CURAWEB: JOINT OPTIMIZATION OF QUALITY, REDUNDANCY, AND DIVERSITY FOR WEB-SCALE PRETRAINING DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-web corpora curated via highly selective filters, such as FineWeb-Edu and DCLM, constitute the core of LLM pretraining data and have significantly advanced LLM performance. However, these pipelines typically rely on singular optimization objectives, which inevitably narrows distributional diversity and marginalizes long-tail knowledge, thereby restricting data coverage and underutilizing the vast potential of the open web. To address this limitation, we propose a novel curation paradigm that shifts from linear pruning to the joint optimization of quality, redundancy, and diversity. This framework synergizes dual-track cleaning (rule-based and model-driven) with hybrid deduplication (n-gram and semantic), while employing a multi-objective sampler to balance informational quality with distributional breadth. Applying this framework to Common Crawl, we construct CURAWEB, a 2T-token English corpus. Unlike existing resources, CURAWEB establishes an industrial-grade standard for data curation by recovering a more holistic data distribution with enhanced diversity and minimal redundancy, achieving broader coverage of long-tail knowledge across diverse domains. At the 3B scale, CURAWEB significantly outperforms state-of-the-art baselines, yielding an average performance gain of 1.8% across a wide range of benchmarks, particularly in knowledge-intensive and reasoning tasks.

1 INTRODUCTION

In the current paradigm of Large Language Models (LLMs) pre-training, the quality and diversity of web-scale corpora, such as Common Crawl, are recognized as the core variables determining the upper bound of a model's emergent capabilities and reasoning performance Xi et al.; Achiam et al. (2023); Liu et al. (2024); Yang et al. (2025); Team et al. (2025); Grattafiori et al. (2024); Adler et al. (2024). However, prevailing data curation frameworks, such as FineWeb-Edu Penedo et al. (2024) and DCLM Li et al. (2024a), traditionally rely on a singular optimization approach, deploying monolithic quality classifiers to filter data. Our research indicates that while this paradigm effectively enhances general stylistic quality, it inevitably narrows distributional diversity. Furthermore, existing pipelines treat curation as isolated, sequential stages—cleaning, deduplication, and sampling—each driven by the same uni-dimensional quality signal. This structural isolation causes the same distributional bias to compound across the entire pipeline, leaving quality-diversity trade-offs systematically unresolved. Together, these limitations render such frameworks inadequate for the rigorous requirements of industrial-scale pre-training.

To address these limitations, we propose a unified data curation framework that shifts the paradigm from disjointed, stage-wise filtering—namely, isolated cleaning, deduplication, and sampling—toward multi-dimensional signal-driven global governance. Specifically, we train specialized models to annotate all pre-cleaned data with fine-grained signals, including multi-aspect quality scores and domain tags. Crucially, these multi-dimensional attributes serve as core dependencies across the entire pipeline, enabling consistent funnel monitoring and strategy optimization. This holistic approach ensures a well-calibrated balance between data quality, domain coverage, and token diversity.

To operationalize this holistic approach, our framework introduces three core technical innovations: First, we refine heuristic filtering Rae et al. (2021); Penedo et al. (2023); Wenzek et al. (2020) by

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.

replacing uniform constraints with domain-specific thresholds, thereby preventing the inadvertent removal of high-value samples in rigorous fields like mathematics and code. To complement these calibrated heuristics, we further incorporate writing quality and content value signals derived from our data understanding model to perform granular preliminary filtering. These domain-aware heuristics and multi-dimensional quality signals enable the retention of significantly more high-value data than conventional baselines while preserving convergence efficiency.

Second, we introduce a hybrid deduplication pipeline to address deep semantic redundancy that evades standard n-gram filters. We identify two critical types of hidden redundancy: *Templatized Clusters* (documents sharing a structural skeleton but differing in localized parameters) and *Redundant Crawling* falling just below the Locality-Sensitive Hashing (LSH) threshold. To handle these boundary cases, we propose a soft semantic deduplication strategy based on a weighted voting mechanism. By accumulating penalty scores across multiple similarity edges, this approach classifies a document as redundant only when its aggregated score exceeds a predefined threshold. This strategy avoids the aggressive over-filtering of hard-threshold approaches Huo et al. (2025), reducing the false positive rate from 37.5% to 28.03% in the highest similarity interval.

Third, we introduce an automated sampling framework that jointly optimizes quality and diversity. Documents are evaluated along two complementary dimensions: writing quality (e.g., coherence and logic), which is quantified via a harmonic mean to heavily penalize any severe single-dimension flaw; and content value, scored via linear summation to reward specialized expertise and high information density. This combined quality metric is further integrated with a diversity score to ensure broad domain representation. To prioritize highly informative content, we adopt Power Sampling, which amplifies the selection probability of top-tier documents.

Experimental results on a 1T-token scale demonstrate that CURAWEB significantly outperforms strong baselines like FineWeb-Edu, Dolma3, and DCLM, particularly in knowledge-intensive and reasoning tasks Cobbe et al. (2021); Amini et al. (2019); Hendrycks et al. (2021); Wang et al. (2024); Lai et al. (2017); Zellers et al. (2019); Bisk et al. (2020); Welbl et al. (2017); Sakaguchi et al. (2021); Mihaylov et al. (2018); Olmo et al. (2025). The robust performance gains across diverse benchmarks validate that our full-stack collaborative optimization effectively maximizes the token efficiency of the training corpus. This provides strong empirical evidence that meticulous data refinement and distribution reshaping are critical paths for advancing the intelligence of LLMs.

2 RELATED WORK

Coarse-Grained Data Filtering A significant portion of raw Common Crawl data consists of low-quality and duplicate content, requiring rule-based filtering and deduplication. Rule-based filtering has steadily evolved from surface-level text statistical features to deep structured extraction. Early frameworks such as CCNet introduced language identification and KenLM perplexity scoring Wenzek et al. (2020), while C4 applied line-level heuristics to remove residual HTML markup and boilerplate noise Dodge et al. (2021). More recently, RefinedWeb and FineWeb2 leverage complete HTML structures to achieve more precise text extraction Penedo et al. (2023; 2025). For deduplication, n-gram fingerprint matching via MinHash-LSH remains dominant Broder (1997); Indyk & Motwani (1998), with recent studies extending to embedding-based semantic deduplication through vector databases to capture deeper content redundancy Abbas et al. (2023). However, these approaches suffer from two fundamental limitations: first, rule-based filters apply uniform thresholds across all domains, inadvertently discarding high-value content in specialized fields such as mathematics and science; second, existing deduplication methods rely on fixed hard thresholds for binary filtering, incurring the risk of over-filtering domain-specific documents that naturally exhibit dense terminological overlap Lee et al. (2022); Abbas et al. (2023); Shen et al. (2023). CURAWEB addresses both limitations by introducing domain-aware heuristic thresholds and a soft semantic deduplication strategy based on a weighted voting mechanism, thereby preserving high-value long-tail content while maintaining robust deduplication effectiveness.

Fine-Grained Content Optimization Web corpora after coarse-grained filtering still suffer from excessive low-value content and severe domain distribution imbalances. Consequently, content quality assessment has shifted from rule-driven paradigms to sophisticated model-driven approaches. Earlier protocols, such as GPT-3 and LLaMA, trained binary classifiers on high-quality reference corpora to distinguish premium documents Brown et al. (2020); Touvron et al. (2023), whereas

RefinedWeb constructed a multi-dimensional scoring system integrating URL authority, content length, and structural features Penedo et al. (2023). FineWeb-Edu and DCLM pushed this trajectory further by employing monolithic quality classifiers trained on highly curated reference sets, achieving strong downstream benchmark performance, albeit at the expense of narrowing distributional diversity Penedo et al. (2024); Li et al. (2024a). To regulate domain distributions, techniques like doremi deploy weighted sampling to amplify the probability of selecting high-quality or low-resource domain documents Xie et al. (2023), while the recently proposed WebOrganizer framework introduces automated topic and format classification for precise domain mixing Wettig et al. (2025). However, these methods traditionally treat quality and diversity as disjoint objectives optimized in isolation, leaving the inherent trade-off between them systematically unresolved. CURAWEB reconciles this tension by unifying multi-aspect quality and diversity signals into a shared multi-dimensional framework, enabling globally consistent joint optimization across the entire data curation pipeline.

3 THE CURAWEB CURATION FRAMEWORK

3.1 OVERVIEW

CURAWEB jointly manages quality, redundancy, and diversity through shared document-level signals. As shown in Figure 1, the pipeline extracts English text, annotates quality and domains, applies domain-aware rule and model filters, removes lexical and semantic duplicates, and samples a 2T-token corpus. The same annotations guide filtering and sampling and track distribution changes throughout the pipeline, allowing quality control to preserve long-tail content. Different stages use these signals for different decisions: domain labels protect technical documents during heuristic filtering, quality dimensions identify specific defects, and content-value scores guide the final sampling distribution. This shared description makes the effects of successive stages comparable while allowing each stage to address a distinct source of data loss or redundancy.

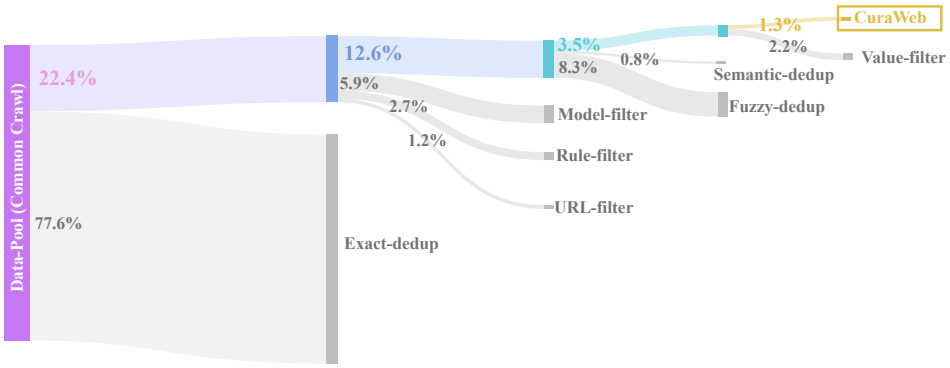


Figure 1: The CURAWEB construction pipeline. Percentages indicate relative document flow or removal at each stage.

