Rillor

RESEARCH REPORT

Quality filtering for language model training data: heuristics and classifiers

How pretraining data filtering works, from heuristic rules and language identification to perplexity and quality classifiers, what dataset ablations measured and what each filter takes out.



13 November 2026 Data services rillor.com/insights/quality-filtering-for-language-model-training-data rillor.com

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Pretraining data filtering decides how much of a web crawl a language model learns from, and the published ablations show that its effect depends on the filter, the threshold, the source data and the benchmarks used to judge it.

- Most crawled text never reaches training. RefinedWeb's pipeline removed nearly 90% of the documents originally in Common Crawl [3], and DCLM-Baseline holds 1.4% of the documents in the pool it was built from [5].
- Heuristic filters are hand-written rules on length, symbols, punctuation and repetition. Removing C4's rules lowered scores on every task its authors tested [1], and three rules FineWeb derived from its own crawl statistics removed about 22% of tokens and raised its aggregate score by about 1% [4].
- Classifier filters score each document with a trained model. In DCLM's comparison at 1 billion parameters, a fastText classifier scored 30.2 on its core benchmarks, against 29.0 for perplexity filtering, which keeps text a reference language model finds predictable, and 27.5 for a heuristic baseline [5].
- FineWeb-Edu kept the pages that a classifier, built from 460,000 ratings by an instruction model of the Llama 3 family, scored 3 or more for educational value [4]. A 1.71-billion-parameter model trained on it rose from 33% to 37% on MMLU and from 46% to 57% on ARC against FineWeb [4].
- Filters remove text that teams may want. A word blocklist removed 42% of African American English documents from C4, against 6.2% of White-aligned English [8], and quality filtering raised both benchmark scores and toxic generation in 28 models of 1.5 billion parameters [10].
- Strict filters trade tokens for scores. FineWeb-Edu's threshold removed 92% of its source data [28], and Nemotron-CC's full 6.3-trillion-token dataset, built with a three-classifier ensemble, synthetic rephrasing and fewer heuristic filters, matched DCLM on MMLU with four times more unique real tokens [18].
- Perplexity filtering has no agreed direction: studies kept low, middle or high perplexity depending on the data [16, 17, 23].
- Small ablations are how the papers test a filter. Rankings from 150-million-parameter models predicted which corpus produced the better 1-billion-parameter model in about 80% of comparisons [22].

Background

What a filter does

A 2024 survey of data selection describes each method as a utility function, which maps a document to a number, and a selection mechanism, which decides from that number whether the document stays, with a threshold as its sensitivity [27]. Heuristic filters compute the number from rules applied to the text itself. Model-based filters compute it with a trained model: a classifier that estimates how close a document is to a reference corpus, a language model whose perplexity measures how predictable the text is, or a large model prompted to rate the document [27]. This report covers both kinds on web text. Deduplication and test-set removal are named where the pipelines place them and are not examined here.

Why web text is filtered

Common Crawl is the usual raw source. Each snapshot holds 20 to 30TB of uncompressed text from about 3 billion pages [7], and C4's authors found that "the majority of the resulting text is not natural language" [1]. The GPT-3 paper judged unfiltered or lightly filtered Common Crawl lower in quality than curated datasets, and its filtering cut 45TB of compressed plaintext to 570GB [6].

What quality means in these sources

No source reviewed defines quality apart from a filter. The Pretrainer's Guide states that its authors "do not have ground truth", and counts as quality or toxic any document that triggers the classifier [10]. Gururangan and colleagues note that the term is often ill-defined in the NLP literature [9]. This report therefore states what each filter removed and what models trained on the rest scored. It does not describe removed text as low in quality.

Method

What was examined. Published methods for pretraining data filtering on web text, and the effects measured by the teams that built them, in papers published from 2019 to 2025, with the 2016 fastText paper and current dataset cards, library code and toolkit documentation.

