62

Table 2: **The same group count can conceal different retained text.** Both columns contain 12,836 AAE-aligned documents. Fixed-share preserves more rare-word token mass but fewer distinct word types. Rare types are the bottom frequency decile of the AAE-aligned training pool's lowercase whitespace types. Marker matches use an existing lexical proxy, not validated dialect labels. These diagnostics are exploratory. Full definitions, both-group summaries, and an untrained stratified-uniform reference are in Appendix B.2 and the supplement.

follows asks whether filtering changes the gap when models receive the same data and training budgets.

3 A WEB COMPARISON AT THE SAME TOKEN BUDGET

The controlled experiment shows that group counts can conceal differences in the documents a model learns from. To examine selection in web pretraining, we train 150M-parameter models from scratch on DCLM-Baseline using the released DataComp-LM quality classifier (Li et al., 2024), while holding the retained data budget and training schedule fixed. The classifier applies an additional top-decile cut to an already-curated pool, so the comparison concerns further curation of DCLM-Baseline rather than its initial construction from raw crawl.

Measuring the gap. We evaluate on 5,000 AAE-aligned and 5,000 white-aligned held-out tweets from TwitterAAE (Blodgett et al., 2016). For each text we compute the mean negative log-likelihood (NLL) per token, then average these values within each group to obtain the dialect gap

$$G = \overline{\mathrm{NLL}}_{\mathrm{AAE}} - \overline{\mathrm{NLL}}_{\mathrm{white}}. \tag{1}$$

Larger values of G indicate worse relative fit to AAE-aligned text, and $\exp(G)$ gives the ratio of the corresponding group perplexities. The labels describe population-level corpus alignment rather than individual authors' identities. Because the natural texts retain differences in topic and content associated with dialect, this outcome cannot isolate linguistic form or establish a task-level disadvantage.

Matching selection and training budgets. The candidate pool contains 166.3M scored documents spanning ten global shards of DCLM-Baseline. Filtering retains the top decile by classifier score, while the uniform control samples across the same ten shards. We shuffle and pack each selected set to an inventory of approximately 16.3B tokens, with inventories matching within 0.4%. Every model then receives 15B training token presentations under a common architecture, tokenizer, optimizer, and update schedule, so no arm completes a pass through its packed inventory. This design estimates the total effect of the selection policy at a matched token budget, allowing document length and other properties of the selected set to vary with the policy. The training recipe appears in Table 6, with packing instances and document counts in Appendix B.

The effect of filtering. The gap rises from 1.158 under region-matched uniform selection to 1.186 under filtering, a difference of $+0.028$ nats with a 95% Welch interval of $[0.011, 0.045]$ across seven filtered and six uniform training seeds. The interval measures training-seed variation on fixed selected datasets; the seed rosters were extended after initial results under rules fixed in advance (Appendix B.1).

The estimated change corresponds to an increase in the perplexity ratio from 3.18 to 3.27, or 2.9% (Figure 1). Byte normalization preserves the direction, giving $+0.009$ nats per byte with interval $[0.003, 0.015]$, although the numerical magnitudes in token and byte units are not directly comparable. Per-group losses are available for the five filtered checkpoints with saved final weights

(Appendix B.1), a smaller roster than that used for the primary gap estimate. Against the six uniform runs, their NLL increases by $+0.031$ nats on AAE-aligned text and $+0.007$ on white-aligned text.

Table 3 brings the checkpoint, extension-only, and byte-normalized readings together. The primary estimate combines final filtered checkpoints with the predeclared average of the last three uniform checkpoints, while using final checkpoints for both arms gives a similar result. All three new filtered runs exceed all three new controls, but the extension-only interval includes zero. Appendix B.1 records the complete rosters and aggregation conventions.

Reading	Filtered/uniform seeds	Unit	Change [95% interval]
Declared checkpoint summaries	$7/6$	nats/token	$+0.028$ $[+0.011, +0.045]$
Both arms at final checkpoint	$7/6$	nats/token	$+0.029$ $[+0.012, +0.045]$
Extension runs alone	$3/3$	nats/token	$+0.026$ $[-0.006, +0.057]$
Byte normalization	$7/6$	nats/byte	$+0.009$ $[+0.003, +0.015]$

Table 3: **Sensitivity of the primary web comparison.** Every row compares filtering with region-matched uniform selection. Intervals are Welch intervals over training seeds on fixed selected datasets. Declared summaries use final filtered checkpoints and the last three uniform checkpoints. Rows reuse runs and are sensitivity readings rather than independent replications.

Control choice and scope. An earlier uniform control sampled a single shard and produced a larger contrast of $+0.050$ nats. The region-matched control draws from the same source regions as the filter, which isolates the selection rule, so we treat it as primary (Appendix B.3).

The direction depends on the pool and budget. In a separate from-scratch experiment on FineWeb (Penedo et al., 2024), the natural-text contrast is -0.017 nats $[-0.035, 0.001]$ across five seed pairs at 124M and 2.46B training tokens (Table 14), and Section 6 examines source-pool history as one explanation.

4 RANDOM-SELECTION CONTROLS

The fixed-share intervention uses dialect labels to hold group counts constant. We also examine cluster-proportional random sampling, CLUSTERRANDOM, which offers a comparison with score-based selection without taking those labels as input. The procedure embeds the full candidate pool with DistilBERT, partitions it into $K = 500$ clusters, and samples randomly within each cluster in proportion to its size. A cluster containing n_k of the pool's N documents receives approximately Bn_k/N places in a retained set of size B, preserving the pre-filter cluster proportions while removing score-based selection within them.

Up to quota rounding, each document has the same marginal inclusion probability, B/N, as under uniform sampling, although clustering controls the number selected from each cluster. Although the procedure takes no dialect labels as input, its representations can still reflect dialect. In the controlled testbed, 48.6% of documents belong to clusters in which at least 80% of the documents come from one alignment group. We therefore describe these as embedding clusters without assuming that they isolate semantic topics. Appendix A.1 documents this association, the original label-informed choice of K, and the subsequent label-free sensitivity check.

Under the production classifier, cluster-proportional random selection returns the testbed gap close to the uniform mean (Table 1). Applying it to DCLM-Baseline with the 150M recipe and matched token budget of Section 3 gives a mean gap of 1.164. Filtering exceeds this by 0.022 nats, with a 95% Welch interval of $[0.008, 0.037]$ across seven filtered and five cluster-random seeds. The contrast supports a reduction in the web gap when selection changes, although CLUSTERRANDOM remains $+0.006$ nats above region-matched uniform selection, with interval $[-0.013, 0.024]$. Random selection thus offers an alternative to the tested quality cut, without evidence that clustering improves on uniform sampling.

Appendix B.3 reports closure ratios against both uniform references; the conclusions here rest on the direct contrasts between selection policies.

5 WHAT ARE THE QUALITY TRADEOFFS?

A smaller dialect gap leaves open how the same selection choices affect other measures of model quality. We examine this on the 150M DCLM-Baseline models used for the web comparison, evaluating the five surviving filtered, six region-matched uniform, and five cluster-random checkpoints at the final training step under one harness. Table 4 reports WikiText-103 perplexity (Merity et al., 2017), held-out Colossal Clean Crawled Corpus (C4) NLL (Raffel et al., 2020), and LAMBADA contextual word-prediction accuracy (Paperno et al., 2016).

Filtering improves both general-text likelihood measures relative to uniform selection while reducing LAMBADA accuracy by 2.03 percentage points. Compared with cluster-random selection, it improves C4 likelihood but reduces LAMBADA accuracy by 1.68 points, with the WikiText-103 interval including zero. Whether stricter quality selection helps therefore depends on the outcome being measured. The LAMBADA result concerns contextual word prediction and does not establish a general reasoning advantage.

The accuracy result also depends on which filtered runs are included. Three of the five surviving checkpoints use a rebuilt packing instance; on these alone, the likelihood contrasts against uniform selection retain their signs and intervals excluding zero, whereas the LAMBADA interval includes zero. The accuracy separation thus depends on pooling all five filtered seeds (Appendix C.2). Its direction also varies across training regimes. In the separate 124M FineWeb experiment, cluster-random selection trails filtering on LAMBADA by about 1.4 points (Table 15), reinforcing the need to evaluate the costs of a selection change within the setting where it will be used.

6 SCOPE ACROSS SCALE, SOURCE POOLS, AND EVALUATION TASKS

The supporting experiments examine whether the selection comparisons extend to other model sizes and scorers, whether reweighting the retained data offers a similar benefit, and whether likelihood differences appear in behavioral tasks. Each comparison addresses a specific boundary of the main results.

Selection and reweighting. Experiments with a frozen GPT-2 perplexity scorer produce a larger selection effect than the production classifier. At 124M, changing selection at approximately the filtered group share again reduces the gap, whereas reweighting the retained documents does not (Appendix Table 11). The 1B experiment compares changing the selected documents with changing their training weights (Table 5). Reweighting leaves the gap close to the filtered level, while sampling from the full pool reduces it in both token and byte units. In these runs, replacing the selected documents therefore helps more than the particular reweighting rule tested. The 1B experiment uses a different base model and scorer within the same short-text training regime and includes no fixed-share arm.