3.2 TEXT EXTRACTION

We process Common Crawl snapshots from 2013–2024 following DCLM Li et al. (2024b) and FineWeb Penedo et al. (2024). Resiliparse Bevendorff et al. (2018) extracts main-body text from WARC files, removing headers, footers, and advertisements. A pretrained FastText classifier Joulin et al. (2017) retains English documents with confidence above 0.5. Global SHA-256 document hashing reduces 184.9B documents to 41.4B unique documents (22.4% retention). An extended URL blacklist, combining public lists with custom rules against spam and adult-content domains, leaves 39.1B documents for annotation and filtering (94.5% retention).

3.3 DATA UNDERSTANDING SYSTEM

We annotate documents before filtering to make both selection decisions and distribution shifts observable. Each document receives nine quality scores: *writing*, *coherence*, *completeness*, *logicality*,

safety, *knowledge*, *education*, *helpfulness*, and *reasoning*. The first four characterize presentation and structure; the last four characterize content value. All use a 0–3 scale except safety, which is binary. Keeping these signals separate allows different stages to target specific defects without conflating presentation with informational value.

Domain labels follow a two-level taxonomy adapted from Google AdSense. We categorize over one million web documents with GPT-4o and merge semantically similar categories, obtaining 26 coarse themes and 105 fine categories. These labels support domain-aware filtering and monitoring of long-tail coverage; the taxonomy is listed in Appendix B. Because annotations are available before cleaning, we can compare retained and rejected documents within each domain and identify filtering policies that disproportionately remove otherwise valuable content.

For scalable annotation, we distill GPT-4o Hurst et al. (2024) labels from 500K stratified documents into a 300M-parameter Transformer with separate linear heads for quality assessment and domain classification. Both tasks share a backbone and run in one forward pass. Uncertainty-based loss weighting Kendall et al. (2018) dynamically adjusts task weights to alleviate gradient imbalance and support stable, balanced convergence. Distillation limits teacher annotation to the sampled documents while making the shared feature system available across the full corpus. Prompts and rubrics appear in Appendix C; the quality-score correlation analysis is in Appendix G.

3.4 Data Filtering

Filtering combines domain-aware heuristics with conservative model-based rejection. The rule stage removes structural noise while protecting technical content; the model stage checks defects that surface statistics cannot reliably detect.

3.4.1 Rule Filter

Uniform heuristics can reject useful technical text: for example, alphabetic-character thresholds penalize formulas and LaTeX. Starting from Gopher Rae et al. (2021), we assess each rule using model annotation and manual verification. We remove five low-precision or over-aggressive rules governing median word length, required-word count, transformed-text length, bullet-point fraction, and ellipsis fraction. Their removal protects short-word technical text, long specifications, and structured outlines. We adjust three thresholds: minimum word count from 50 to 25, the symbol-to-word ratio cutoff from 0.1 to 0.38, and the alphabetic-character ratio cutoff from 0.8 to 0.26.

Documents identified as Mathematics, STEM, or Code bypass heuristics and proceed directly to model filtering. The bypass only skips heuristic rejection; protected domains remain subject to the subsequent quality checks. On sampled data, the revised rules achieve 90% filtering precision and retain 90–95% of high-quality documents. This stage retains 34.2B of the 39.1B input documents, compared with 12.7B under the original uniform thresholds. The complete rules, modification rationale, and filtering evaluation are in Appendix D.

3.4.2 Model Filter

Relaxed heuristics preserve coverage but can also admit low-quality text. We therefore use the annotation model's *writing*, *coherence*, *completeness*, and *safety* scores to reject documents that fall below an empirically selected threshold in any dimension. This intersection of quality requirements removes malformed, incoherent, incomplete, or unsafe content while leaving content-value ranking to the sampling stage. It reduces the corpus from 34.2B to 23.2B documents. This conservative cascade addresses the trade-off introduced by rule relaxation: heuristics preserve potentially useful candidates, and model scores check semantic defects that character statistics cannot resolve. A strong score in one checked dimension does not compensate for failing another.

3.5 Hybrid Deduplication

We first apply MinHash-LSH over n-grams to remove near-duplicates with high lexical overlap, then cluster document embeddings to identify residual semantic redundancy. Two recurring cases motivate the second stage: *templatized clusters*, whose documents share a structural skeleton but vary in local parameters, and *redundant crawls*, whose small updates fall below the LSH threshold. A

single cosine-similarity cutoff can also discard distinct technical texts with shared terminology. For example, tax-rate pages and salary calculators can differ in local values while preserving the same template, yielding large repetitive clusters after lexical deduplication. Repeated crawls retain minor updates or formatting differences that can obscure redundant content. We identify these patterns through intra-cluster similarity analysis and manual spot checks.

We select embeddinggemma-300m Team et al. (2024) using a GPT-4o-annotated benchmark over similarities $[0.90, 1.00]$ in 0.01 intervals. It provides the strongest discrimination and most consistent monotonic relationship between similarity and duplication among the four evaluated models. Calibration is needed because the same cosine threshold can correspond to different duplication rates in different embedding spaces Ethayarajh (2019). We use observed interval-level false positive rates to configure the evidence weights. Comparisons and redundancy examples appear in Appendix E.

Soft semantic deduplication. Within each semantic cluster, we process candidates in batches. A strict intra-batch threshold first resolves conflicts, retaining earlier-index representatives as leaders. Each leader then contributes a similarity-dependent penalty to unprocessed documents, and a document is marked redundant when its accumulated penalty reaches the target score of 10. Interval weights increase from 2 to 8 over the calibrated range $[0.90, 1.00]$; for example, $[0.98, 0.99)$ receives 7 and $[0.99, 1.00]$ receives 8. One edge in the highest interval contributes 8 points and remains insufficient, whereas two such edges contribute 16 points and meet the target. Lower-similarity edges require more confirmations. The strict intra-batch conflict check is separate.

After each batch, candidates and documents marked redundant leave the unprocessed set; remaining targets retain their accumulated scores for subsequent batches. Accumulating evidence reduces sensitivity to individual false matches, lowering the measured false positive rate from 37.5% to 28.03% in the highest interval and from 34.8% to 21.89% in $[0.98, 0.99)$. Appendix E reports the full weights, error rates, and algorithm.

3.6 Data Sampling

From 6.4B deduplicated documents, we select a training corpus under a fixed token budget. Building on SampleMix Xi et al., we combine semantic diversity with content value, then use power sampling to favor informative documents.

Diversity scoring. We group document embeddings with K-Means ($K \approx 49{,}000$). For document x_i in cluster C_j, we combine intra-cluster compactness and inter-cluster separation:

$$d_i = d_{\text{intra},j} \times d_{\text{inter},j} \tag{1}$$

The compactness term assigns lower scores to overly dense clusters, while the separation term rewards semantically distinct regions. Documents within the same cluster share this diversity score, while their quality scores remain document-specific. This discourages oversampling redundant content and complements document-level quality.

Content-value scoring. After the model filter enforces basic quality requirements, we define content value q_i as the sum of the *knowledge*, *education*, *helpfulness*, and *reasoning* scores. This additive score rewards specialized expertise without requiring uniformly high scores on every value dimension: a technically informative document can contribute knowledge even when it offers little practical guidance or instructional content. Separating this reward from the earlier defect checks avoids using presentation alone as a proxy for information value. We remove documents with $q_i \leq 1$.

Power sampling. We fuse normalized quality and diversity into the sampling weight:

$$s_i = \alpha \cdot q_i + (1 - \alpha) \cdot d_i \tag{2}$$

where $\alpha = 0.6$ balances quality and diversity. Power Sampling replaces SampleMix's softmax:

$$p_i = \frac{\tilde{x}_i^n \cdot \sigma(k(\tilde{x}_i - a))}{\sum_j \tilde{x}_j^n \cdot \sigma(k(\tilde{x}_j - a))} \tag{3}$$

Here $\tilde{x}_i \in [0, 1]$ is the min-max normalized value of s_i, n controls the power-law emphasis on high scores, σ is the sigmoid function, k controls gate steepness, and a sets the gate center. Larger k makes the cutoff sharper, and scores below a receive progressively smaller weights. The power

term emphasizes high-scoring documents, while the gate suppresses low-scoring ones. Quality and diversity thus affect the same document-level probability: high content value raises priority, and semantic distinctiveness can increase the representation of less redundant regions. Together, they shape the training distribution within the token budget.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Baselines and evaluation. We compare against four widely used public corpora: DCLM Li et al. (2024b), FineWeb-Edu Penedo et al. (2024), Nemotron-CC Su et al. (2025), and Dolma3 Soldaini et al. (2024), using the same token budget. Evaluation uses `lm-evaluation-harness` Gao et al. (2023) and covers mathematical reasoning, multidisciplinary knowledge, reading comprehension, and commonsense reasoning. The ten tasks are GSM8K-Platinum Cobbe et al. (2021), MathQA Amini et al. (2019), MMLU Hendrycks et al. (2021), MMLU-Pro Wang et al. (2024), RACE Lai et al. (2017), HellaSwag Zellers et al. (2019), PIQA Bisk et al. (2020), SCIQ Welbl et al. (2017), Winogrande Sakaguchi et al. (2021), and OpenBookQA Mihaylov et al. (2018). MMLU and MMLU-Pro use 5-shot evaluation. We report individual task accuracy and the unweighted macro-average across all ten benchmarks.

Implementation Details. We train all models from scratch at the 3B parameter scale. To isolate the pre-training corpus as the experimental variable, we use the same architecture and training hyperparameters across runs (Table 5). The primary comparison uses a fixed budget of 200B tokens. We then extend CURAWEB and DCLM, the two best-performing configurations at 200B, to 1T tokens to examine whether the performance advantage persists during longer training.