Sources. Thirty-five sources, each opened and viewed on 8 November 2026: dataset papers [1, 2, 3, 4, 5, 7, 11, 12, 18, 20, 24], the GPT-3 [6] and Llama 3 [21] papers, selection studies [13, 14, 15, 16, 17, 22, 23], audits of what filters remove [8, 9, 10, 19, 25], a survey [27], the fastText paper [26], dataset and model cards [28, 29, 30], library source code [31, 32] and toolkit documentation [33, 34, 35].

How. Papers were read in full text, except Nemotron-CC (see Limits), and each figure was taken from the passage or table cited, with the model size, token budget and benchmark that produced it. Results are not averaged across papers, since each trains its own models on its own data. Where sources give different figures for the same thing, both appear in Finding 9. No filter was run for this report.

Findings

1. Heuristic rules test length, symbols, punctuation and repetition

C4. Built from Common Crawl's April 2019 extracted text, C4 kept only lines ending in a period, exclamation mark, question mark or end quotation mark, and only lines of at least 5 words [1]. It discarded pages with fewer than 3 sentences, pages containing a word from a public list of obscene and offensive words, pages with "lorem ipsum" and pages with a curly bracket, and removed lines mentioning Javascript or cookie and privacy policies [1]. A language detector kept pages classified as English with a probability of at least 0.99 [1]. C4 against its unfiltered variant [1]:

Benchmark	C4 (745GB)	C4 unfiltered (6.1TB)
GLUE	83.28	81.46
SuperGLUE	71.36	68.04
SQuAD	80.88	78.78

The unfiltered variant still passed the language detector, so it kept some filtering [1].

Gopher. MassiveText's quality filter applies to MassiveWeb, the web text that made up 48% of the training data [2]. The other subsets received only content filtering, deduplication and test-set filtering [2]. A document is removed when it fails any rule below [2], and the datatrove library's Gopher filter uses the same defaults [31].

Rule	Document removed when
Length	Under 50 or over 100,000 words
Mean word length	Outside 3 to 10 characters
Symbols	More than 0.1 hash symbols or ellipses per word
Bullets	Over 90% of lines start with a bullet point
Ellipses	Over 30% of lines end with an ellipsis
Alphabetic words	Under 80% of words contain a letter

Rule	Document removed when
Stop words	Fewer than two of the, be, to, of, and, that, have, with

Its repetition rules remove documents where duplicate lines or paragraphs exceed 30% of the count or 20% of the characters, and where the most frequent 2-gram covers more than 20% of characters, with the limit falling step by step to 0.10 for duplicated 10-grams [2]. The authors chose simple rules to limit the potential for bias [2].

RefinedWeb. RefinedWeb applies Gopher's repetition and document rules, then line-wise corrections that remove lines such as social media counters and navigation buttons, and drops a document when corrections remove more than 5% of it [3]. After filtering, 23% of Common Crawl's documents remained [3].

FineWeb. FineWeb's base filtering combined a URL blocklist for adult content, English language identification at a score of 0.65 and the MassiveText rules at their original thresholds, leaving about 36 trillion tokens from 96 snapshots [4]. Tested on one 2019 crawl, C4's curly bracket and word-length rules removed 2.8% and 4.3% of tokens for small gains, the lorem ipsum, Javascript and policy rules each under 0.5%, and the terminal punctuation rule about 30% [4]. All of those except terminal punctuation beat that rule alone while removing about 7% in total, so FineWeb adopted them [4].

FineWeb then collected over 50 statistics on two versions of its 2013-48 snapshot, one deduplicated within the snapshot and one deduplicated across all snapshots, treated as higher and lower quality, and set thresholds where the lower-quality distribution was denser [4]. Of 16 candidate rules tested in 28-billion-token ablations, three stayed [4]:

FineWeb rule	Tokens removed
Fraction of lines ending in punctuation at or below 0.12	10.14%, against 30% for C4's rule
Fraction of characters in duplicated lines at or above 0.1	12.47%, where MassiveText's threshold is 0.2
Fraction of lines shorter than 30 characters at or above 0.67	3.73%

Together the three removed about 22% of tokens and raised the aggregate score by about 1%, above C4's rules while removing less [4]. Applied to 96 snapshots, the pipeline produced the 15-trillion-token FineWeb [4].