Released web models. Released filter ablations provide further context up to 7B (Magnusson et al., 2025; Emami Gohari et al., 2026). Their natural-text point estimates widen the gap on already-cleaned pools and narrow it on raw-crawl-derived pools, with the deepest cut in the GneissWeb 7B

Arm (n)	WikiText-103 ppl ↓	C4 NLL ↓	LAMBADA acc. (%) ↑
Filtered (5)	21.55 ± 0.14	3.196 ± 0.006	38.74 ± 1.13
Uniform, region-matched (6)	21.77 ± 0.14	3.219 ± 0.003	40.77 ± 0.46
CLUSTERRANDOM (5)	21.71 ± 0.17	3.214 ± 0.004	40.42 ± 1.00
Filtered minus comparison arm: difference [95% interval]			
Uniform, region-matched	-0.22 $[-0.40, -0.03]$	-0.023 $[-0.029, -0.016]$	-2.03 $[-3.41, -0.65]$
CLUSTERRANDOM	-0.16 $[-0.39, +0.07]$	-0.018 $[-0.025, -0.011]$	-1.68 $[-3.25, -0.11]$

Table 4: **Quality outcomes in the primary DCLM-Baseline experiment.** Arm values are means $\pm$ seed sample standard deviations; contrasts use Welch intervals. LAMBADA contrasts are percentage points. The five filtered checkpoints pool two original-pack and three rebuilt-pack runs. All evaluations use final checkpoints.

1B intervention	Gap (nats/token)	Gap (nats/byte)
Uniform selection	0.162 ± 0.028	0.291 ± 0.011
Score selection	0.577 ± 0.011	0.432 ± 0.003
Reweight the score-selected set	0.573 ± 0.021	0.429 ± 0.008
CLUSTERRANDOM selection	0.188 ± 0.020	0.300 ± 0.004

Table 5: **Changing the selected documents and reweighting them have different outcomes at 1B.** Each arm retains 29,491 tweets and continues Llama-3.2-1B for 4,000 updates. Values are means $\pm$ SD over five training seeds. Reweighting uses the same retained set as score selection; CLUSTERRANDOM draws from the full candidate pool. This experiment uses a frozen perplexity scorer and includes no fixed-share arm.

ablation suite yielding $+0.139$ nats. These comparisons share suites and base models, so they should not be counted as independent replications of our controlled design. Appendix D explores source-pool history as an explanation for the differing signs.

Behavioral outcomes. The likelihood widening in the primary 150M experiment has no established counterpart in our behavioral evaluation. On four-way content-word cloze, the filtered-minus-uniform change in the between-group error gap is $+0.63$ percentage points $[-0.12, 1.40]$ (Table 17). The deepest released 7B comparison gives a larger change of $+4.92$ points, with a three-seed t interval of $[2.79, 7.04]$, although the groups start from unequal error rates. To account for that difference, we compare the result with a null that multiplies both groups' error by the white-aligned group's observed factor. The excess is then $+1.26$ points $[-1.76, 3.78]$, leaving degradation beyond this proportional baseline unresolved. Open-vocabulary top-1 word prediction and constrained continuation also have unresolved overall contrasts at the deep cut. Appendix C.4 defines the probes and reports the full results.

Paired rewrites. A secondary evaluation pairs 4,992 natural AAE-aligned tweets with GPT-4o-mini rewrites into white-aligned text (Furquan, 2025). Using five filtered and six uniform final checkpoints, the 150M gap change is $+0.012$ nats $[0.002, 0.021]$, while the FineWeb estimate is $+0.023$ $[-0.001, 0.047]$. Although the pairs reduce content variation, the rewrites also change grammatical regularity, explicitness, and predictability, so a preference for generated prose could contribute to the differential selection effect. We use this comparison as a supporting check on the natural-text result, with no claim that it isolates dialectal form.

7 RELATED WORK

Filtering and dialect disparities. Prior corpus audits quantify unequal removal and quality judgments across social groups and language varieties (Dodge et al., 2021; Gururangan et al., 2022; Lucy et al., 2024; Deas et al., 2025), alongside evidence of uneven data quality in multilingual corpora (Kreutzer et al., 2022). Studies of trained models examine the consequences of detoxification (Welbl et al., 2021; Xu et al., 2021) and broader pretraining-data choices (Longpre et al., 2024), while Hofmann et al. (2024) investigate dialect prejudice in model decisions. The corpus audits measure what a filter keeps from each group but train no model on the result, and the trained-model studies compare data choices without holding either the retained group counts or the training budget fixed. Neither line of work separates a group's share of the data from the choice of documents within that share, or tests reweighting against reselection. We contribute a matched-budget comparison of document selection together with an intervention that holds group counts exactly fixed, which shows that retained proportions cannot fully explain the testbed effect. Following Blodgett et al. (2020), we treat alignment labels and likelihood gaps as corpus-level measures rather than statements about speakers.

Balancing and selecting training data. Methods for constructing training data use domain weights, distribution matching, perplexity, and learned quality judgments (Xie et al., 2023a;b; Thrush et al., 2025; Wettig et al., 2024; Sachdeva et al., 2026). Group-robust learning and reweighting address disparities using available examples (Sagawa et al., 2020; Liu et al., 2021; Kirichenko

et al., 2023; Lahoti et al., 2020). Our fixed-share intervention changes which documents are available while holding group counts constant, highlighting the importance of examining selection before examples are discarded.

Cluster-based sampling. Cluster-balanced sampling and quality-diversity selection are established approaches (Shao et al., 2024; Liu et al., 2025; Zhang et al., 2025), and embeddings have also been used for subgroup discovery and diverse selection (Sohoni et al., 2020; Abbas et al., 2023; Tirumala et al., 2023). We use proportional stratified random sampling to preserve the candidate pool's embedding-cluster proportions. Its role here is to provide intervention evidence and measure the resulting tradeoffs.

8 Limitations

The two central experiments support conclusions at different levels. The balanced, short-text continued-pretraining testbed shows that group proportions alone cannot explain its selection effect, while the from-scratch web comparison establishes widening in one curated pool. We have not tested the fixed-share explanation in web pretraining, and the FineWeb comparison is unresolved. The controlled result should therefore guide further investigation of web selection rather than serve as an established explanation for it. The 63% figure measures the reduction against the filter's own documents at fixed counts; we have not trained a group-stratified uniform arm at the same $12{,}836$ and $16{,}655$ counts, which would separate the share lever from the document lever and is the next experiment. Our evaluation also limits what can be inferred about dialect. TwitterAAE alignment labels are demographic and linguistic proxies rather than validated identities, and the two groups of natural tweets differ in register, topic, and linguistic features. Their likelihood contrast includes associated content differences and cannot separate dialect from differences in informality within social media. Generated rewrites reduce some of this variation while introducing their own confounds. We have not established generalization to spoken African American Language, other dialects, or individual users' experiences, and the likelihood ordering need not carry over to behavioral tasks.

The web estimates depend on small seed rosters that were extended after initial results, without an adjustment for that history in the nominal intervals. Some evaluations are limited to surviving checkpoints, and the filtered extension uses a rebuilt packing instance. These intervals measure training-seed variation conditional on the selected data and evaluation set; repeated training on one uniform or cluster-random subset does not capture variability over independent selection draws. The source-region control cannot supply that missing replication. Nor do we have a trained length-matched control. Filtered web documents average about $1{,}240$ tokens against $1{,}308$ under region-matched uniform selection, leaving length as a possible contributor to the total effect.

Finally, the retained-document diagnostics and cluster–dialect association do not identify which aspect of coverage mediates the improvement. We have not established an advantage for embedding clusters over uniform sampling, arbitrary partitions, or length strata. Cluster-random selection requires suitable candidate data and has mixed quality tradeoffs, which limits its use as a general replacement for quality filtering.

9 Conclusion

Retained group proportions provide an incomplete account of how quality filtering affects model fit. In our controlled continued-pretraining experiment, changing the surviving documents reduces the dialect gap even when group counts remain fixed, with better fit to both groups than the filter's own documents provide. A separate matched-budget web experiment finds widening under an additional quality cut on DCLM-Baseline. In the testbed, reweighting the documents kept by a perplexity filter does not reproduce the benefit of changing them. Together, these findings support evaluating the documents selected within groups alongside the proportions those groups contribute to training.

Reproducibility statement

Every controlled continued-pretraining experiment fits on a single 24 GB A10G GPU, though the 1B runs were parallelized over four GPUs. Each of the twenty-one 150M from-scratch runs, spread over

four arms, trained on four L40S GPUs. Results are reported over multiple seeds, with code, pinned environments, seed records, available run logs, and saved results in the supplementary package. Its index maps the tables and figure to supporting artifacts, and CPU-only verification scripts reproduce the core comparisons. The package also includes the manuscript sources and the Supplementary Analysis Record, whose Section S5 gives the selection bookkeeping.