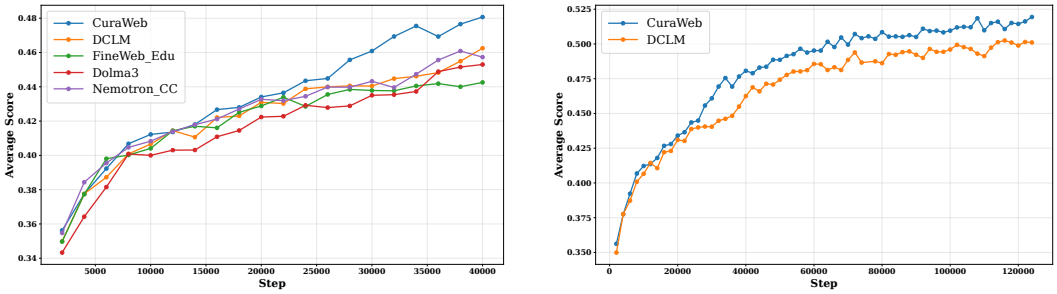
Table 1: Downstream task performance evaluated at 200B training tokens. The best results are indicated in bold, and the second-best results are underlined.

Dataset	GSM8K	MathQA	MMLU	MMLU-Pro	RACE	HellaSwag	PIQA	SCIQ	Wino.	OBQA	Avg.
DCLM	3.72	24.02	39.15	11.11	**36.94**	<u>71.63</u>	77.37	90.30	**66.46**	41.80	<u>46.25</u>
FineWeb_Edu	3.64	<u>25.53</u>	28.02	9.20	35.02	67.16	76.28	90.20	63.30	**44.20**	44.26
Nemotron_CC	2.56	24.96	36.25	10.04	36.75	**72.76**	**78.13**	**92.10**	63.46	40.30	45.73
Dolma3	<u>3.80</u>	24.92	<u>40.56</u>	<u>12.08</u>	36.65	66.94	75.84	88.70	62.90	40.60	45.30
CuraWeb	**8.19**	**26.00**	**47.16**	**14.90**	<u>36.84</u>	71.09	<u>77.86</u>	<u>90.90</u>	<u>63.85</u>	<u>43.90</u>	**48.07**

4.2 MAIN RESULTS

Table 1 and Figure 2a compare downstream performance at 200B tokens. CURAWEB achieves 48.07% average accuracy, exceeding the strongest baseline, DCLM (46.25%), by 1.82 percentage points. From the results, we have the following key observations. **(1) Reasoning and knowledge-intensive tasks.** The largest gains occur on benchmarks that require mathematical reasoning or broad knowledge. Relative to the strongest baseline, Dolma3, CURAWEB improves GSM8K by 4.39 points (8.19 versus 3.80), MMLU by 6.60 points (47.16 versus 40.56), and MMLU-Pro by 2.82 points (14.90 versus 12.08). These gains span mathematical reasoning and multidisciplinary questions, consistent with prioritizing knowledge-rich content. **(2) Performance across specialized baselines.** The baselines have different strengths: FineWeb-Edu leads on OpenBookQA (44.20), while Nemotron-CC leads on HellaSwag, PIQA, and SCIQ. CURAWEB remains competitive on these tasks, including 43.90 on OpenBookQA and 77.86 on PIQA, while improving the overall average. Its broad ranking does not imply uniform dominance: HellaSwag remains below DCLM and Nemotron-CC. The main benefit is a stronger overall balance, with especially large gains in reasoning and knowledge. **(3) STEM-related evaluation.** Compared with FineWeb-Edu, CURAWEB gains 0.47 points on MathQA and 0.70 on SCIQ. Together with its stronger GSM8K result, these observations are consistent with preserving useful technical content during curation. The corpus-level STEM retention analysis below provides complementary evidence, while the controlled ablations examine the individual pipeline components. **(4) Scaling behavior.** Figure 2b tracks CURAWEB and DCLM as training increases from 200B to 1T tokens. The average-accuracy margin grows from 1.82 to 1.95 points, and the two curves continue to improve over the training range. Thus, the advantage is sustained beyond the initial comparison budget and does not disappear during longer training. This

comparison tests the two leading configurations under matched model settings; it supports the use of the curated corpus at larger token budgets without attributing the gain to any single filtering or sampling decision.



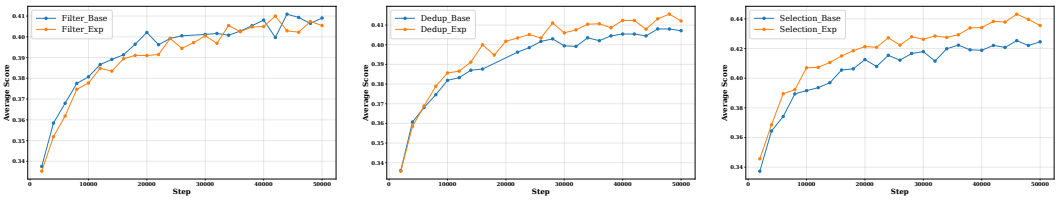
(a) Downstream performance at 200B tokens.

(b) Scaling trends up to 1T tokens.

Figure 2: Corpus comparison at 200B tokens and scaling to 1T tokens.

4.3 Ablation Studies

We conduct three controlled ablations, replacing one component at a time while keeping the remaining pipeline fixed. Figure 3 compares the resulting downstream performance curves for filtering, deduplication, and sampling.



(a) Effect of Data Filtering.

(b) Effect of Deduplication.

(c) Effect of Sampling Strategy.

Figure 3: Ablations comparing the full pipeline with (a) rule-only filtering, (b) removal of semantic deduplication, and (c) uniform sampling, with all other components fixed.

Filtering: preserving high-quality data yield. The full filtering pipeline retains twice as much high-quality data as the rule-only configuration while achieving similar downstream performance (Figure 3a). This result highlights a distinction between data yield and accuracy at a fixed training budget. Domain-aware heuristics and model-based checks preserve a larger usable pool without lowering the observed training curve, supporting the recall-oriented design of Section 3.4.

Deduplication: removing residual redundancy. Removing semantic deduplication shifts the training curve downward (Figure 3b). Lexical matching alone therefore leaves redundant content that affects downstream performance. The result complements the duplicate-count analysis: semantic matching removes additional documents after fuzzy deduplication, including templated pages and slightly modified recrawls. Its effects are visible in both corpus size and performance.

Sampling: allocating the training budget. Replacing quality-diversity sampling with uniform sampling also reduces performance (Figure 3c). Even after filtering and deduplication, the remaining documents are not equally useful under a fixed token budget. Content-value scoring and Power Sampling prioritize informative documents, while diversity scores account for semantic coverage. These ablations show that the components contribute in different ways: filtering expands usable data, and deduplication and sampling improve the observed training outcomes.

5 ANALYSIS

We analyze coverage, diversity, and quality to understand how the pipeline reshapes the corpus alongside the downstream gains. Domain and lexical comparisons use equal-sized samples of 50M documents per corpus; full-pipeline document counts are reported separately. The principal comparisons appear below, with per-domain plots, additional redundancy analysis, and sampling diagnostics in Appendices H, I, and J.

5.1 DATA COVERAGE

Distribution changes during filtering. Rule filtering produces small distribution shifts, with major categories such as Arts, Shopping, and Online Communities decreasing by about 0.2–0.7 percentage points. Model filtering is more selective: Adult, Online Communities, Games, and Reference decline more, while long-tail categories such as Religion and Apparel are less affected. These patterns support complementary roles for structural heuristics and model-based quality checks. Full retained and removed distributions appear in Appendix H.

Comparison with public corpora. CURAWEB reaches L1/L2 entropies of 4.295/6.075 bits, above FineWeb-Edu (3.733/5.477) and Dolma3 (4.086/6.034), but slightly below DCLM and Nemotron-CC (Table 2). Its three largest domains account for 26.5% of documents, compared with 48.2% for FineWeb-Edu. Thus, it preserves a less concentrated distribution than the educationally focused baseline while remaining competitive with the broader corpora.

Table 2: Domain distribution entropy across corpora (50M documents each). Higher values indicate less concentrated domain distributions.

Dataset	L1 entropy (bits)	L2 entropy (bits)
Common Crawl	4.442	5.917
DCLM	4.308	6.122
Nemotron-CC	4.380	6.148
Dolma3	4.086	6.034
FineWeb-Edu	3.733	5.477
CuraWeb	4.295	6.075

Coverage across pipeline stages. L1 entropy remains near 4.42 through rule filtering (4.419), model filtering (4.428), fuzzy deduplication (4.427), and semantic deduplication (4.417). It then decreases to 4.330 after value filtering and 4.295 after sampling. Most of the reduction therefore accompanies content-value prioritization, while cleaning and deduplication preserve broad coverage. Figure 4 shows the full trajectory, complementing the downstream evaluation of this quality-coverage trade-off.

5.2 REDUNDANCY AND LEXICAL DIVERSITY

Complementary duplicate removal. Fuzzy deduplication reduces the corpus from 23.29B to 7.97B documents; semantic deduplication removes a further 19.4%, leaving 6.42B and a cumulative reduction of 72.4%. In mathematics and physics, semantic matching identifies an additional 682.4K duplicates, bringing the total to 1.65M. For utility-oriented websites, detection increases from 2.33M to 3.93M. These results show that semantic matching captures templates and recrawls missed by lexical overlap; the ablation in Figure 3b confirms a corresponding training benefit.

Lexical diversity through curation. The unique n-gram ratio $\mathcal{R}_n$ divides the unique count by the total count. In 50M-document samples, it rises from 0.000373 after rule filtering to 0.000560 after semantic deduplication. Sampling increases unique n-grams from 4.73M to 5.13M, but lowers the ratio to 0.000352 as the total grows from 8.5B to 14.6B. The final sample therefore has greater absolute lexical coverage despite a lower unique-to-total ratio.

Comparison with baseline diversity. FineWeb-Edu has the lowest unique count (3.18M) and ratio (0.000213). CURAWEB's ratio matches Nemotron-CC and exceeds DCLM (0.000301) and Dolma3 (0.000324), although Dolma3 has the largest unique count (5.67M). These complementary metrics

show competitive diversity alongside the strongest average downstream score. Complete counts and stage-level statistics are provided in Table 10 in Appendix I.

5.3 Data Quality and STEM Retention

Quality across stages and corpora. Table 3 compares the final corpora. CuraWeb's composite score is 0.521, above Dolma3 (0.457), DCLM (0.448), and Nemotron-CC (0.392). It leads in PracticalHelpfulness and ranks second in EducationValue, Knowledge, and Reasoning. FineWeb-Edu scores higher overall (0.610), with lower domain entropy. Within our pipeline, the composite increases from 0.300 after fuzzy deduplication to 0.302 after semantic deduplication, 0.436 after value filtering, and 0.521 after sampling; earlier cleaning stages are not monotonic (Appendix J).