Dolma and the Llama 3 report. Citing arguments against model-based quality filters, Dolma used all of Gopher's rules plus C4's rule removing paragraphs without terminal punctuation [11]. Gopher's rules tagged 15.23% of UTF-8 characters for removal and the punctuation rule 22.73% [11]. On 1.2-billion-parameter models, the punctuation rule alone outscored C4's full set and Gopher's full set, and the combination scored

higher still [11]. The Llama 3 report adds a duplicated n-gram coverage ratio that catches repeated logging or error messages, counts of obscene words for adult sites missing from blocklists, and a token-distribution divergence that flags documents with many outlier tokens [21].

2. Language identification makes the first large cut

CCNet's fastText language identifier, trained on Wikipedia, Tatoeba and SETimes for 176 languages, assigns a language when its score exceeds 0.5 and discards the page otherwise [7]. The published model files are 126MB, or 917kB compressed [33].

The cut is large. At a threshold of 0.65, RefinedWeb discarded about half of its documents for not being English, before quality rules removed 24% of the rest and deduplication 12% [3]. Dolma's English threshold of 0.5 removed 61.7% of the data by bytes across 25 snapshots, and the CCNet pipeline as a whole removed 84.2% of the content, from 175.1TB to 27.7TB [11]. FineWeb2 found that about 40% of documents met FineWeb's English threshold and built its multilingual corpus from the other 60% [20].

Thresholds vary. mC4 keeps scores above 70% and CC-100 uses 50% [20], Nemotron-CC combines pycld2 with fastText at 0.3 [18], and the Llama 3 report sorts its multilingual documents into 176 languages with a fastText-based model [21]. FineWeb2 found that Arabic and Russian did better above 0.8 and Swahili around 0.3, a removal rate of almost 65%, and set each language's threshold at the median of its scores minus one standard deviation, bounded between 0.3 and 0.9 [20].

Mislabeling is common in multilingual crawls. An audit of 205 language-specific corpora found at least 15 with no usable text and many mislabeled [19]. Filtering noisy corpora with a Transformer language identifier raised median precision from 13.8% to 43.9% at a 77.5% loss in recall [19]. For Lingala precision rose from 8% to 80% and for Oromo from 2% to 33%, each losing half of its correct in-language sentences [19].

3. Perplexity filtering: which end to keep depends on the data

CCNet trained 5-gram language models on Wikipedia for 48 languages and split each of those languages' pages into head, middle and tail by perplexity [7]. The authors found perplexity a relatively good proxy for quality. Keyword lists scored high perplexity, but so did valid text far from Wikipedia, such as blog comments in a spoken style and specialized forums [7]. They kept every bucket and set thresholds per language, because perplexity distributions differ [7]. Word vectors trained on English pages scored 62.0 from the tail, 74.2 from the middle and 77.9 from the head [7].

Later corpora reuse the buckets. Dolma read them as 21.9% high, 28.5% medium and 49.6% low quality, and its heuristic rules did not change these proportions, suggesting that model-based filters capture other signals [11].

Studies that prune by perplexity disagree on direction.

Study	Reference model	Text kept	Result
Muennighoff et al. [23]	Trained on Wikipedia	Lowest-perplexity 25%	Effective on an 84B-token budget. On C4, effective only at 4.2B parameters
Marion et al. [16]	124M trained on Common Crawl or Wikipedia, 6B to 52B on web-scale data	Middle of the range	Keeping 30% or 50% beat keeping 70% with the better-performing reference models
Ankner et al. [17]	125M	High (Pile) or medium (Dolma), 50% kept	3B model up to 2.04 points higher, up to 1.45x fewer steps
DCLM [5]	154M, Wikipedia, books and peS2o	Low perplexity	29.0 CORE at 1B, against 27.5 for the heuristic baseline

Marion and colleagues found that perplexity outperformed two more expensive scores, error L2-norm and memorization [16]. Ankner and colleagues report that pruning is sensitive to domain mix, and the Pile holds only 15.61% general web data [17]. QuRating's authors note that low-perplexity data includes simple, repetitive text that is easy to predict [14], and Gopher's authors left likelihood ranking as future work, because high-likelihood samples "are not necessarily high quality" [2].