ETHICS STATEMENT

This work examines a representational disadvantage in data curation using existing data, with the aim of understanding and reducing it. Our dialect experiments draw on the public TwitterAAE corpus (Blodgett et al., 2016), following its convention of describing text as AAE-aligned or white-aligned. These labels express alignment inferred from linguistic content and demographic priors, without establishing individual identity. We acknowledge the African-American Language / White Mainstream English terminology and the interpretive caveats of Deas et al. (2025) and Mire et al. (2025) in measuring a filtering disadvantage on an already under-represented variety.

AI USE STATEMENT

Generative AI tools served as a coding assistant, helped brainstorm ideas, helped find related work, and helped smooth the prose. The authors verified every number, table, figure, and citation against the released artifacts and take responsibility for the final text and claims.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient Learning at Web-scale through Semantic Deduplication. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2303.09540. arXiv:2303.09540.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1119–1130, 2016. doi: 10.18653/v1/D16-1120. URL https://aclanthology.org/D16-1120/. arXiv:1608.08868.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476, 2020. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485/. arXiv:2005.14050.

Nicholas Deas, Blake Vente, Amith Ananthram, Jessica A. Grieser, Desmond U. Patton, Shana Kleiner, James R. Shepard, III, and Kathleen McKeown. Data Caricatures: On the Representation of African American Language in Pretraining Corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29192–29217, 2025. doi: 10.18653/v1/2025.acl-long.1416. URL https://aclanthology.org/2025.acl-long.1416/. arXiv:2503.10789.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit Optimizers via Block-wise Quantization. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2110.02861. arXiv:2110.02861.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1286–1305, 2021. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98/. arXiv:2104.08758.

Hajar Emami Gohari, Swanand Ravindra Kadhe, Yousaf Shah, Constantin M. Adam, Abdulhamid Adebayo, Praneet Adusumilli, Farhan Ahmed, Nathalie Baracaldo, Santosh Subhashrao Borse, Yuan-Chi Chang, Xuan-Hong Dang, Nirmit Desai, Revital Eres, Ran Iwamoto, Alexei A. Karve, Yan Koyfman, Wei-Han Lee, Changchang Liu, Boris Lublinsky, Takuya Ohko, Pablo Pesce, Maroun Touma, Shiqiang Wang, Shalisha Witherspoon, Herbert Woisetschläger, David Wood, Kun-Lung Wu, Issei Yoshida, Syed Zawad, Petros Zerfos, Yi Zhou, and Bishwaranjan Bhattacharjee. GneissWeb: Preparing High Quality Data for LLMs at Scale. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4c61f578ff07830f5c37378dd3ecb0d-Abstract-Conference.html. arXiv:2502.14907.

Furquan. Dialectic Preferences Bias Dataset: AAE–SAE Parallel Tweets. Hugging Face dataset, 2025. URL https://huggingface.co/datasets/furquan/dialectic-preferences-bias-aae-sae-parallel/tree/75a43acfa7e0ff0c453c914143dc4afa762a9bf6. Revision 75a43acfa7e0ff0c453c914143dc4afa762a9bf6, last modified 2025-02-16; MIT license.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, 2024. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/. arXiv:2402.00838.

Suchin Gururangan, Dallas Card, Sarah K. Dreier, Emily K. Gade, Leroy Z. Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 2562–2580, 2022. doi: 10.18653/v1/2022.emnlp-main.165. URL https://aclanthology.org/2022.emnlp-main.165/. arXiv:2201.10474.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. AI Generates Covertly Racist Decisions About People Based on Their Dialect. *Nature*, 633(8028):147–154, 2024. doi: 10.1038/s41586-024-07856-5. URL https://www.nature.com/articles/s41586-024-07856-5. arXiv:2403.00742.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2204.02937. arXiv:2204.02937.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 2022. doi: 10.1162/tacl_a_00447. URL https://aclanthology.org/2022.tacl-1.4/.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H. Chi. Fairness without Demographics through Adversarially Reweighted Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 728–740, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/07fc15c9d169ee48573edd749d25945d-Abstract.html. arXiv:2006.13114.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 14200–14282, 2024. doi: 10.52202/079017-0455. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/19e4ea30dded58259665db375885e412-Abstract-Datasets_and_Benchmarks_Track.html. arXiv:2406.11794.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792. PMLR, 2021. URL https://proceedings.mlr.press/v139/liu21f.html. arXiv:2107.09044.

Fengze Liu, Weidong Zhou, Binbin Liu, Zhimiao Yu, Yifan Zhang, Haobin Lin, Yifeng Yu, Bingni Zhang, Xiaohuan Zhou, Taifeng Wang, and Yong Cao. QuaDMix: Quality-Diversity Balanced

Data Selection for Efficient LLM Pretraining. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2504.16511. arXiv:2504.16511.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3245–3276, 2024. doi: 10.18653/v1/2024.naacl-long.179. URL https://aclanthology.org/2024.naacl-long.179/. arXiv:2305.13169.

Li Lucy, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7393–7420, 2024. doi: 10.18653/v1/2024.acl-long.400. URL https://aclanthology.org/2024.acl-long.400/. arXiv:2401.06408.

Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D. Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, Noah A. Smith, Pang Wei Koh, and Jesse Dodge. DataDecide: How to Predict Best Pretraining Data with Small Experiments. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 42487–42502. PMLR, 2025. URL https://proceedings.mlr.press/v267/magnusson25a.html. arXiv:2504.11393.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Byj72udxe. arXiv:1609.07843.

Meta. Llama 3.2 Model Card. Meta Llama model repository, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. Rejected Dialects: Biases Against African American Language in Reward Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7483–7502, 2025. doi: 10.18653/v1/2025.findings-naacl.417. URL https://aclanthology.org/2025.findings-naacl.417/. arXiv:2502.12858.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1525–1534, 2016. doi: 10.18653/v1/P16-1144. URL https://aclanthology.org/P16-1144/. arXiv:1606.06031.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 30811–30849, 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html. arXiv:2406.17557.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL https://www.jmlr.org/papers/v21/20-074.html. arXiv:1910.10683.

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. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/4b8bd97555d4060104a6b84b80920bab-Abstract-Conference.html. arXiv:2402.09668.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1911.08731. arXiv:1911.08731.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. In *Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS*, 2019. URL https://arxiv.org/abs/1910.01108. arXiv:1910.01108.

Yunfan Shao, Linyang Li, Zhaoye Fei, Hang Yan, Dahua Lin, and Xipeng Qiu. Balanced Data Sampling for Language Model Training with Clustering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14012–14023, 2024. doi: 10.18653/v1/2024.findings-acl.833. URL https://aclanthology.org/2024.findings-acl.833/. arXiv:2402.14526.

Nimit S. Sohoni, Jared A. Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 19339–19352, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/e0688d13958a19e087e123148555e4b4-Abstract.html. arXiv:2011.12945.

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, 2024. doi: 10.18653/v1/2024.acl-long.840. URL https://aclanthology.org/2024.acl-long.840/. arXiv:2402.00159.

Stanford CRFM. HELM TwitterAAE Scenario: Dataset Construction and Loading. HELM source code, 2025. URL https://github.com/stanford-crfm/helm/blob/4db8410f66b38360465220b7a93513b2ea606ea0/src/helm/benchmark/scenarios/twitter_aae_scenario.py. Revision 4db8410f66b38360465220b7a93513b2ea606ea0.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. Improving Pretraining Data Using Perplexity Correlations. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2c625366ae28066fcb1827b44517d674-Abstract-Conference.html. arXiv:2409.05816.

Kushal Tirumala, Dániel Simig, Armen Aghajanyan, and Ari S. Morcos. D4: Improving LLM Pretraining via Document De-Duplication and Diversification. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 53983–53995, 2023. doi: 10.52202/075280-2348. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a8f8cbd7f7a5fb2c837e578c75e5b615-Abstract-Datasets_and_Benchmarks.html. arXiv:2308.12284.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. Challenges

in Detoxifying Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2447–2469, 2021. doi: 10.18653/v1/2021.findings-emnlp.210. URL https://aclanthology.org/2021.findings-emnlp.210/. arXiv:2109.07445.

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*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52915–52971. PMLR, 2024. URL https://proceedings.mlr.press/v235/wettig24a.html.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69798–69818, 2023a. doi: 10.52202/075280-3059. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/dcba6be91359358c2355cd920da3fcbd-Abstract-Conference.html.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data Selection for Language Models via Importance Resampling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34201–34227, 2023b. doi: 10.52202/075280-1482. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6b9aa8f418bde2840d5f4ab7a02f663b-Abstract-Conference.html.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. Detoxifying Language Models Risks Marginalizing Minority Voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2390–2397, 2021. doi: 10.18653/v1/2021.naacl-main.190. URL https://aclanthology.org/2021.naacl-main.190/. arXiv:2104.06390.

Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Jiantao Qiu, Lei Cao, Ju Fan, Ye Yuan, Guoren Wang, and Conghui He. Harnessing Diversity for Important Data Selection in Pretraining Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b588d9b67932b459ea66ff6e2804c6b3-Abstract-Conference.html. arXiv:2409.16986.