Table 3: Content-value scores of the final corpora. Best results are bold; second-best results are underlined. Full pipeline scores appear in Table 11.

Dataset	EducationValue	Knowledge	PracticalHelpfulness	Reasoning	Composite_score
FineWeb-Edu	**1.963**	**1.387**	<u>0.562</u>	**1.202**	**0.610**
Dolma3	1.233	0.983	0.527	0.923	0.457
DCLM	1.230	0.887	0.486	0.936	0.448
Nemotron-CC	1.114	0.684	0.398	0.640	0.392
CuraWeb	<u>1.498</u>	<u>1.164</u>	**0.751**	<u>0.987</u>	<u>0.521</u>

Retention of high-value technical content. For documents scoring at least 2 on the 0–3 scale, CuraWeb's mean retention across Computer Science, Programming, Mathematics, and Physics is 0.432, compared with 0.427 for FineWeb-Edu and 0.261–0.376 for the other baselines (Table 4). CuraWeb leads in Computer Science and Mathematics; FineWeb-Edu leads in Programming and Physics. This indicates competitive technical-content retention alongside greater domain entropy than FineWeb-Edu.

Table 4: High-quality document retention rate ($\geq$2 out of 3) across key STEM domains.

Dataset	Computer Science	Programming	Mathematics	Physics	Avg
FineWeb-Edu	0.429	**0.451**	0.405	**0.424**	0.427
Dolma3	0.350	0.332	0.426	0.396	0.376
DCLM	0.354	0.336	0.407	0.358	0.364
Nemotron-CC	0.252	0.208	0.301	0.282	0.261
CuraWeb	**0.436**	0.425	**0.448**	0.419	**0.432**

Filtering and sampling effects. Model filtering reduces 34.18B documents to 23.29B, removing 31.9%. Across the pipeline, the STEM share grows from 19.74% to 33.58%, and the math share from 1.656% to 3.832% (Table 12). These composition changes differ from retention rates and are not monotonic across stages. After sampling, Health accounts for 9.11%, Science for 7.83%, and Jobs and Education for 7.04%. The redistribution accompanies higher content-value scores and absolute lexical coverage, with modest decreases in entropy and the unique-to-total ratio. Detailed domain shifts and STEM composition are in Appendix J.

6 Conclusion

We present CuraWeb, a 2T-token English corpus built through joint management of quality, redundancy, and diversity. Shared document-level quality and domain signals connect domain-aware filtering, hybrid lexical and semantic deduplication, and quality-diversity sampling. Controlled 3B-model experiments show a 1.82-point average accuracy gain over DCLM at 200B training tokens and a sustained 1.95-point advantage at 1T. The ablations and corpus analyses illustrate complementary effects: preserving usable data, removing residual repetition, and directing the training budget toward informative content. Together, these results support coordinated curation that retains broad coverage while improving the value of the final training distribution.

AI USE STATEMENT

Generative AI tools were used to improve the language and presentation of the manuscript. GPT-4o was also used for document quality and domain annotation and to label the similarity-interval benchmark for embedding-model selection. The authors conceived the core ideas, designed the algorithms, implemented the methods, and conducted the experiments. All AI-assisted revisions were reviewed by the authors, who accept full responsibility for the final manuscript, its claims, and all associated code and artifacts.

REPRODUCIBILITY STATEMENT

To support reproducibility, we use publicly available Common Crawl snapshots and evaluation benchmarks, with references provided in the main text. The methodology section details CURAWEB's data understanding system, domain-aware filtering, hybrid deduplication, and quality–diversity sampling. The appendix provides the reported training configurations and hyperparameters, domain taxonomy, annotation prompts, heuristic filtering rules, and soft deduplication pseudocode.

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 arXiv:2303.09540*, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2357–2367, 2019.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Elastic chatnoir: Search engine for the clueweb and the common crawl. In *European conference on information retrieval*, pp. 820–824. Springer, 2018.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Andrei Z Broder. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pp. 21–29. IEEE, 1997.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 55–65, 2019.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

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

Bi Huo, Bin Tu, Cheng Qin, Da Zheng, Debing Zhang, Dongjie Zhang, En Li, Fu Guo, Jian Yao, Jie Lou, et al. dots. llm1 technical report. *arXiv preprint arXiv:2506.05767*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604–613, 1998.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomáš Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th conference of the European chapter of the association for computational linguistics: volume 2, short papers*, pp. 427–431, 2017.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7482–7491. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00781. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.html.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 785–794, 2017.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024a.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Scott Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F. 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 M. Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Raghavi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. 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, Alex 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 Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/19e4ea30dded58259665db375885e412-Abstract-Datasets_and_Benchmarks_Track.html.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2381–2391, 2018.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. Fineweb2: One pipeline to scale them all — adapting pre-training data processing to every language. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=jnRBe6zatP.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, et al. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*, 2023.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 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.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*, 2024.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 2459–2475. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.123/.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Meituan LongCat Team, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, et al. Longcat-flash technical report. *arXiv preprint arXiv:2509.01322*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37: 95266–95290, 2024.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the twelfth language resources and evaluation conference*, pp. 4003–4012, 2020.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*, 2025.

Xiangyu Xi, Deyang Kong, Jian Yang, Jiawei Yang, Zhengyu Chen, Wei Wang, Jingang Wang, Xunliang Cai, Shikun Zhang, and Wei Ye. Samplemix: A sample-wise pre-training data mixing strategey by coordinating data quality and diversity.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

A Training Configuration

Table 5 lists the architecture and optimization hyperparameters used for training.

Table 5: Training hyperparameters.

Hyperparameter	Value
Parameters	3B
Architecture	LLaMA (RMSNorm, SwiGLU, RoPE, GQA)
Layers	32
Hidden size	4096
FFN hidden size	14336
Attention heads	32
KV heads (GQA)	8
Sequence length	8192
Vocabulary size	131072
Precision	BF16
Optimizer	Adam Kingma & Ba (2014)(β_1=0.9, β_2=0.95, ϵ=10^{-8})
Learning rate	3×10^{-4} (cosine decay, min 3×10^{-5})
Warmup steps	122,070
Global batch size	960
Weight decay	0.1
Gradient clipping	1.0
Training tokens	200B / 1T

B Domain Taxonomy

Table 6 presents the 26 primary domain (L1-domain) categories utilized in our data understanding system. Adapted from the Google AdSense classification hierarchy, each L1 category is further subdivided into fine-grained secondary (L2) subcategories, yielding a total of 105 distinct categories.

Table 6: 26 L1-domain categories.

ID	Domain	ID	Domain
1	`Adult`	14	`Internet_and_Telecom`
2	`Arts_and_Entertainment`	15	`Jobs_and_Education`
3	`Autos_and_Vehicles`	16	`Law_and_Government`
4	`Beauty_and_Fitness`	17	`News`
5	`Books_and_Literature`	18	`Online_Communities`
6	`Business_and_Industrial`	19	`People_and_Society`
7	`Computers_and_Electronics`	20	`Pets_and_Animals`
8	`Finance`	21	`Real_Estate`
9	`Food_and_Drink`	22	`Science`
10	`Games`	23	`Sensitive_Subjects`
11	`Health`	24	`Shopping`
12	`Hobbies_and_Leisure`	25	`Sports`
13	`Home_and_Garden`	26	`Travel_and_Transportation`

C Annotation Prompts

Quality Scoring Prompt The following prompt is used to obtain supervision labels for the quality dimensions. Table 7 summarizes the scoring rubric: eight dimensions use a 0–3 scale, while Safety is binary.

Quality Scoring Prompt

```
You are a large language model training data annotator tasked
with evaluating the quality of text data and scoring the text
from each of the following dimensions separately.

The dimensions include four expression-related dimensions:
Writing, Coherence, Completeness, and Logicality; four value-
oriented dimensions: Knowledge, EducationValue,
PracticalHelpfulness, and Reasoning; and one safety-related
dimension: Safety.

Name: Writing
Description: Evaluate the spelling, grammar, punctuation, and
overall writing quality of the text.
Scoring Criteria:
0: Severe Writing Errors: the text is either unreadable or
incomprehensible because the content is dominated by serious
writing errors, including garbled text, abnormal symbols, lack of
 proper spacing/punctuation, repetitive content, or grammar/
spelling mistakes.
1: Readable with Obvious Errors: the text contains multiple
obvious errors in spelling, grammar, or punctuation that
interrupt the flow of reading and impede overall comprehension,
but the content remains generally readable and understandable.
2: Good Writing with Minor Flaws: the text is well-written with
only a few ignorable mistakes that have little impact on reading
and understanding. The writing flows naturally and effectively
communicates its message despite these small imperfections.
3: Excellent Writing: the text demonstrates exemplary writing
quality without writing errors. It is polished, precise, and
highly effective in its use of language.

...

Requirements:
Evaluate the texts based on the above dimensions, provide
evaluation reasoning first, and then give a quality evaluation
score. Output format in JSON:

{
  "Evaluation Reasoning":
  {
    "Writing": "...",
    "Coherence": "...",
    "Completeness": "...",
    "Logicality": "...",
    "Knowledge": "...",
    "EducationValue": "...",
    "PracticalHelpfulness": "...",
    "Reasoning": "...",
    "Safety": "..."
  },
  "Writing": X,
  "Coherence": X,
  "Completeness": X,
  "Logicality": X,
  "Knowledge": X,
  "EducationValue": X,
  "PracticalHelpfulness": X,
  "Reasoning": X,
  "Safety": X
}
```

```
Evaluate the following text as a whole: $TEXT$
```

Table 7: Scoring rubric for quality annotation dimensions.