4. Classifier filters: the reference data defines quality

Reference-corpus classifiers. The GPT-3 paper's classifier was a logistic regression on hashed token features, trained with WebText, Wikipedia and a web books corpus as positives and unfiltered Common Crawl as negatives [6]. A document was kept when a Pareto draw with alpha 9 exceeded one minus its score, which kept mostly high-scoring documents along with some out of distribution [6]. The Pile repeated the method with a fastText bigram classifier, choosing OpenWebText2 as positives over the whole Pile [24]. The PaLM and GLaM model families used a quality classifier with Wikipedia, books and a few selected websites as references, and the LLaMA family used pages cited as references in Wikipedia [10]. The Llama 3 report trained fastText classifiers to find text that would be referenced by Wikipedia, and Roberta-based classifiers on quality predictions from a chat model of the Llama 2 family, with DistilRoberta producing document scores for efficiency [21].

DSIR's authors call this approach heuristic classification and note that it does not guarantee data distributed like formal text [15]. DSIR instead resamples with importance weights over hashed n-gram features. It selected 100 million documents from the Pile in 4.5 hours, and its distribution measure correlated with downstream accuracy at r = 0.82 across 8 methods [15].

Classifiers trained on model ratings. FineWeb-Edu sampled 460,000 pages from one 2024 snapshot and had an instruction model of the Llama 3 family rate educational value from 0 to 5, focused on grade-school and middle-school knowledge so that highly technical pages would not be favored [4]. A regression head on a small frozen embedding model, trained on 410,000 ratings and validated on 50,000, reached F1 of 82% at a threshold of 3 [4]. Scoring all 15 trillion FineWeb tokens took 6,000 H100 GPU hours [4]. The classifier's model card reports, on 46,867 held-out samples, overall accuracy of 0.71 and macro F1 of 0.50, with recall of 0.35 for class 4 and 0.01 for class 5 [29]. A jury of three annotator models shifted scores upward, and a classifier trained on jury ratings performed worse [28].

QuRating used pairwise judgments from a model of the GPT-3.5 family on writing style, required expertise, facts and trivia, and educational value, turned into ratings with a Bradley-Terry model, and annotated a 260-billion-token corpus [14]. Ask-LLM asks an instruction-tuned model whether a paragraph is informative for pretraining and uses the probability of "yes" as the score [13].

DCLM's comparison. DCLM trained its classifiers on about 400,000 documents, half positive, with negatives drawn at random from its RefinedWeb reproduction [5]. DCLM's Table 4 compares the methods at 1 billion parameters [5]:

Method, 1B parameters	CORE	EXTENDED
RefinedWeb reproduction (heuristics)	27.5	14.6
Top 20% by PageRank	26.1	12.9
SemDedup	27.1	13.8
Classifier on BGE embeddings	27.2	14.0
AskLLM	28.6	14.3
Perplexity filtering	29.0	15.0
Top-k average logits	29.2	14.7
fastText, OH-2.5 and ELI5	30.2	15.4

At 7 billion parameters, keeping the top 10% by fastText score, the OH-2.5 and ELI5 reference data reached 41.0 CORE, against 35.7 for Wikipedia, 34.7 for OpenWebText2 and 37.5 for an approximation of the GPT-3 paper's references [5].

Ensembles. Nemotron-CC found that the FineWeb-Edu and DCLM classifiers recall only around 10% of high-quality tokens [18]. It combined two classifiers trained on educational scores from Mixtral-family and Nemotron-family instruction models with the DCLM classifier, mapped each score to 20 buckets, took the maximum and grouped the buckets into five quality levels by testing them in continued pretraining [18].