A DATA, EVALUATION, AND TRAINING

This appendix contains the methods and results needed for the main argument. The *Supplementary Analysis Record*, included as `paper/analysis_record.pdf` in the code-and-results supplement, preserves the full analysis history, exploratory audits, alternate conventions, and additional experiments. The package's `TOC.md` indexes the current manuscript; historical indices are labeled separately. Experimental code paths are relative to the organized code supplement, whose top-level directories include `training/`, `selection/`, and `eval/`.

Natural-text data and alignment labels. We load `lighteval/TwitterAAE`, a balanced 100,000-tweet dataset derived from TwitterAAE (Blodgett et al., 2016) and used for the Holistic Evaluation of Language Models (HELM) TwitterAAE task. HELM documents selecting the 830,000 tweets with the highest AAE proportions and 7.3 million with the highest white proportions, then randomly sampling 50,000 from each subset (Stanford CRFM, 2025); we have not independently verified that the mirror contains exactly that sample. We impose no additional posterior threshold. The original mixed membership model averages census demographics across each user's geolocated messages to form user-level priors, then infers message-level mixtures using linguistic content. The corpus authors validated linguistic properties of the inferred AAE-aligned text; these population-level labels do not identify individual speakers. We perform no independent human validation of individual labels.

A seed-42 permutation partitions the data into the balanced 90k training pool and a held-out 5k+5k evaluation set. The controlled testbed, primary web experiment, and released-checkpoint comparisons use that natural-text holdout. The FineWeb from-scratch family instead uses unpermuted head slices, sharing 9.9% of AAE-aligned evaluation texts with the frozen holdout. Comparisons within each family use the same evaluation texts. A controlled-testbed near-duplicate exclusion check is reported in the Supplementary Analysis Record; author identifiers are not available in the mirror used here, although the original corpus includes user IDs.

Paired rewrites. The 4,992 pairs are from the released `furquan/dialectic-preferences-bias-aae-sae-parallel` dataset (Furquan, 2025), revision `75a43acf`. Its natural AAE-aligned tweets have more than ten words and non-ASCII characters removed; the counterparts were generated with GPT-4o-mini. The dataset card reports no validation of the rewrites. Changes in punctuation, grammatical regularity, explicitness, and predictability accompany the dialect rewrite.

Controlled training. The 124M models start from GPT-2 and use 4,000 updates, batch size 32, constant learning rate 5×10^{-5}, and a 64-token sequence cap. All selected arms retain 29,491 documents. The DCLM classifier uses a one-shot top-32.8% cut. Stress-scorer experiments use a frozen GPT-2 perplexity ranking with five top-80% rounds. The 1B experiments continue from Llama-3.2-1B (Meta, 2024) with the same update count, batch, learning rate, and sequence cap, using 8-bit AdamW (Dettmers et al., 2022) and mixed precision. The nine main 1B arms share the 29,491-document budget; earlier score-selection rounds retain more documents. This is a short-text experiment, not an evaluation of long-context capability.

Names and selection rules. The *fixed-share* arm changes documents while retaining the filtered group counts; it is stored as `COMPHOLD` in the artifacts. The *restored-share* arm restores the source group proportions while retaining top-scored documents within each group; its artifact name is `ORACLEQUOTA`. Both use alignment labels. CLUSTERRANDOM refers throughout to proportional sampling over embedding clusters with random within-cluster selection. The source label `NOFILTER` denotes matched-count uniform selection in the 1B ladder, but denotes an unselected reference in other experiment families; the displayed names follow the actual construction.

Cluster-count choice. The initial $K = 500$ choice used a labeled retention statistic on $K \in \{50, 200, 500\}$. A later label-free rule, $K = \mathrm{round}(\sqrt{n/2})$ snapped to the nearest available count, selects $K = 200$. The rule and three-seed evaluation were specified before training. Its mean natural gap was 0.3751, compared with 0.3750 for the corresponding $K = 500$ reference. This sensitivity check does not establish an advantage over uniform or arbitrary-stratum sampling.

Units and validity masks. All central contrasts use a common tokenizer within their experiment. Table headings use CI for confidence interval, ppl for perplexity, and acc. for accuracy. Per-token NLL is averaged within texts and then within alignment groups. Stored per-byte rescores divide a text's total loss by the UTF-8 byte length of its decoded span, truncated at 64 tokens, before averaging. The 1B outcomes exclude 176 AAE-aligned and 51 white-aligned texts without valid per-token losses, with the same masks in every arm. Table 12 reports both units. Raw levels should not be compared across tokenizers, base models, or evaluation draws.

A.1 From-scratch web recipe and embedding implementation

Table 6 records the 150M Open Language Model (OLMo) configuration used for the primary web experiment. The single-shard uniform, region-matched/extension, and cluster-random templates agree on these fields; the launch scripts substitute run names, seeds, and data paths. The extension launch also supports resuming the same run. The retained inventories and packing differences are documented in Appendix B.

Configuration	Value
Layers / hidden width / attention heads	12 / 768 / 12
Position / normalization / activation	RoPE / RMSNorm / SwiGLU
Configured MLP ratio / tied embeddings	8 / no
Context length / global batch	2,048 tokens / 192 sequences
Tokens per update / updates	393,216 / 38,157
Optimizer / peak learning rate	AdamW / 0.0042
Adam betas / weight decay	$(0.9, 0.95)$ / 0.1
Schedule / warmup	cosine / 386 updates
Final learning-rate fraction / gradient clipping	0.01 / 1.0
Precision / GPUs per run	bfloat16 mixed precision / four L40S
Vocabulary / embedding rows	50,280 / 50,304
Attention, residual, embedding dropout	0

Table 6: **Primary 150M from-scratch recipe.** The recipe follows the OLMo training stack (Groeneveld et al., 2024; Soldaini et al., 2024). Tokenizer: `allenai/gpt-neox-olmo-dolma-v1_5`. The training seed sets initialization and data order; document selection and packing are fixed per arm. This recipe is distinct from the short-text continued-pretraining configuration. RoPE denotes rotary positional embeddings, RMSNorm root mean square normalization, SwiGLU the Swish-gated linear unit, and MLP a multilayer perceptron.

Code paths here are relative to the root of the organized code supplement. The configuration directory is the supplement's training configurations, with templates `olmo_150m_uniform_single_shard.yaml`, `olmo_150m_region_matched.yaml`, and `olmo_150m_cluster_random.yaml`. The region-matched template is also used for the filtered extension, with the selected data paths substituted.

Embedding long web documents. The web selector uses `distilbert-base-uncased`. It normalizes whitespace, retains the first 2,000 characters, and tokenizes with truncation at 256 tokens. It averages the token embeddings using the attention mask and stores the vectors in 16-bit floating-point precision (fp16), without L2 normalization. It therefore represents a document prefix rather than aggregating chunks over the full document. MiniBatchKMeans uses $K = 500$, seed 42, batch size 4,096, `n_init=1`, and one streamed partial-fit pass with shard and buffer shuffling. The embedding and clustering code is in the supplement.

In the short-text testbed, embeddings instead use a 64-token cap, masked mean pooling, and L2 normalization. MiniBatchKMeans uses $K = 500$, batch size 4,096, and `n_init=5`. The 1B continued-pretraining runs retain the shared 64-token training cap to study the same short-text regime; they do not test long-context performance.

Cluster composition. On the stored 90,000-document testbed partition, weighted majority-label purity is 77.19%: each cluster is assigned its majority alignment label, and the fraction of documents matching those labels is computed in-sample. Moreover, 48.59% of documents lie in clusters

containing at least 80% one alignment group, and 74.03% of AAE-aligned documents lie in AAE-majority clusters. These descriptive summaries establish association between embeddings and alignment. They are neither held-out prediction accuracy nor a causal explanation of mitigation, and they do not describe the web clusters.

Fixed-share construction and interpretation. The fixed-share intervention begins with the full pool and applies five successive retention rounds. Each round samples 80% of its current documents, up to rounding, in proportion to the current cluster sizes. Using dialect labels, it then swaps selected and unselected documents within clusters to target the group counts produced by the frozen filter at the same retention depth. Only the retained documents enter the next round. The construction records include lexical diagnostics and a revision to their validity criteria without a change to the selected sets. These diagnostics do not independently identify a coverage mediator. The final $29{,}491$-document sets each contain $12{,}836$ AAE-aligned documents under filtering and fixed-share selection.

B Primary comparisons and uncertainty

B.1 Web selection, packing, and seed rosters

The primary web experiment scores $166{,}294{,}541$ documents from ten global shards of DCLM-Baseline. The top-decile threshold, 0.97971, retains $16{,}629{,}659$ documents. Packing takes a shuffled prefix to approximately 16.3B tokens: 16.31B in the original filtered pack and 16.26B for region-matched uniform selection. The former contains $13{,}152{,}704$ documents; the latter $12{,}432{,}842$. The single-shard uniform inventory contains approximately 12.50M documents and 16.26B tokens. Each arm receives the same 15B training presentations. The region-matched control varies selection within the same source regions; the earlier single-shard control also changes source composition.