Dimension	Scoring Criteria
Writing	**0**: Severe writing errors; the text is unreadable or incomprehensible due to garbled text, abnormal symbols, repetition, or serious grammar/spelling errors. **1**: Readable with obvious errors; multiple spelling, grammar, or punctuation errors interrupt reading and impede comprehension. **2**: Good writing with minor flaws; only a few negligible mistakes that do not affect understanding. **3**: Excellent writing; polished, precise, and free of writing errors.
Coherence	**0**: Severely incoherent; ideas are randomly arranged with abrupt shifts between unrelated or conflicting topics. **1**: Partially coherent; some basic connections exist, but poor transitions and fragmented organization make the topic hard to follow. **2**: Generally coherent; ideas are mostly logically ordered with only minor disruptions or inconsistencies. **3**: Excellent coherence; ideas flow smoothly with clear thematic development and effective transitions.
Completeness	**0**: Severely incomplete; major necessary components are missing, leaving substantial unanswered questions. **1**: Partially complete; basic information is present, but core messages or important components are omitted or underdeveloped. **2**: Generally complete; most key points are covered, with only minor missing details. **3**: Comprehensively complete; all necessary points are addressed with appropriate depth and cohesion.
Logicality	**0**: Severe logic errors; the text contains contradictions, unclear causality, fallacies, or unsupported claims. **1**: Partially logical; basic consistency exists, but flaws, insufficient support, or contradictions weaken the argument. **2**: Generally sound logic; arguments are mostly valid and supported, with only minor lapses. **3**: Exemplary logical reasoning; claims are well-supported, connections are explicit, and no contradictions or fallacies are present.
Knowledge	**0**: Minimal knowledge value; the text is superficial, sparse, or mostly uninformative. **1**: Basic knowledge with limited depth; factually adequate but lacks analytical depth, novelty, or conceptual integration. **2**: Substantial knowledge; demonstrates good conceptual understanding, professional terminology, and meaningful connections. **3**: Advanced knowledge; shows exceptional depth, originality, professional insight, and sophisticated conceptual integration.

Dimension	Scoring Criteria
EducationValue	**0**: Negligible educational value; irrelevant to education or dominated by inappropriate or promotional content. **1**: Limited educational utility; contains some relevant information but is poorly aligned, disorganized, or unsuitable for teaching. **2**: Moderately educational; suitable for school-level use, with clear presentation and substantial educational content. **3**: Excellent educational resource; well-suited for teaching, with clear reasoning, learning value, and appropriate depth.
Practical	**0**: No practical value; vague, irrelevant, or lacking actionable guidance and implementable suggestions. **1**: Limited practical application; contains some useful elements but lacks specificity, guidance, or adaptability. **2**: Moderately helpful; provides actionable steps, useful recommendations, and problem-solving structure. **3**: Exceptionally practical; offers clear step-by-step instructions, concrete examples, and executable methods.
Reasoning	**0**: No reasoning value; contains no reasoning-related content or has fundamental reasoning errors. **1**: Limited reasoning value; attempts reasoning but includes noticeable errors or conclusions unsupported by premises. **2**: Generally sound reasoning; mostly accurate reasoning with appropriate evidence-conclusion connections. **3**: Exemplary reasoning; rigorous, methodical, well-supported, and free from logical fallacies.
Safety	**0**: Unsafe content; involves serious issues such as cheating, drugs, gambling, pornography, violence, horror, anti-human activity, or illegal transactions. **1**: Mainly safe; the main content is unrelated to unsafe issues, even if a few unsafe phrases appear.

Domain Classification Prompt The following prompt is used to classify web page contents into predefined domain categories.

Domain Classification Prompt

```
You are a highly skilled web page domain classifier. Your task is
 to identify the correct domain for a given web page content
based on predefined categories.

The top-level domains are:
  1. Adult
  2. Arts_and_Entertainment
  3. Autos_and_Vehicles
  4. Beauty_and_Fitness
  5. Books_and_Literature
  6. Business_and_Industrial
  7. Computers_and_Electronics
  8. Finance
  9. Food_and_Drink
  10. Games
  11. Health
  12. Hobbies_and_Leisure
  13. Home_and_Garden
  14. Internet_and_Telecom
```

```
      15. Jobs_and_Education
      16. Law_and_Government
      17. News
      18. Online_Communities
      19. People_and_Society
      20. Pets_and_Animals
      21. Real_Estate
      22. Science
      23. Sensitive_Subjects
      24. Shopping
      25. Sports
      26. Travel_and_Transportation

    The second-level domains are:
    ...

    Each domain is structured as:
    /PARENT_DOMAIN/SUB_DOMAIN
```

D Comprehensive Comparison of Gopher and CuraWeb Heuristic Rules

Table 8 presents the complete set of heuristic filtering rules, comparing the original Gopher Rae et al. (2021) thresholds with our recalibrated thresholds. Rules designated as "Removed" were eliminated due to an excessive risk of over-filtering high-quality content, while those labeled "Retained" maintain their original thresholds.

Calibration and evaluation. We evaluate individual rules on sampled documents using model-aided annotation and manual verification. Five rules are removed because of low filtering precision or excessive rejection of technical content, and three thresholds are relaxed. For example, symbol and alphabetic-character constraints otherwise reject mathematical formulas and LaTeX, while length and bullet-point rules can remove technical specifications and structured outlines. Documents assigned to Mathematics, STEM, or Code bypass the rule stage and proceed to model filtering.

On the sampled evaluation set, the revised rule filter achieves 90% filtering precision and retains 90–95% of high-quality documents. It retains 34.2B of the 39.1B input documents, compared with 12.7B under the original uniform thresholds. These results motivate the recalibration; Table 8 provides both the complete configuration and the rationale for each change.

E Hybrid Deduplication: Calibration and Algorithm

Residual redundancy after lexical deduplication. Manual spot checks and intra-cluster similarity analysis identify two recurring patterns. *Templatized clusters* share a structural skeleton but vary localized parameters, as in jurisdiction-specific tax pages, salary calculators, and color-code reference sheets. Their differing n-gram profiles can evade fuzzy deduplication; individual clusters contain hundreds of thousands of documents, with duplication rates reaching 99.96%. *Redundant crawling* produces copies of the same webpage with small updates or formatting changes that remain below the LSH threshold.

Embedding-model calibration. Cosine-similarity thresholds depend on the embedding space Ethayarajh (2019), and specialized documents may share terminology without being duplicates. We therefore construct a GPT-4o-annotated benchmark spanning $[0.90, 1.00]$ in 0.01 intervals and compare four models. Jina-v3 Sturua et al. (2024) exhibits prefix bias and a false positive rate above 40% in high-similarity intervals. Qwen3-0.6B and Qwen3-4B Yang et al. (2025) show weaker overall discrimination. We select embeddinggemma-300m Team et al. (2024) for its discrimination and consistent monotonic relationship between cosine similarity and duplication.

Interval weights and observed error rates. Table 9 gives the empirically calibrated weights and corresponding false positive rates. Soft deduplication reduces FPR from 37.5% to 28.03% in

Table 8: Complete heuristic rule comparison: Gopher original vs. CuraWeb optimized.

Rule	Gopher Original	CuraWeb	Status	Motivation
Document-level rules				
word_count (min)	≥ 50	≥ 25	Relaxed	Retain concise knowledge snippets
word_count (max)	$\leq 100{,}000$	$\leq 100{,}000$	Retained	—
mean_word_length	3–10	3–10	Retained	—
symbol_to_word_ratio (hash/ellipsis)	> 0.1	> 0.38	Relaxed	Tolerate symbol-dense formulaic content
alphabetic_char_ratio	< 0.80	< 0.26	Relaxed	Prevent STEM data loss from LaTeX/formulas
stop_word_filter	≥ 2 stop words	≥ 2 stop words	Retained	—
Removed rules (high over-filtering risk)				
median_word_length	≤ 3	—	Removed	Reduces misjudgment of short-word technical texts
required_word_count	≤ 2	—	Removed	Eliminates redundant low-precision constraint
transformed_text_len	$\geq 10{,}000$	—	Removed	Prevents purging of long technical specifications
fraction_bullet_points	≥ 0.9	—	Removed	Preserves well-structured technical outlines
fraction_ellipses	≥ 0.3	—	Removed	Removes low-precision constraint
Repetition removal rules (retained from Gopher, Table A1)				
Duplicate line fraction	> 0.30	> 0.30	Retained	—
Duplicate paragraph fraction	> 0.30	> 0.30	Retained	—
Duplicate line char fraction	> 0.20	> 0.20	Retained	—
Duplicate paragraph char fraction	> 0.20	> 0.20	Retained	—
Top 2-gram char fraction	> 0.20	> 0.20	Retained	—
Top 3-gram char fraction	> 0.18	> 0.18	Retained	—
Top 4-gram char fraction	> 0.16	> 0.16	Retained	—
Duplicate 5–10-gram char fraction	> 0.15–0.10	> 0.15–0.10	Retained	—
STEM bypass mechanism (CuraWeb only)				
Mathematics / STEM / Code domains	N/A	Skip rule filter	Added	Directly forward to model-based filter

$[0.99, 1.00]$ and from 34.8% to 21.89% in $[0.98, 0.99)$. Multiple confirmations reduce sensitivity to individual false matches. These are empirical results; correlated similarity edges do not by themselves imply a formal exponential error bound.

Table 9: Similarity-interval weights and false positive rates (FPR) for single-edge decisions and soft deduplication ($T_{\text{target}} = 10$).

Cosine similarity	Weight w_k	Edge FPR	Soft FPR
$[0.99,\ 1.00]$	8.0	37.5%	**28.03%**
$[0.98,\ 0.99)$	7.0	34.8%	**21.89%**
$[0.96,\ 0.98)$	6.0	42.2%	23.13%
$[0.95,\ 0.96)$	5.0	54.5%	29.70%
$[0.94,\ 0.95)$	4.0	65.8%	35.49%
$[0.92,\ 0.94)$	3.0	71.6%	34.45%
$[0.90,\ 0.92)$	2.0	85.0%	44.37%

Batched weighted-score accumulation. Algorithm 1 operates independently within each K-Means cluster. It processes candidates in batches, first resolving intra-batch conflicts at a strict threshold t_{strict} and retaining earlier indices as leaders $\mathcal{L}$. Each leader ℓ then contributes a penalty to unprocessed targets $u \in U$:

$$w^* = \max\{w_k \mid \text{sim}(E_\ell, E_u) \geq t_k,\ (t_k, w_k) \in \mathcal{T}\} \tag{4}$$

We add w^* to $S[u]$ and mark u as redundant when $S[u] \geq T_{\text{target}} = 10$. As every interval weight is below 10, this score-based decision requires more than one contributing edge. The strict intra-batch conflict check remains a separate step.