Cost. fastText trains on more than a billion words in under ten minutes on a standard multicore CPU [26], and the survey notes that a classifier at this scale should be lightweight [27]. Another curation toolkit's classifiers, including a DeBERTa quality classifier that labels documents low, medium or high, require GPUs [35]. RedPajama-V2 publishes raw, unfiltered data with 46 quality measures computed for each document, so that rules such as C4's become special cases a user applies later [12].

5. Measured gains, under each paper's conditions

Source	Filter tested	Model and data	Result as reported
Gopher [2]	Heuristic quality rules	1.4B, 5GB subsamples	Gains across the board on three validation sets
RefinedWeb [3]	Gopher rules, line corrections	Zero-shot aggregate	52.7 to 54.3, and 56.2 with deduplication
FineWeb [4]	Three new rules	1.71B, 28B tokens	Aggregate up about 1%
FineWeb-Edu [4]	Classifier, threshold 3	1.71B, 350B tokens	MMLU 33% to 37%, ARC 46% to 57%
DCLM [5]	fastText, top 10%	7B, 2.6T tokens	64% MMLU, 6.6 points above MAP-Neo with 40% less compute
Nemotron-CC [18]	Ensemble plus synthetic rephrasing, high-quality subset	8B, 1T tokens	MMLU 5.6 points above DCLM
QuRating [14]	Educational value ratings	1.3B, 30B tokens	In-context learning 1.8% above uniform selection
Ask-LLM [13]	Ratings from a Flan-T5-XL model	T5-Large (800M) on C4	60% of the data trained better and 70% faster than all of it

The gains vary by source data. With filtering alone, RefinedWeb's rules moved OSCAR-21.09 by 0.4 points, OSCAR-22.01 by minus 0.4, C4 by 0.5, The Pile by 0.8 and RefinedWeb's own raw data by 1.6, while deduplication alone, tested on the other four, brought a steady gain on each [3]. On C4, a perplexity filter helped only at 4.2 billion parameters, and the ROOTS filter on OSCAR did no better than unfiltered OSCAR on that study's benchmark [23]. The Pretrainer's Guide concludes that filtering benefits are not predictable from text characteristics [10].

6. What filters remove besides their target

Blocklists. C4's blocklist removed text from and about minority individuals at higher rates [8]. Of 50 clusters of excluded documents, only 16 were largely sexual, and others covered science, medicine and health, and legal and political documents [8]. Of 50 excluded documents checked for each term, 22% for "lesbian" and 36% for "gay" were non-offensive or non-sexual [8]. Removal rates were 42% for African American English, 32% for Hispanic-aligned English, 6.2% for White-aligned English and 7.2% for other English, and 97.8% of the documents left are White-aligned [8]. Gopher used a safe-search filter instead of word lists for this reason, citing that study [2].

Classifiers. Replicating the GPT-3 filter on U.S. high school newspapers, Gururangan and colleagues found it more likely to rate papers from larger schools in wealthier, more educated and urban ZIP codes as high quality, and its scores unaligned with factuality or literary acclaim [9]. QuRating's authors document topical, social-role and geographic biases in their ratings, and a bias toward English [14]. Nemotron-CC found that heuristic filters removed 18.1% of the tokens the FineWeb-Edu classifier rated high quality, and applied heuristics only to its lower-quality splits [18].

Toxicity. The Pretrainer's Guide found that removing toxic data trades fewer toxic generations for reduced generalization, and that quality filtering raises both toxic generation and downstream performance [10]. Its quality filter improved toxicity identification and most question answering despite removing 10% or more of the data [10]. Question answering peaked at its mildest threshold, while toxicity identification still improved after 55% of the data was gone [10]. An earlier study found that basic toxicity interventions reduce coverage of texts about, and dialects of, marginalized groups, and that after strong toxicity reduction human raters often disagree with high automatic toxicity scores [25]. Dolma's strict toxicity threshold removed 29.1% to 34.9% of content and generally scored above its permissive threshold, which removed 5.5% to 7.3%, yet Dolma chose the permissive one to meet its token target [11]. The FineWeb-Edu team avoided both toxicity classifiers and similarity to a gold source such as Wikipedia, citing dialect removal and over-classification of identity terms [28].