Packing instance and surviving checkpoints. The original filtered pack index was lost. Its three extension seeds use a rebuilt pack of the same frozen selected set, containing $13{,}096{,}160$ documents at the matched token target. This changes which shuffled prefix is trained. The original four filtered seeds average 1.1852 nats and the three rebuilt-pack seeds 1.1875. The difference is small relative to the seed spread, but the experiment does not independently estimate packing variability. Filtered seeds 44 and 45 survive only as saved gap values. Per-group losses, quality, paired-rewrite, and cloze evaluations therefore use the other five filtered checkpoints. Against six uniform runs, their per-group NLL increases are 0.031 nats AAE-aligned and 0.007 white-aligned, point estimates without separate intervals.

Arm	Gaps in seed order	Mean $\pm$ SD
Filtered	1.1716, 1.1754, 1.1939, 1.2000, 1.1915, 1.1811, 1.1898	1.1862 ± 0.0104
Uniform, region-matched	1.1489, 1.1746, 1.1388, 1.1745, 1.1651, 1.1460	1.1580 ± 0.0155
ClusterRandom	1.1482, 1.1691, 1.1721, 1.1557, 1.1739	1.1638 ± 0.0112
Uniform, one shard	1.1339, 1.1365, 1.1386	1.1363 ± 0.0024

Table 7: **Every primary web training seed.** Seeds start at 42 and increase consecutively within each arm. Filtered and single-shard values use final checkpoints; region-matched uniform and ClusterRandom use their last-three-checkpoint averages. Values are natural-text gaps in nats per token. The filtered extension uses a rebuilt packing instance. These are training runs on the selected sets, not independent selection draws.

Sequential design and aggregation. The filtered roster was extended from seeds 42–45 to 42–48; the region-matched uniform roster from 42–44 to 42–47. ClusterRandom was extended from 42–44 to 42–46. Every extension followed a written rule with no outcome-based seed replacement or exclusion, but the decisions followed observed results. Pooled fixed-sample intervals do not adjust for that sequential design. The three new filtered and three new uniform runs alone differ by 0.0256

Contrast	Convention	n_1/n_2	Difference	95% interval
Filtered − region-matched	token, declared	$7/6$	$+0.0282$	$[+0.0112, +0.0451]$
Filtered − region-matched	token, final	$7/6$	$+0.0286$	$[+0.0119, +0.0453]$
Filtered − region-matched	byte, declared	$7/6$	$+0.0090$	$[+0.0032, +0.0147]$
Filtered − CLUSTERRANDOM	token, declared	$7/5$	$+0.0224$	$[+0.0078, +0.0370]$
CLUSTERRANDOM − region-matched	token, declared	$5/6$	$+0.0058$	$[-0.0125, +0.0241]$
Filtered − one shard	token, final	$7/3$	$+0.0499$	$[+0.0401, +0.0596]$
Closure: (filtered − ClusterRandom)/(filtered − uniform)				
Reference	Roster state	$n_F/n_C/n_U$	Ratio	95% interval
One shard	declared references	$4/5/3$	0.438	$[0.17, 0.65]$
Region-matched	declared references	$4/5/3$	0.689	$[0.27, 1.79]$
One shard	extended references	$7/5/3$	0.449	$[0.25, 0.64]$
Region-matched	extended references	$7/5/6$	0.795	$[0.43, 1.47]$

Table 8: **Primary estimates and essential sensitivity comparisons.** Difference intervals are Welch intervals over training seeds; the sequential-design qualification applies. Declared aggregation combines final filtered checkpoints with last-three-checkpoint uniform and ClusterRandom values. Closure uses the same five ClusterRandom seeds in every row; its intervals are the recorded percentile bootstrap intervals over 10^5 resamples. The declared single-shard ratio retains its archived rounded denominator, 0.0489. Complete alternate readings remain in the Supplementary Analysis Record.

nats, with full separation; a one-sided exact permutation test at three versus three has minimum attainable $p = 0.05$.

The declared aggregation reads the last three saved checkpoints for uniform and CLUSTERRANDOM, and final checkpoints for the filtered and single-shard arms. This preserves the inherited arm-specific summaries; it is not a claim that different aggregation is preferable. The extension protocol retained the filtered final-checkpoint convention to match its frozen roster and required both uniform aggregations. The all-final-checkpoint contrast is reported alongside it in Table 8. Intervals are Welch seed-level intervals unless another method is named. They condition on the selected sets and evaluation fixture. Training seeds are not independent data selection draws.

B.2 THE FIXED-SHARE INTERVENTION

All five filtered and fixed-share runs are matched on group counts and training recipe. Their gap reduction is 0.07294 nats with Welch 95% interval $[0.06505, 0.08084]$. A paired-t sensitivity interval is $[0.06432, 0.08156]$, and all five seedwise reductions are positive. The three-seed uniform reference is the existing random-retention baseline, not a five-seed arm; it supplies the denominator of the 63% summary. It does not enter the direct five-versus-five fixed-share contrast.

Seed	Filtered	Fixed-share	Restored-share	CLUSTERRANDOM
42	0.4884	0.4184	0.4063	0.3806
123	0.4963	0.4178	0.3911	0.3779
456	0.4917	0.4133	0.3930	0.3630
789	0.4835	0.4078	0.3848	0.3693
1011	0.4828	0.4207	0.3913	0.3715

Table 9: **Production-classifier testbed seeds.** Every arm retains $29{,}491$ documents; filtered and fixed-share each contain $12{,}836$ AAE-aligned documents. Values are natural-text gaps in nats per token. The uniform reference uses seeds $42, 123, 456$, with mean 0.3721 and SD 0.0135. Arm-level summaries appear in Table 1.

Per-group outcomes. Relative to filtering, fixed-share changes AAE-aligned NLL by -0.13369 nats (95% Welch interval $[-0.15139, -0.11599]$) and white-aligned NLL by -0.06074 ($[-0.07426, -0.04723]$). Both component losses were saved for every filtered and intervention

seed. The original three-seed uniform evaluator saved only their difference and did not save the trained models, so its component losses cannot be recovered from those records.

Within-group document diagnostics. Table 10 compares the saved selected indices, using the same seed-42 pool order as training. The surface measurements were computed without model inference. GPT-2 token counts use the full text, before the training cap; tokens per byte are averaged per document. Marker counts use the existing 35-pattern lexicon, with case-insensitive word-boundary matching. A multiword match counts as one match; the denominator is the whitespace-word count. Capitalization and punctuation are literal character statistics, not validated grammaticality or dialect measures. The cached TwitterAAE release contains only tweet strings, so original AAE posterior distributions are unavailable here.

The lexical rows were already in the construction records and are reproduced from the source texts. Types are lowercase whitespace unigrams or bigrams from the 45k AAE-aligned training pool. Rare types are the lowest-frequency 10% of unigram types, with stable pool-order tie breaking. Rare-token share is their token mass divided by all word tokens, normalized by the same share in the full AAE-aligned pool. Fixed-share raises this ratio from 0.957 to 1.036, but lowers unigram coverage from 39.16% to 36.88% and bigram coverage from 36.08% to 33.48%. No single coverage mediator is identified. The original construction diagnostics informed the analysis; these are not independent confirmatory tests of an explanation.

The stratified-uniform reference samples 12,836 AAE-aligned and 16,655 white-aligned documents without clustering, averaging five draws. These draws characterize selected data only: no models were trained on them, and they do not supply the missing trained stratified-uniform control. The supplement includes all four intervention arms' surface summaries for both groups and the numerical feature bank, without redistributing tweet text.

AAE-aligned retained documents	Filtered	Fixed-share	Stratified uniform
Mean characters	45.39	41.06	41.95
Mean whitespace words	7.68	6.98	7.16
Mean GPT-2 tokens	15.71	14.59	14.67
Mean tokens/byte per document	0.351	0.369	0.364
Documents above 64 tokens (%)	0.95	1.07	0.90
Documents with a marker (%)	17.86	16.62	15.48
Marker matches per 100 words	2.92	2.95	2.67
Uppercase share of letters (%)	8.94	11.46	11.28
Punctuation share of characters (%)	7.37	7.40	7.21
Pool unigram types covered (%)	39.163	36.883	37.434
Pool bigram types covered (%)	36.075	33.484	34.106
Rare pool types covered (%)	29.284	28.785	28.363
Rare-token share / pool share	0.957	1.036	0.994

Table 10: **What changes within the retained AAE-aligned group?** Every column holds the AAE-aligned count at 12,836. Stratified uniform averages five untrained selection draws (seeds 0–4); it is a descriptive reference, not a trained control. Surface features were computed without model inference. The four lexical rows reproduce the saved construction diagnostics. These quantities describe fixed selected sets; they do not estimate a causal mediator or selection-policy uncertainty. Definitions and limitations are in Appendix B.2.