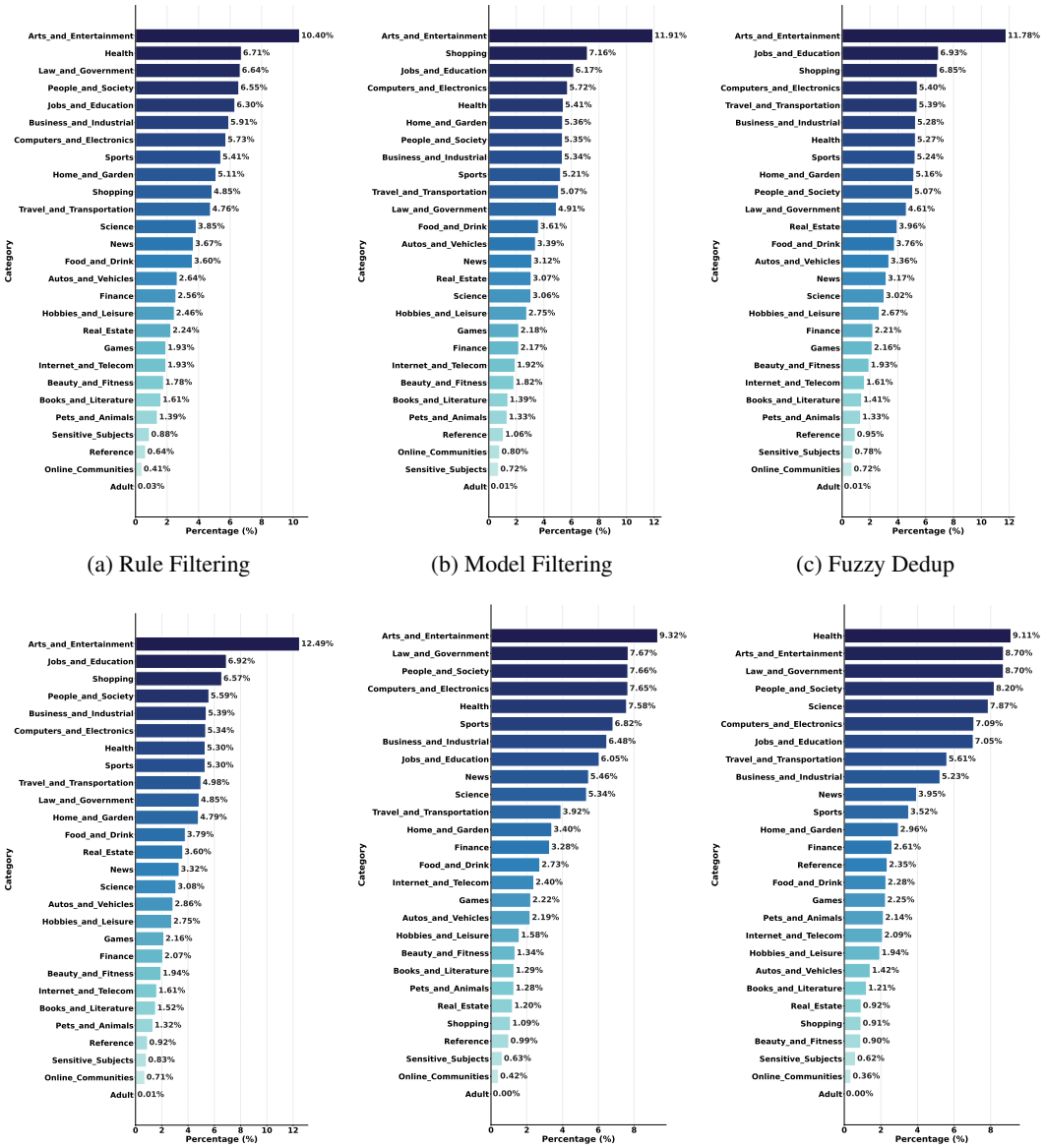
Algorithm 1 The Soft Deduplication Algorithm via Weighted Score Accumulation

Require: Embedding matrix $E \in \mathbb{R}^{N \times d}$ (L2-normalized), similarity interval weights $\mathcal{T} = \{(t_1, w_1), \ldots, (t_K, w_K)\}$, target score T_{target}, strict threshold t_{strict}, batch size B
Ensure: Binary array $\texttt{is_duplicate} \in \{0, 1\}^N$

$\texttt{is_processed} \leftarrow \mathbf{0}^N$
$\texttt{is_duplicate} \leftarrow \mathbf{0}^N$
$S \leftarrow \mathbf{0}^N$ ▷ Accumulated penalty scores
while $\exists\, i : \texttt{is_processed}[i] = 0$ **do**
  $U \leftarrow \{i \mid \texttt{is_processed}[i] = 0\}$ ▷ Unprocessed nodes
  $C \leftarrow U[:B]$ ▷ Current batch of candidates
  // Step 1: Intra-batch conflict resolution
  for each pair $i < j$ with $i, j \in C$ **do**
    if $\text{sim}(E_i,\, E_j) \geq t_{\text{strict}}$ **then**
      kill j within batch ▷ Retain earlier index as leader
    end if
  end for
  $\mathcal{L} \leftarrow \{i \in C \mid i \text{ not killed}\}$ ▷ Valid leaders
  // Step 2: Score accumulation
  for each leader $\ell \in \mathcal{L}$ **do**
    for each unprocessed target $u \in U,\ u \neq \ell$ **do**
      $w^* \leftarrow \max\{w_k \mid \text{sim}(E_\ell, E_u) \geq t_k,\ (t_k, w_k) \in \mathcal{T}\}$
      $S[u]\ +\!= w^*$
      if $S[u] \geq T_{\text{target}}$ **then**
        $\texttt{is_duplicate}[u] \leftarrow 1$
      end if
    end for
  end for
  $\texttt{is_processed}\big[C \cup \{u \mid \texttt{is_duplicate}[u] = 1\}\big] \leftarrow 1$
end while
return $\texttt{is_duplicate}$

F Domain Distribution Evolution Across the Pipeline

Figure 4 illustrates the evolution of the first-level (L1) domain distribution across the CuraWeb pipeline. Through rule filtering, model filtering, and deduplication (Figures 4a–4d), L1 entropy remains near 4.42 bits. It falls to 4.330 after value filtering and 4.295 after final sampling, reflecting the subsequent redistribution toward higher-value content. In contrast, the final Importance Sampling stage (Figure 4f) successfully drives a controlled distribution shift, significantly enriching the corpus with high-value, knowledge-intensive domains such as `Health`, `Science`, and `Jobs_and_Education` to enhance downstream model performance.



(a) Rule Filtering (b) Model Filtering (c) Fuzzy Dedup

(d) Semantic Dedup (e) Value Filtering (f) Importance Sampling

Figure 4: Evolution of first-level domain distribution across the pipeline.

G Quality-Signal Correlation Analysis

Figure 5 reports Pearson correlations among the eight non-binary quality scores. The signals are complementary but not statistically independent: writing and coherence correlate at 0.69, as do knowledge and reasoning, while writing and practical helpfulness correlate at 0.01. This variation supports retaining separate dimensions for structural filtering and content-value ranking rather than treating all scores as interchangeable. The binary safety score is omitted from this comparison.

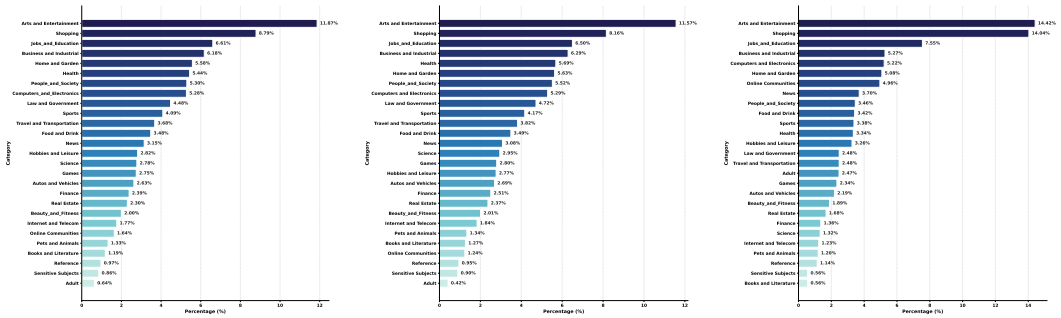


Figure 5: Pearson correlations among the eight non-binary quality dimensions. Safety is excluded.

H Detailed Domain Coverage Analysis

Data coverage evaluates whether the dataset encompasses the comprehensive knowledge scope required for foundational model training. Focusing on rule-based and model-based filtering—the two stages responsible for the highest data compression—we analyze how each filtering stage affects the L1-level domain distribution.

Impact of Rule-Based Filtering on Domain Distribution. Figures 6a and 6b present the L1-level domain distributions before and after rule-based filtering, while Figure 6c shows the domain composition of the filtered documents. The filtering rate is approximately 10%, and the resulting distributional perturbation is minimal: major categories such as `Arts`, `Shopping`, and `Online_Communities` exhibit only slight decreases of 0.2%–0.7%, and no L1-level domain is significantly compressed. Examining the filtered documents, the fastText templated-content classifier primarily targets advertisement-related content, whereas the Gopher rules remove documents relatively uniformly across domains, consistent with the high-recall design of the rule-based filtering stage.