Personal information. Dolma replaced email addresses, IP addresses and phone numbers in documents with 5 or fewer such spans (0.02% of documents) and removed documents with more (0.001%), and removal against replacement had no effect on model performance in its ablations [11].

7. How much to filter

FineWeb-Edu's threshold ablation favored 3 on average [4]. Its dataset card reports that this threshold removed 92% of the data and leaves 1.3 trillion tokens [28]. A threshold of 2 leaves 5.4 trillion tokens and still beats FineWeb, and a higher threshold improves knowledge and reasoning benchmarks while lowering HellaSwag and PIQA [28]. DCLM's 7-billion-parameter runs scored 41.0 CORE when keeping the top 10%, 39.8 at 15% and 38.7 at 20% [5]. DCLM-Baseline holds 3.8 trillion tokens from a 240-trillion-token pool [5], and its card describes it as intended for research use only, not for production models or for domains such as code and math [30].

Nemotron-CC argues that removing 90% of the data limits long-horizon training, which reached 15 trillion tokens for the Llama 3.1 family [18]. The paper's Table 2 sorts its tokens into five levels [18]:

Nemotron-CC level	Tokens	Share
High	553B	12.63%
Medium-High	504B	11.52%
Medium	2,023B	46.24%
Medium-Low	894B	20.43%
Low	402B	9.18%

Repeating data is the alternative to loosening filters. Up to 4 epochs of repeated data changed loss negligibly against unique data, in runs up to 9 billion parameters and 900 billion tokens [23]. The same study recommends reserving filtering for noisy datasets [23]. Ask-LLM reports gains over full-data training even when rejecting 90% of the data [13]. QuRating found that selecting only the top-rated documents produced models that excel at some tasks and underperform on others, and that sampling at temperature 2.0 gave more consistent results [14].

8. Testing and recording a filter

Small ablations. FineWeb trained two 1.71-billion-parameter models per dataset version, identical apart from their data, on different random subsets and seeds, and averaged their scores [4]. Filtering ablations ran on about 28 billion tokens, roughly compute-optimal for that size, and all of FineWeb's ablations together took more than 70 models and an estimated 80,000 H100 GPU hours [4]. Benchmarks were kept when they showed low variance between samples of the same data, rose steadily during training and scored above

random [4]. DataDecide trained models up to 1 billion parameters on 25 corpora with three seeds, and found that no scaling-law method among 8 baselines beat single-scale predictions on the compute-decision frontier [22].

Human judgment. In DCLM, 16 annotators labeled about 500 documents three times each, with 71% average agreement [5]. The fastText filters scored about 73% ROC-AUC against the majority labels [5]. AskLLM scored about 82%, yet produced weaker data, about 28.5 CORE against more than 31 for several fastText classifiers [5].

Decontamination. Removing detected MMLU and HellaSwag overlaps from DCLM-Baseline did not lower model performance [5], while the C4 audit found benchmark data in C4's English set [8].

Records. In the datatrove library, a filter returns either a keep decision or a drop with a named reason, such as a document too short, the run counts dropped documents per reason, and an optional writer saves the dropped documents [31, 32]. Another curation toolkit adds score columns without removing documents, so thresholds can be set after examining score distributions, and its default logs report aggregate counts without the rejection reason for each record [34]. Its classifiers can store prediction scores as metadata [35].