Supporting stress-scorer and 1B results. Table 11 gives the subset of supporting interventions discussed in the main text. The stress scorer produces a much larger effect than the production classifier; the 1B experiment does not include a fixed-share arm. Full arm ladders, exposure controls, and independent testbed-pool resamples are retained in the Supplementary Analysis Record. Those resamples do not estimate selection-draw uncertainty for the primary web experiment.

Additional 1B arm definitions. The score is negative frozen-scorer NLL. Cluster quotas with score fill (`CLUSTERQUOTA`) keep the highest scores within proportional embedding-cluster allocations. Equal group quotas (`ORACLE_QUOTA`) keep the highest scores within each alignment group,

Selection	Seeds	Natural gap (nats/token)
GPT-2 124M, frozen perplexity scorer		
Uniform	3	0.372 ± 0.014
Score selection	5	0.869 ± 0.011
Approximately fixed share	5	0.536 ± 0.015
Reweight retained examples	5	0.892 ± 0.012
CLUSTERRANDOM	5	0.375 ± 0.005
Llama-3.2-1B continued pretraining		
Uniform	5	0.162 ± 0.028
Score selection	5	0.577 ± 0.011
Reweight retained examples	5	0.573 ± 0.021
CLUSTERRANDOM	5	0.188 ± 0.020

Table 11: **Supporting intervention ordering.** Values are means $\pm$ seed SD. All listed selected arms retain $29{,}491$ documents at the shared update budget of their experiment. The stress-scorer fixed-share construction matches proportions approximately (AAE shares 0.295–0.296), whereas the production-classifier comparison in Table 1 matches exact group counts. No fixed-share arm was run at 1B. Absolute levels are not comparable across models or tokenizers. Full ladders and exposure controls are preserved in the Supplementary Analysis Record.

1B arm	Gap (nats/token)	Gap (nats/byte)
Uniform retention	0.1621 ± 0.0277	0.2909 ± 0.0105
Score selection, round 1	0.2274 ± 0.0206	0.2835 ± 0.0072
Score selection, round 2	0.3346 ± 0.0180	0.3260 ± 0.0065
Score selection, round 3	0.4388 ± 0.0202	0.3690 ± 0.0074
Score selection, round 4	0.5229 ± 0.0245	0.4030 ± 0.0074
Score selection, round 5	0.5769 ± 0.0108	0.4324 ± 0.0028
CLUSTERRANDOM	0.1884 ± 0.0201	0.3000 ± 0.0044
Cluster quotas, score fill	0.3237 ± 0.0294	0.3399 ± 0.0082
Equal group quotas, score fill	0.2855 ± 0.0208	0.3351 ± 0.0074
Soft score sampling	0.3165 ± 0.0099	0.3415 ± 0.0034
Disagreement reserve	0.4386 ± 0.0199	0.3812 ± 0.0066
Random reserve	0.5641 ± 0.0266	0.4288 ± 0.0076
Reweight retained documents	0.5728 ± 0.0209	0.4293 ± 0.0080

Table 12: **Unit normalization in the 1B continued-pretraining experiment.** Values are means $\pm$ seed sample standard deviations over five runs from the saved per-text rescore. Per-byte loss divides each text's total loss by the UTF-8 byte length of its decoded span truncated at 64 tokens, then averages over texts using the same validity masks as the per-token column. The first depth rung changes from $+0.0653$ nats/token to -0.0074 nats/byte relative to uniform retention. The legacy artifact name `NOFILTER` denotes the matched-count uniform arm in this experiment. The saved per-text results are in the supplement.

dividing the final budget equally up to rounding. Soft score sampling (SOFTSAMPLE) selects the top B values of $s_i/T + g_i$, with independent Gumbel noise and T equal to the pool's scorer-NLL standard deviation. Disagreement reserve (PROT_R5) protects the 20% of current documents with largest absolute NLL disagreement between two masked-training GPT-2 views; score selection fills the remainder of each 80% retention budget. Random reserve (RANDPROT_R5) protects a random 20% instead. Reweighting (IWEIGHT) keeps the score-selected set and samples training documents with weights proportional to $\exp((\mathrm{NLL}_i - \overline{\mathrm{NLL}})/T)$, clipped at their first and 99th percentiles. Cluster and reserve constructions apply five rounds. The artifact alias NOFILTER denotes matched-count uniform selection in this 1B experiment. In the production-classifier testbed, DCLM_COMPHOLD and DCLM_ORACLEQUOTA mean fixed-share and restored-share, respectively; neither name denotes the missing group-stratified uniform training control.

B.3 CLUSTERRANDOM COMPARISONS AND THE ORIGINAL PREDICTION

The direct filtered-minus-CLUSTERRANDOM contrast is 0.0224 nats on seven versus five seeds. CLUSTERRANDOM minus region-matched uniform is 0.0058 nats, with interval $[-0.0125, 0.0241]$; no advantage over uniform is resolved. The fraction of widening removed is

$$\text{Closure} = \frac{G_{\text{filtered}} - G_{\text{cluster-random}}}{G_{\text{filtered}} - G_{\text{uniform}}}. \tag{2}$$

The original outcome bands and prediction of greater-than-half closure used the single-shard reference, the only uniform control then available. The prediction was not met: closure is 0.438 on the five-seed CLUSTERRANDOM roster against the declared reference rosters. Designating the region-matched denominator primary occurred after both results existed. Its design advantage does not turn this into a prospectively specified successful prediction. Table 8 preserves both references and both roster states. Closure intervals resample the three arm seed sets and recompute the ratio; unstable denominators can produce wide intervals or values above one.

Two initial CLUSTERRANDOM seeds were evaluated before the third finished, under the already-fixed estimand, roster, aggregation, and outcome bands. Their subsequent rescoring reproduced the stored arrays exactly. The two-seed extension reproduced all 117 packed-shard hashes. The complete evaluation trail and alternate readings are preserved in Section S3 of the Supplementary Analysis Record.

C QUALITY, EVALUATION TASKS, AND EXTERNAL SCOPE

C.1 THE SECOND FROM-SCRATCH SETTING: FINEWEB

The fourteen FineWeb runs use randomly initialized GPT-2 124M with sequence length $1{,}024$, $5{,}000$ optimizer steps, and 2.46B training tokens from a pinned 14.9M-document pool. The filtered and uniform arms have five seed pairs, CLUSTERRANDOM three seeds, and the token-matched uniform sensitivity arm one seed. The dialect outcomes average checkpoints $4{,}000$, $4{,}500$, and $5{,}000$. The natural-text contrast is -0.017 nats $[-0.035, 0.001]$, and the paired-rewrite contrast is $+0.023$ $[-0.001, 0.047]$. These unresolved contrasts remain part of the main result's scope.

FineWeb quality outcomes. Table 15 uses final checkpoints, a different aggregation from the dialect table above. It reports six distinct metrics; the duplicated WikiText-2 row in the original battery is omitted because it used the WikiText-103 test split. Seedwise direction counts are descriptive, not confidence intervals or independent tests of algorithmic superiority.

Arm	Paired-rewrite gap	Natural gap	FineWeb NLL	WikiText-103 ppl
DCLM top-10%, seed 42	1.945	1.133	3.696	48.3
Uniform, seed 42	1.919	1.145	3.571	49.2
DCLM top-10%, seed 43	1.960	1.142	3.701	49.2
Uniform, seed 43	1.918	1.154	3.570	49.2
DCLM top-10%, seed 44	1.944	1.137	3.705	52.5
Uniform, seed 44	1.935	1.174	3.568	49.7
DCLM top-10%, seed 45	1.950	1.144	3.689	48.0
Uniform, seed 45	1.909	1.143	3.572	47.9
DCLM top-10%, seed 46	1.931	1.135	3.693	49.1
Uniform, seed 46	1.933	1.159	3.563	48.0
CLUSTERRANDOM, seed 42	1.919	1.144	3.560	48.5
CLUSTERRANDOM, seed 43	1.923	1.162	3.567	48.5
CLUSTERRANDOM, seed 44	1.925	1.171	3.572	47.1
Uniform, token-matched, seed 42	1.908	1.146	3.574	48.3

Table 13: **All fourteen from-scratch runs report the paired-rewrite and natural-text gaps.** Outcomes average the last three checkpoints. The held-out FineWeb set is a uniform sample of the pool, so the filtered arm's higher NLL measures training-distribution mismatch.