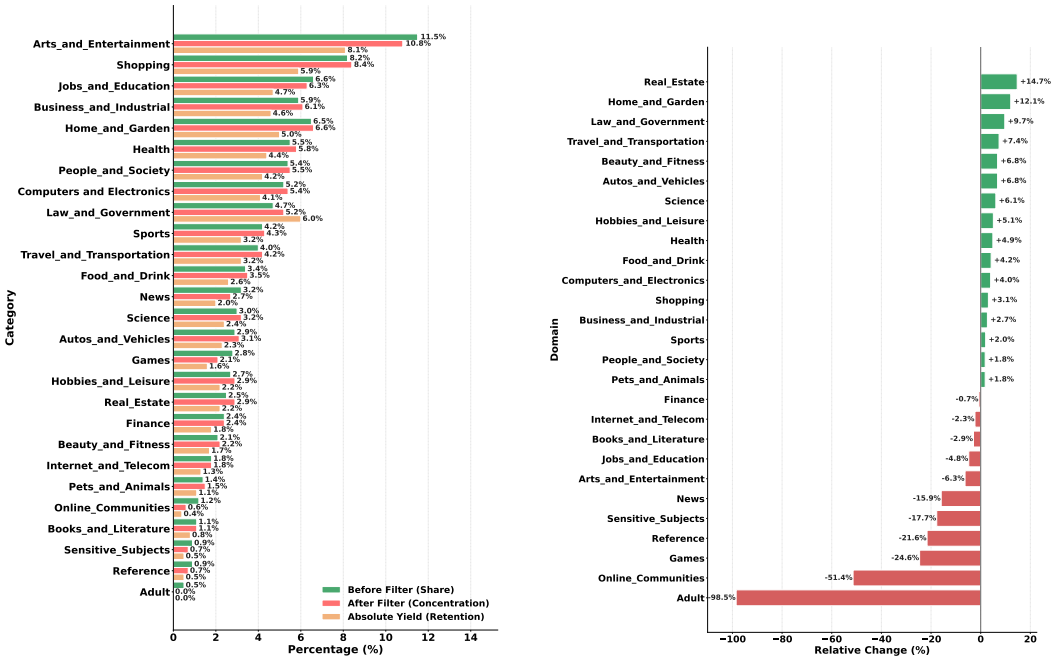


(a) Before filtering. (b) After filtering. (c) Removed documents.

Figure 6: Domain L1 distribution before and after rule-based filtering: (a) before filtering, (b) after filtering, and (c) removed documents.

Impact of Model-Based Filtering on Domain Distribution. Figure 7a presents the domain distributions across three stages: original, after rule-based filtering, and after model-based filtering. Model-based filtering removes an additional 24.8% of documents, exhibiting selective rather than uniform compression: `Adults`, `Online_Communities`, `Games`, and `Reference` are filtered at substantially higher rates than the overall average, while long-tail vertical domains such as `Religion` and `Apparel` are barely affected. The primary rejection criterion is failure in the format dimension—structural deficiencies such as garbled text, sentence fragmentation, or excessive noise—rather than issues related to factual correctness or opinion quality.

Comparison of Domain Coverage with Baseline Corpora. Under a controlled setting of 50 million documents, we compare the L1-level domain distributions of CuraWeb with four mainstream external datasets (see Figure 8). CuraWeb exhibits a relatively balanced distribution: its top three categories (`Health`, `Arts_and_Entertainment`, `Law_and_Government`) together account for only 26.5%. In contrast, FineWeb-Edu's top three categories alone reach 48.2%, reflecting its strong concentration on educational content at the cost of coverage in domains such as law, business, and entertainment. Dolma3 shows a technology and entertainment bias (`Computers_and_Electronics` and `Arts_and_Entertainment` together: 26.4%), while DCLM and Nemotron-CC exhibit similar distribution patterns dominated by `Arts_and_Entertainment`. Regarding long-tail domain retention, CuraWeb consistently achieves higher proportions in niche yet high-value areas such as `Science/Biological_Sciences`, `Finance`, and `Reference` compared with all external datasets.



(a) L1-Domain distribution across pipeline stages. (b) Relative change rate of L1-domain distribution.

Figure 7: Domain-level effects of model-based filtering: (a) L1-domain distribution across pipeline stages; (b) relative change rate of the domain L1 distribution after model-based filtering.

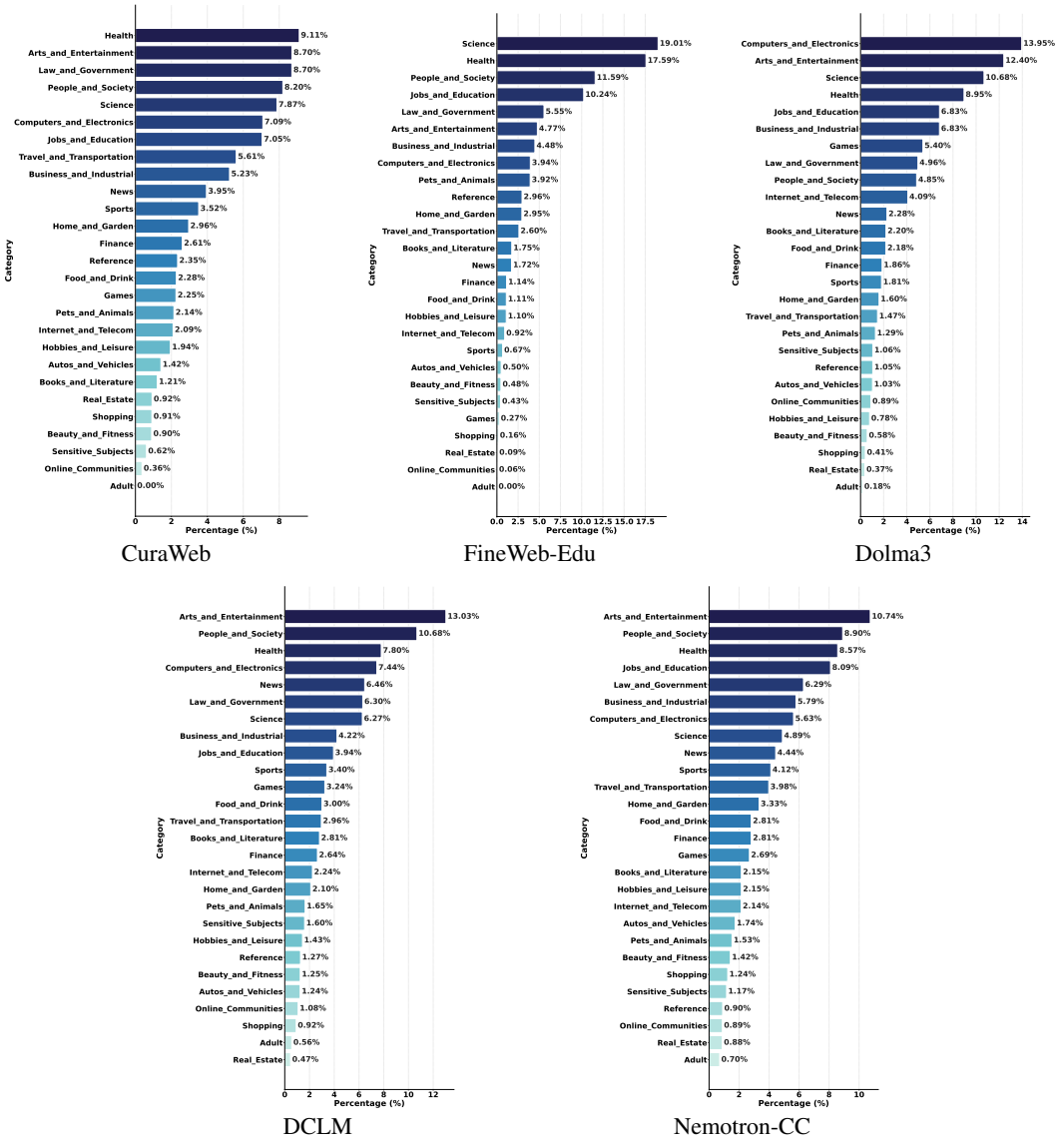


Figure 8: Top-27 L1-domain distributions across five datasets under equal-volume (50M documents) comparison. CuraWeb has a less concentrated distribution than FineWeb-Edu, which is heavily concentrated in Science (19.01%) and Health (17.59%).

I Detailed Redundancy and Lexical Diversity Analysis

We quantify the diversity of CuraWeb through deduplication counts, domain distribution entropy, and lexical diversity. Domain and lexical comparisons use 50M documents per corpus; deduplication counts describe the full pipeline.

Deduplication Redundancy. We validate the cascading benefits of our two-stage deduplication pipeline. The original corpus of 23.29B documents is reduced to 7.97B after MinHash fuzzy deduplication (compression ratio: 65.8%), and further to 6.42B after semantic deduplication (an additional reduction of 19.4%), yielding a cumulative deduplication rate of 72.4%.

Semantic deduplication proves particularly effective at capturing near-duplicates that evade MinHash-based fuzzy matching. Specifically, in mathematics and physics corpora, it identifies an additional 682.4K duplicates beyond fuzzy deduplication, bringing the total to 1.65M. Meanwhile, for utility-oriented websites, it expands the duplicate detection coverage from 2.33M to 3.93M documents. Notably, the average Jaccard similarity of the remaining pairs drops from 0.057% to 0.049%, confirming the successful elimination of low-similarity templated noise. These measurements show that semantic deduplication broadens the coverage of residual redundancy.

Domain Distribution Entropy. Table 2 reports the domain distribution entropy of CuraWeb and baseline corpora. CuraWeb achieves a L1-level domain entropy of 4.295 bits and a L2-level domain entropy of 6.075 bits, substantially higher than FineWeb-Edu (3.733 / 5.477 bits) and Dolma 3 (4.086 / 6.034 bits), while remaining comparable to DCLM (4.308 / 6.122 bits) and Nemotron-CC (4.380 / 6.148 bits). To understand how this entropy level is achieved and maintained throughout the pipeline, we track entropy dynamics across all stages. The results show that filtering and deduplication stages introduce negligible entropy loss: rule-filter (4.419) $\rightarrow$ model-filter (4.428) $\rightarrow$ fuzzy-dedup (4.427) $\rightarrow$ semantic-dedup (4.417) $\rightarrow$ value-filter (4.330) $\rightarrow$ final sampling (4.295). The cumulative entropy reduction of approximately 0.13 bits originates primarily from value filtering and final sampling, where a minor entropy reduction is traded for a higher proportion of high-value documents and improved downstream performance (Table 1).

Lexical Diversity. We measure lexical diversity using the unique n-gram ratio $\mathcal{R}_n$, defined as the ratio of unique n-grams to total n-grams. Table 10 reports $\mathcal{R}_n$ across pipeline stages and baselines. Within the CuraWeb pipeline, $\mathcal{R}_n$ increases monotonically from 0.000373 at the rule filtering stage to 0.000560 after semantic deduplication, demonstrating that each deduplication stage contributes positively to lexical diversity. In the final sampling stage, $\mathcal{R}_n$ decreases to 0.000352 due to the density dilution effect of power-law resampling, which shifts corpus volume toward high-value long-form documents and increases total n-gram count from 8.5B to 14.6B. Importantly, the number of unique n-grams increases to 5.13M—the highest among the pipeline stages—indicating that the sampled corpus provides broader lexical coverage than earlier pipeline stages. Compared with baselines, FineWeb-Edu achieves the lowest $\mathcal{R}_n$ (0.000213) and unique n-gram count (3.18M), consistent with its narrow domain distribution. Although Dolma3 achieves the largest unique n-gram count (5.67M), its overall quality score is substantially lower. When considering entropy, $\mathcal{R}_n$, and downstream performance jointly, CuraWeb consistently ranks within the top tier across all diversity metrics without sacrificing quality.

Table 10: Lexical diversity ($\mathcal{R}_n$) across pipeline stages and baselines (50M documents each).

Stage / Dataset	**Documents**	**Total n-grams**	**Unique n-grams**	$\mathcal{R}_n$
CuraWeb pipeline stages				
Rule Filter	50,000,000	12,548,809,607	4,677,970	0.000373
Model Filter	50,000,000	9,898,961,669	4,662,598	0.000471
Fuzzy Dedup	50,000,000	9,140,074,677	4,893,923	0.000535
Semantic Dedup	50,000,000	8,455,364,183	4,733,528	0.000560
Value Filter	50,000,000	12,872,152,074	4,815,215	0.000374
Final	50,000,000	14,573,357,160	5,129,070	0.000352
Baselines				
Nemotron-CC	50,000,000	11,659,696,700	4,101,843	0.000352
FineWeb-Edu	50,000,000	14,944,623,936	3,179,844	0.000213
Dolma3	50,000,000	17,486,244,342	5,665,350	0.000324
DCLM	50,000,000	15,554,026,736	4,677,300	0.000301

J Detailed Quality and Sampling Analysis

To systematically evaluate the quality of CuraWeb, we analyze our data curation pipeline from four perspectives: (1) **corpus quality assessment**, how the overall quality of our final corpus compares with existing open-source datasets across multiple quality dimensions; (2) **filtering precision**, whether our hybrid rule+model strategy accurately removes low-quality content without over-discarding high-value documents; (3) **high-value content retention**, how well our pipeline preserves information-dense documents across key STEM domains compared with existing corpora; and (4) **sampling-driven quality optimization**, how importance sampling further reshapes the corpus distribution toward higher knowledge density.

Corpus Quality Assessment. We evaluate overall corpus quality from two complementary angles. Within our pipeline, the overall quality score increases monotonically across stages: from 0.300 after fuzzy deduplication, through 0.302 after semantic deduplication and 0.436 after value filtering, finally reaching 0.521 after importance sampling—the highest among all internal pipeline variants, showing an increase across these later stages. Compared with external datasets, our final corpus (0.521) outperforms Dolma3 (0.457), DCLM (0.448), and Nemotron-CC (0.392). As shown in Table 11, CuraWeb ranks second in Knowledge (1.164, below FineWeb-Edu's 1.387) and first in PracticalHelpfulness (0.751). Although FineWeb-Edu achieves a higher overall score (0.610), this advantage stems from its strong concentration on educational content, at the cost of substantial domain diversity loss ($H_{\text{L1}} = 3.73$, the lowest among all compared datasets) and a notably lower PracticalHelpfulness score (0.562). Overall, CuraWeb achieves the most balanced quality profile across all dimensions.

Table 11: Normalized value scores of different corpora across value-oriented evaluation dimensions. The best results are indicated in bold, and the second-best results are underlined.

Dataset	EducationValue	Knowledge	PracticalHelpfulness	Reasoning	Composite_score
CuraWeb pipeline stages					
Rule Filter	1.110	0.675	0.438	0.512	0.380
Model Filter	0.894	0.400	0.293	0.322	0.313
Fuzzy Dedup	0.859	0.347	0.252	0.289	0.300
Semantic Dedup	0.846	0.352	0.247	0.305	0.302
Value Filter	1.249	0.942	0.488	0.797	0.436
Final	<u>1.498</u>	<u>1.164</u>	**0.751**	<u>0.987</u>	<u>0.521</u>
Baselines					
FineWeb-Edu	**1.963**	**1.387**	<u>0.562</u>	**1.202**	**0.610**
Dolma3	1.233	0.983	0.527	0.923	0.457
DCLM	1.230	0.887	0.486	0.936	0.448
Nemotron-CC	1.114	0.684	0.398	0.640	0.392

Filtering Precision. We compare two filtering strategies in terms of document retention and content quality. The pure rule-based strategy retains 34.18B documents, while the rule+model hybrid strategy further compresses the corpus to 23.29B documents, removing an additional 31.9% of documents. However, pure rule-based methods such as Gopher rely on surface-level textual features—including punctuation density, special character ratios, and line-length distributions—to assess quality, leading to systematic misclassification of STEM documents containing mathematical formulas, chemical symbols, or code snippets. In contrast, our rule+model hybrid strategy leverages model-based signals to identify semantically low-quality content, effectively distinguishing between "symbol-dense but high-quality" and genuinely low-quality documents. As shown in Table 12, the proportion of STEM documents increases from 19.74% at the rule filtering stage to 33.58% after importance sampling, while math-related documents rise from 1.656% to 3.832%, demonstrating that our hybrid strategy accurately removes low-quality content while actively protecting high-value professional documents.

High-Value Content Retention. Beyond filtering precision, we further evaluate how well our pipeline retains high-value documents across key STEM domains. As reported in Table 4, CuraWeb achieves an average retention rate of 0.432 for high-quality documents (score ≥ 2 under a 0–3 scoring scale) across Computer Science, Programming, Mathematics, and Physics, consistently outperforming DCLM (0.364), Dolma3 (0.376), and Nemotron-CC (0.261), while remaining comparable to FineWeb-Edu (0.427). Notably, although FineWeb-Edu achieves a slightly higher retention rate in Programming (0.451), its L1-domain entropy is substantially lower ($H_{\text{L1}} = 3.73$, the lowest among

all compared datasets). CURAWEB thus maintains competitive STEM retention rates while preserving a significantly more balanced domain distribution.

Table 12: Proportions of code, math, and STEM documents across pipeline stages.

Stage	Code Ratio	Math Ratio	STEM Ratio
Rule Filter	0.886%	1.656%	19.74%
Model Filter	0.801%	1.701%	20.38%
Fuzzy Dedup	0.915%	1.789%	25.11%
Semantic Dedup	0.847%	1.546%	19.49%
Value Filtering	1.600%	2.713%	30.46%
Final (w/ importance sampling)	**2.544%**	**3.832%**	**33.58%**

Sampling-Driven Quality Optimization. The importance sampling stage clusters documents and adjusts per-cluster sampling weights according to value signals, enabling targeted enhancement of high quality content. As shown in Figure 9, the most upsampled domains are concentrated in knowledge-intensive categories, including `Jobs_and_Education/Education` (6.53%), `Health/Health_Conditions` (5.18%), `Law_and_Government/Legal` (3.68%), and `Computers_and_Electronics/Programming` (2.59%), while low-information-density categories such as `Shopping` and `Real_Estate` experience substantial declines. This is further reflected in the overall domain distribution: after sampling, `Health` rises to the largest category (9.11%), and the proportions of `Science` (7.83%) and `Jobs_and_Education` (7.04%) increase significantly, whereas `Arts_and_Entertainment` drops from the top position (9.32%) before sampling. These structural shifts confirm that importance sampling effectively reallocates corpus weights toward domains with higher value density, consistent with our design objective. At the lexical level, the sampled corpus contains 5.13M unique n-grams and 14.6B total n-grams, representing a substantial scale-up compared to the semantic deduplication stage (4.73M unique n-grams and 8.5B total n-grams), while sustaining a high lexical diversity ($\mathcal{R}_n = 0.000352$). This trajectory shows increased lexical coverage alongside higher value scores, while the lower unique-to-total ratio reflects the larger n-gram denominator.

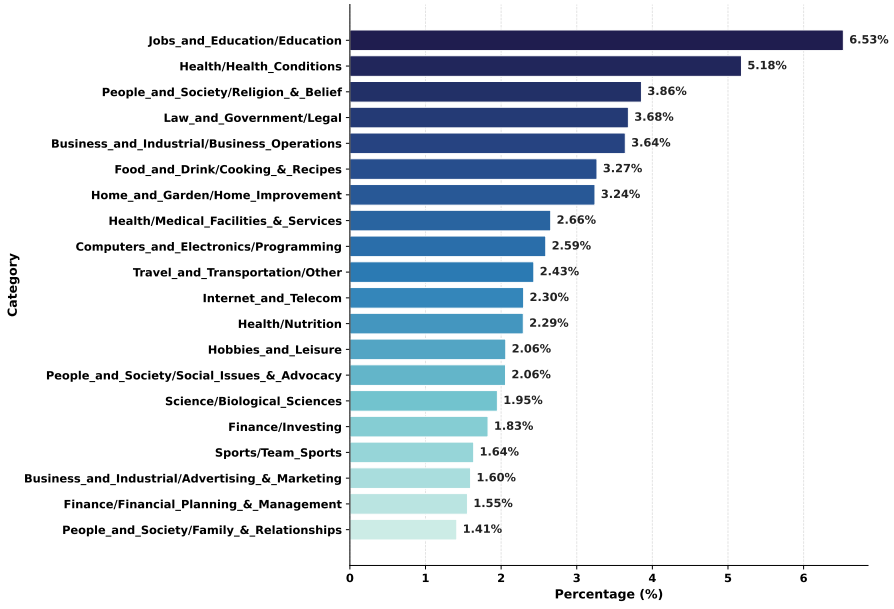


Figure 9: Top-20 domain categories by upsampling ratio in the CURAWEB importance sampling stage. Categories with high value density such as `Education`, `Health`, and `Programming` receive the highest upsampling weights.