Pool and design	Evaluation	Replication	Gap change
DCLM-Baseline	natural text, 150M / 15B	independent arms	$+0.028$ $[+0.011, +0.045]$
DCLM-Baseline	paired rewrites, 150M / 15B	5 vs 6 seeds	$+0.012$ $[+0.002, +0.021]$
FineWeb	natural text, 124M / 2.46B	5 pairs	-0.017 $[-0.035, +0.001]$
FineWeb	paired rewrites, 124M / 2.46B	5 pairs	$+0.023$ $[-0.001, +0.047]$

Table 14: **Matched-budget, from-scratch web comparisons on the natural-text and paired-rewrite gaps.** Natural-text rows carry seed-level 95% intervals. The paired-rewrite rows are secondary evidence. The DCLM-Baseline paired-rewrite row uses five surviving filtered and six uniform final checkpoints; the natural-text row uses the extended seven-versus-six roster and declared aggregation. The DCLM-Baseline natural-text row compares seven filtered seeds with six region-matched uniform seeds, both rosters extended under a rule written down before evaluation. Against the single-shard uniform arm the change is $+0.049$ at that comparison's four and three seeds. Appendix Table 8 gives every unit, checkpoint convention, and roster for both contrasts. Appendix Tables 7 and 13 give every endpoint.

	filter − unif.		CLUSTERRANDOM − unif.		CLUSTERRANDOM − filter	
Metric	mean	favorable	mean	favorable	mean	favorable
WikiText-103 ppl	$+0.150$	$2/5$	-1.796	$3/3$	-1.883	$2/3$
LAMBADA final NLL	-0.320	$5/5$	-0.063	$2/3$	$+0.243$	$0/3$
LAMBADA accuracy	$+0.023$	$5/5$	$+0.008$	$2/3$	-0.014	$0/3$
C4-val NLL†	$+0.118$	$0/5$	-0.006	$3/3$	-0.128	$3/3$
high-score holdout	-0.007	$4/5$	-0.006	$2/3$	-0.003	$3/3$
low-score holdout†	$+0.321$	$0/5$	-0.006	$2/3$	-0.329	$3/3$

Table 15: **Six quality metrics at the final FineWeb from-scratch checkpoint.** The WikiText-2 duplicate of the WikiText-103 evaluation is omitted. Signs are oriented so that a negative NLL or perplexity delta and a positive accuracy delta mean that the first arm is better. "Favorable" counts the seeds in which it is, out of five for the filter contrast and three for the CLUSTERRANDOM ones. †C4-val and the low-score holdout read the filtered model off its own training distribution, so its deficit there measures distribution shift. The three delta columns rest on different seed rosters, so they do not compose additively.

C.2 QUALITY AND PAIRED REWRITES IN DCLM-BASELINE

Table 16 evaluates all nineteen surviving final checkpoints under one fp32 harness. Five checkpoints with previously committed natural-text results first reproduced those gaps to four decimals. The filtered roster comprises two original-pack and three rebuilt-pack checkpoints. The table retains their separate summaries and the packing sensitivity intervals. These outcomes are descriptive, with no pre-specified outcome band. The paired-rewrite contrast uses the generated counterparts described in Appendix A; it does not isolate dialectal form.

On the three rebuilt filtered checkpoints alone, differences from the region-matched control remain resolved for WikiText-103 perplexity (-0.2646, $[-0.4766, -0.0527]$), C4 NLL (-0.0211, $[-0.0372, -0.0050]$), and LAMBADA final-word NLL ($+0.1390$, $[0.0171, 0.2608]$). LAMBADA accuracy is unresolved on that subset (-0.0216, $[-0.0532, 0.0100]$, as fractions): its separation depends on the pooled roster. The rebuilt-only rewrite contrast is $+0.0153$ nats, $[0.0002, 0.0304]$, with a lower bound close to zero. Pack-instance comparisons have only two original-pack seeds. All control-versus-control quality intervals include zero.

Arm	n	WikiText-103 ppl	LAMBADA acc.	LAMBADA final NLL	C4-val NLL	Paired-rewrite gap
Filtered, rebuilt pack	3	21.5008 ± 0.1076	0.3861 ± 0.0136	3.1436 ± 0.0562	3.1977 ± 0.0073	1.7084 ± 0.0070
Filtered, original pack	2	21.6246 ± 0.1827	0.3894 ± 0.0114	3.0999 ± 0.0348	3.1940 ± 0.0012	1.6995 ± 0.0067
Uniform, region-matched	6	21.7655 ± 0.1351	0.4077 ± 0.0046	3.0046 ± 0.0289	3.2188 ± 0.0033	1.6931 ± 0.0037
CLUSTERRANDOM	5	21.7092 ± 0.1717	0.4042 ± 0.0100	3.0245 ± 0.0359	3.2142 ± 0.0040	1.6958 ± 0.0053
Uniform, one shard	3	21.7448 ± 0.1782	0.4051 ± 0.0030	3.0070 ± 0.0197	3.2139 ± 0.0039	1.6924 ± 0.0025

Anchor	Contrast	Difference	95% interval
WikiText-103 ppl	filtered pooled − region-matched	−0.2151	[−0.4028, −0.0274]
	filtered pooled − CLUSTERRANDOM	−0.1589	[−0.3873, +0.0696]
	rebuilt pack − original pack	−0.1238	[−1.0062, +0.7586]
LAMBADA acc.	filtered pooled − region-matched	−0.0203	[−0.0341, −0.0065]
	filtered pooled − CLUSTERRANDOM	−0.0168	[−0.0325, −0.0011]
	rebuilt pack − original pack	−0.0033	[−0.0422, +0.0356]
LAMBADA final NLL	filtered pooled − region-matched	+0.1215	[+0.0606, +0.1825]
	filtered pooled − CLUSTERRANDOM	+0.1016	[+0.0374, +0.1658]
	rebuilt pack − original pack	+0.0437	[−0.0862, +0.1735]
C4-val NLL	filtered pooled − region-matched	−0.0226	[−0.0295, −0.0157]
	filtered pooled − CLUSTERRANDOM	−0.0180	[−0.0252, −0.0108]
	rebuilt pack − original pack	+0.0037	[−0.0134, +0.0209]
Paired-rewrite gap	filtered pooled − region-matched	+0.0117	[+0.0023, +0.0212]
	filtered pooled − CLUSTERRANDOM	+0.0091	[−0.0008, +0.0190]
	rebuilt pack − original pack	+0.0089	[−0.0144, +0.0323]

Table 16: **Quality and paired rewrites in the 150M DCLM-Baseline experiment.** All nineteen surviving final checkpoints use one fp32 harness. Arm values are means ± seed SD; contrasts have unpaired Welch 95% intervals. Accuracy is a fraction. "Filtered pooled" combines two original-pack and three rebuilt-pack seeds. The paired-rewrite gap uses 4,992 natural/generated pairs and does not isolate dialectal form. These values are not directly comparable to the 124M FineWeb battery in Table 15. Appendix C.2 reports the packing sensitivity and roster limits.

C.3 BEHAVIORAL EVALUATION IN THE PRIMARY WEB EXPERIMENT

The 150M cloze evaluation uses the same four-way content-word construction as the released-checkpoint probes below. The primary tier contains 2,000 items per group; the secondary tier contains 2,800. Their overlap means they are not independent replications. The protocol fixed the primary tier, roster, and control before scoring these nineteen checkpoints, after an earlier two-arm descriptive reading was visible. The primary contrast is unresolved. The secondary tier's positive additive result is retained here without replacing the pre-specified primary result.

For filtered arm F and control U, the additive contrast is $\Delta = (e_{F,A} - e_{U,A}) - (e_{F,W} - e_{U,W})$, where e is the group error rate. Proportional excess is $E = e_{F,A} - e_{U,A} \times (e_{F,W}/e_{U,W})$. This asks whether AAE-aligned error rises beyond applying the white-aligned error multiplier to its own

baseline. Each checkpoint first reproduced its saved natural-text gap. Within the filtered roster, the original/rebuilt packing contrast is $+0.24$ points on the cloze gap, descriptively.

Arm error rates, primary tier (2,000 items per side), percent			
Arm	Seeds	AAE-aligned error	White-aligned error
Filtered (DCLM top-10%)	5	25.92	7.95
Uniform retention, region-matched	6	25.07	7.72
ClusterRandom	5	25.85	7.85
Uniform retention, one shard	3	25.33	7.88
Contrasts, primary tier			
Contrast		Additive Δ, points	Proportional excess E, points
Filtered − region-matched (primary)		$+0.63$ $[-0.12, +1.40]$	$+0.12$ $[-1.00, +1.23]$
Filtered − ClusterRandom		-0.03 $[-0.64, +0.61]$	-0.26 $[-1.67, +0.83]$
Filtered − one shard		$+0.52$ $[-0.20, +1.28]$	$+0.37$ $[-0.85, +1.31]$
ClusterRandom − region-matched		$+0.66$ $[-0.13, +1.44]$	–
Contrasts, secondary tier (2,800 items per side)			
Filtered − region-matched		$+0.79$ $[+0.08, +1.47]$	$+0.33$ $[-0.76, +1.44]$
Filtered − ClusterRandom		-0.06 $[-0.45, +0.32]$	–
Filtered − one shard		$+0.44$ $[-0.29, +0.98]$	–
ClusterRandom − region-matched		$+0.85$ $[+0.10, +1.60]$	–

Table 17: **Forced-choice cloze on the nineteen surviving 150M checkpoints.** Error rates are percentages; contrasts are percentage points. Brackets give 95% bootstrap intervals over unpaired training seeds on the fixed items, recomputing proportional excess in each resample. Dashes mean the saved analysis did not report that contrast. The primary tier and control were fixed before scoring. Both primary intervals include zero; the secondary tier does not replace that result. The ClusterRandom and filtered arms have similar gaps on both overlapping tiers. This is a secondary behavioral evaluation, not the primary natural-text outcome.

C.4 Released-checkpoint context and probes

The released ablations vary a filter within a suite and recipe. They provide scale and source-pool context, not independent replications of the fixed-share intervention. Related cells share base models and recipes. Table 20 gives natural-text endpoints and distinguishes evaluation-item uncertainty from uncertainty that also resamples training seeds.

The behavioral fixtures use the same natural holdout. M1 is open-vocabulary top-1 word prediction over 3,760 AAE-aligned and 4,517 white-aligned texts: from the fourth word onward, a greedy continuation at each word boundary is compared with the held-out word after case-folding and punctuation removal. M2 is four-way content-word cloze, with one mid-sentence target in 2,000 items per group and three distractors from the same group's corpus, matched in frequency within a factor of two. Candidates are alphabetic words of at least three characters, appearing at least three times and outside a short stoplist. Whole-sequence likelihood ranks the four alternatives. M3 is constrained continuation over 500 prompts per group, with one true and three distractor continuations.

Every model in a comparison receives identical items. The deep-cut M2 rescore using mean per-token likelihood gives $+4.0$ points, positive at all three seeds; the prompted multiple-choice variant sits near chance. These checks and the subgroup breakdowns do not establish a general behavioral deficit for the 150M primary experiment.

Seed uncertainty and unequal base error. For M2, the three seedwise additive differences are $4.35, 5.90, 4.50$ points at the deep cut, $0.05, 0.85, 1.40$ at the shallow cut, and $2.75, 2.95, 0.90$ on DataDecide. Three-seed t intervals, rather than item-only intervals, support our additive conclusions. An alternative hierarchical bootstrap gives $[3.35, 6.47]$, $[-0.45, 2.02]$, and $[0.58, 3.80]$, respectively; the last conflicts with the small-sample t reading, so we retain the latter and make no positive DataDecide claim.

At both 7B cuts, base error is 14.80% AAE-aligned versus 4.23% white-aligned. White-aligned error becomes 5.70% deep and 4.07% shallow. For DataDecide, base errors are 17.42% and 4.93%, and white-aligned error becomes 6.38%. Applying these white-side factors to the AAE-aligned baselines defines the proportional null above. The deep seedwise excesses ($0.61, 4.48, -1.64$ points) disagree in sign. We do not infer a general behavioral deficit from the positive deep additive result.

Released-suite scope. The DataDecide qc-10% and qc-20% arms use Ai2's reproduction of the DCLM classifier; orig-DCLM 10% uses the original classifier. The 7B GneissWeb comparisons use three filtered and three base seeds, each trained for 350B tokens. FineWeb-Edu comparisons give positive natural-text gap changes even though the in-house FineWeb contrast does not. Differences in filter, model, and training budget prevent attribution of these signs to source-pool history alone. Further readout and subgroup analyses are retained in Section S3 of the Supplementary Analysis Record.

Instrument	Deep 7B, +0.139	Shallow 7B, +0.047	DataDecide 1B, +0.060
M1, top-1 word	+0.01 [−0.40, +0.41]	+0.65 [+0.33, +0.97]	−0.09 [−0.44, +0.27]
content words	−0.35 [−0.76, +0.05]	+0.40 [+0.07, +0.75]	−0.15 [−0.53, +0.23]
function words	+0.99 [+0.23, +1.75]	+1.39 [+0.78, +2.01]	+0.02 [−0.65, +0.70]
M2, four-way cloze	+4.92 [+3.77, +6.10]	+0.77 [−0.08, +1.62]	+2.20 [+1.10, +3.30]
M3, continuation	−0.73 [−3.60, +2.13]	+0.60 [−1.47, +2.60]	+1.13 [−1.60, +3.93]

Table 18: **Released-pair behavioral probes with item-level uncertainty.** Values are accuracy difference-in-differences in percentage points; positive means greater filtering cost on AAE-aligned text. Intervals are 95% paired bootstraps over evaluation items. They do not capture training-seed uncertainty: for M2, use the three-seed t intervals in Table 19. In particular, the DataDecide item interval above excludes zero but its seed interval does not; we make no positive claim for that cell. Column headings give natural-text gap changes in nats/token. These are released filter ablations, not matched-budget experiments.

Reading	Deep 7B, +0.139	Shallow 7B, +0.047	DataDecide 1B, +0.060
Additive, in points	+4.92 [+2.79, +7.04]	+0.77 [−0.92, +2.45]	+2.20 [−0.61, +5.01]
Excess over proportional	+1.26 [−1.76, +3.78]	+1.18 [−0.52, +2.73]	−1.47 [−4.26, +0.93]

Table 19: **Released-pair cloze: additive and proportional readings.** The additive row uses 95% Student-t intervals on three seedwise difference-in-differences, with the same method in every column. The second row uses 10^4 item-bootstrap resamples of base error rates pooled over seeds; the nonlinear statistic need not equal the mean of seedwise values. All proportional intervals include zero. The deep-cut additive result therefore does not establish a deficit beyond proportional degradation. The DataDecide additive result is unresolved over seeds. Appendix C.4 defines the null and reports base errors.

Pair (filtered − control), natural instrument	Δgap	Per-text 95% CI	Seed-outer 95% CI	seeds	vs prop.
already-cleaned pools					
DCLM qc-20% − DCLM-Baseline, 1B/100BT	+0.062	[+0.048, +0.076]	[+0.038, +0.086]	9/9	2.31
DCLM qc-10% − DCLM-Baseline, 1B/100BT	+0.060	[+0.045, +0.076]	[+0.029, +0.091]	9/9	1.23
FineWeb-Edu 10% − base, 1B/100BT	+0.120	[+0.101, +0.139]	[+0.092, +0.148]	9/9	0.79
FineWeb-Edu 3% − base, 1B/100BT	+0.164	[+0.140, +0.188]	[+0.131, +0.197]	9/9	0.74
FineWeb-Edu − FineWeb, 1.8B/350BT†	+0.124	[+0.099, +0.149]	n/a	n/a	0.65
+C4 rules − base CC, 1.8B/28BT†	+0.039	[+0.022, +0.056]	n/a	n/a	1.45
IBM GneissWeb bloom ablation, 7B/350BT	+0.0467	[+0.0369, +0.0563]	[+0.021, +0.073]	9/9	n/a
IBM GneissWeb FineWeb-Edu-score-2, 7B/350BT	+0.1388	[+0.1252, +0.1526]	[+0.116, +0.162]	9/9	1.57
raw-crawl-derived pools					
orig-DCLM 10% − Falcon+CC, 1B/100BT	−0.063	[−0.076, −0.051]	[−0.103, −0.027]	0/9	n/a
repro-DCLM qc-10% − Falcon+CC	−0.032	[−0.045, −0.018]	[−0.072, +0.005]	1/9	n/a
repro-DCLM qc-20% − Falcon+CC	−0.020	[−0.031, −0.010]	[−0.057, +0.013]	3/9	n/a
DCLM − uniform, 160M/3.2BT†	−0.022	[−0.039, −0.005]	n/a	n/a	n/a
DCLM − uniform, 410M/8.2BT†	−0.030	[−0.046, −0.013]	n/a	n/a	n/a
DCLM − uniform, 1B/20BT†	−0.022	[−0.039, −0.005]	n/a	n/a	n/a

Table 20: **Natural-text gaps on released checkpoints.** Gaps and intervals are nats/token, AAE-aligned minus white-aligned. Per-text intervals use 10^4 paired item bootstraps after pooling training seeds; seed-outer intervals use hierarchical bootstraps over seeds then items. † marks one-checkpoint arms with no training-seed replication. "Seeds" counts positive filtered-by-base pairings out of nine, not nine independent runs. "vs prop." divides the gap change by the change expected if both groups scaled by the white-side NLL factor; n/a denotes a reversed sign or near-zero denominator. Seed-level Welch intervals include zero for all three raw-crawl-derived cells and exclude zero for the already-cleaned multi-seed cells. Related cells share base models and recipes; they do not identify pool history as the cause of heterogeneous effects.

D Exploratory analyses and open explanations

Pool history is a hypothesis about the heterogeneous signs, not an identified explanation. Our exploratory retention audit uses a 35-pattern marker lexicon. Its only web audit with recorded counts examined 20 documents flagged by a shortened 12-term variant; four were judged plausibly AAE by one labeler. The descriptive Wilson 95% interval is approximately $[8.1\%, 41.6\%]$. This does not validate the full lexicon, establish precision across pools, or supply labels for a web fixed-share intervention. The audit has no demonstrated predictive lift beyond its calibration pools. It is not needed to establish either central trained-model comparison.

The full lexicon, sampled-audit limitations, pool/depth tables, and failed predictions are preserved in Section S3 of the Supplementary Analysis Record, along with synthetic-junk experiments. The latter examine a constructed quality tradeoff and do not establish a deployable improvement for the primary web pool. Sections S4–S5 record the allocation/temperature family and frozen-scorer identities, including the required group/complement partition assumption. These are supporting analyses rather than additional headline contributions.