9. Where the sources disagree

1. **Heuristics or classifiers.** DCLM finds model-based filtering "key to assembling a high-quality training set" [5]. Dolma used heuristics only, and its results suggest the two kinds capture different signals [11]. Nemotron-CC stopped applying heuristics to its high-quality split [18].
2. **Perplexity direction.** DCLM's perplexity filter [5] and Muennighoff and colleagues [23] keep low perplexity, while CCNet splits pages by perplexity and keeps every bucket [7]. Marion and colleagues favor the middle [16], and Ankner and colleagues keep high or medium perplexity by dataset [17].
3. **Human agreement and data value.** AskLLM matched human labels closely but produced weaker data in DCLM [5], while Ask-LLM's own paper reports gains over full data on C4 with T5-family models [13]. The setups differ.
4. **Toxicity filters.** Dolma applied one at a permissive threshold [11]. The FineWeb-Edu team applied none [28].
5. **FineWeb-Edu's annotation count.** The paper says 460,000 [4], the dataset card 500,000 [28] and the classifier card 450,000 for training plus a 46,867-sample hold-out [29].
6. **DCLM-Baseline's size.** The paper says 3.8 trillion tokens [5], the dataset card 4 trillion tokens in 3 billion documents [30].
7. **QuRating's rater.** QuRating names a GPT-3.5-family model [14], while the survey names a GPT-4-family model [27]. This report follows QuRating.

Implications

For teams building a pretraining corpus. Text extraction and language identification set much of the result before any quality rule runs [3, 5, 11]. Published thresholds were tuned in ablations on English web text at one model scale, such as FineWeb's 1.71 billion parameters [4], and FineWeb2 needed per-language thresholds [20]. FineWeb and DCLM chose their thresholds by training small models on differently filtered data and comparing benchmark scores [4, 5], and Nemotron-CC grouped its quality buckets by continued pretraining of a partly trained 8-billion-parameter model [18].

For teams licensing or buying filtered data. A filtered dataset carries its filters' choices, including what they removed from dialects and topics [8]. The FineWeb-Edu card notes that code is likely not prevalent and that specialized sources may need adding [28], and DCLM-Baseline's card limits it to research use [30]. Per-rule removal counts and stored scores, which datatrove and RedPajama-V2 produce, show what was dropped and allow filtering again [12, 31, 32].

For evaluation leads. The C4 audit found benchmark data in C4 [8], and DCLM found no score loss from removing detected overlaps [5].

Rillor's training data cleaning and filtering filters out records outside a client's scope, by source, date, language or type, to the rules the client sets, and the results of its checks ship with every version of the cleaned data.

Limits of this analysis

- The filter ablations reviewed trained models of up to 8 billion parameters [5, 18]. None measures filters on larger models.
- Most results are for English. Multilingual evidence comes from FineWeb2 [20] and one audit [19].
- Each paper uses its own benchmarks and token budgets, so scores across rows are not comparable.
- The bias audits cover C4's blocklist [8], a GPT-3-style classifier [9] and a PaLM-family quality classifier with a toxicity scorer [10]. The FineWeb paper compares gender, age and religion terms in FineWeb and FineWeb-Edu and finds the biases it examined not strong [4], but none of the sources reviewed measures what the FineWeb-Edu or DCLM classifiers remove by dialect or identity.
- Compute costs are limited to the figures the sources give [4, 15, 26]. No dollar costs are stated.
- For Nemotron-CC, only the abstract, Sections 2.1 and 2.2 and Tables 1, 2 and 4 were read [18].
- No filter was run and no dataset was measured for this report.

Sources

1. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Colin Raffel, Noam Shazeer, Adam Roberts and others, Journal of Machine Learning Research 21, arXiv:1910.10683, version 4, 19 September 2023, viewed 8 November 2026. https://arxiv.org/abs/1910.10683

2. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. Jack W. Rae, Sebastian Borgeaud, Trevor Cai and others, arXiv:2112.11446, version 2, 21 January 2022, viewed 8 November 2026. https://arxiv.org/abs/2112.11446

3. The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only. Guilherme Penedo, Quentin Malartic, Daniel Hesslow and others, arXiv:2306.01116, 1 June 2023, viewed 8 November 2026. https://arxiv.org/abs/2306.01116

4. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal and others, NeurIPS 2024 Datasets and Benchmarks Track, arXiv:2406.17557, version 2, 31 October 2024, viewed 8 November 2026. https://arxiv.org/abs/2406.17557

5. DataComp-LM: In search of the next generation of training sets for language models. Jeffrey Li, Alex Fang, Georgios Smyrnis and others, arXiv:2406.11794, version 4, 21 April 2025, viewed 8 November 2026. https://arxiv.org/abs/2406.11794

6. Language Models are Few-Shot Learners. Tom B. Brown, Benjamin Mann, Nick Ryder and others, NeurIPS 2020, arXiv:2005.14165, version 4, 22 July 2020, viewed 8 November 2026. https://arxiv.org/abs/2005.14165

7. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau and others, arXiv:1911.00359, version 2, 15 November 2019, viewed 8 November 2026. https://arxiv.org/abs/1911.00359

8. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. Jesse Dodge, Maarten Sap, Ana Marasović and others, EMNLP 2021, arXiv:2104.08758, version 2, 30 September 2021, viewed 8 November 2026. https://arxiv.org/abs/2104.08758

9. Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection. Suchin Gururangan, Dallas Card, Sarah K. Dreier and others, arXiv:2201.10474, version 2, 26 January 2022, viewed 8 November 2026. https://arxiv.org/abs/2201.10474

10. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. Shayne Longpre, Gregory Yauney, Emily Reif and others, arXiv:2305.13169, version 2, 13 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2305.13169

11. Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. Luca Soldaini, Rodney Kinney, Akshita Bhagia and others, ACL 2024, arXiv:2402.00159, version 2, 6 June 2024, viewed 8 November 2026. https://arxiv.org/abs/2402.00159

12. RedPajama: an Open Dataset for Training Large Language Models. Maurice Weber, Daniel Fu, Quentin Anthony and others, NeurIPS 2024 Datasets and Benchmarks Track, arXiv:2411.12372, 19 November 2024, viewed 8 November 2026. https://arxiv.org/abs/2411.12372

13. How to Train Data-Efficient LLMs. Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang and others, arXiv:2402.09668, 15 February 2024, viewed 8 November 2026. https://arxiv.org/abs/2402.09668

14. QuRating: Selecting High-Quality Data for Training Language Models. Alexander Wettig, Aatmik Gupta, Saumya Malik and Danqi Chen, ICML 2024, arXiv:2402.09739, version 3, 17 July 2024, viewed 8 November 2026. https://arxiv.org/abs/2402.09739

15. Data Selection for Language Models via Importance Resampling. Sang Michael Xie, Shibani Santurkar, Tengyu Ma and Percy Liang, NeurIPS 2023, arXiv:2302.03169, version 3, 18 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2302.03169

16. When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale. Max Marion, Ahmet Üstün, Luiza Pozzobon and others, arXiv:2309.04564, 8 September 2023, viewed 8 November 2026. https://arxiv.org/abs/2309.04564

17. Perplexed by Perplexity: Perplexity-Based Data Pruning With Small Reference Models. Zachary Ankner, Cody Blakeney, Kartik Sreenivasan and others, arXiv:2405.20541, 30 May 2024, viewed 8 November 2026. https://arxiv.org/abs/2405.20541

18. Nemotron-CC: Transforming Common Crawl into a Refined Long-Horizon Pretraining Dataset. Dan Su, Kezhi Kong, Ying Lin and others, ACL 2025, arXiv:2412.02595, version 2, 30 May 2025, viewed 8 November 2026. https://arxiv.org/abs/2412.02595

19. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. Julia Kreutzer, Isaac Caswell, Lisa Wang and others, Transactions of the Association for Computational Linguistics 10, arXiv:2103.12028, version 4, 21 February 2022, viewed 8 November 2026. https://arxiv.org/abs/2103.12028

20. FineWeb2: One Pipeline to Scale Them All: Adapting Pre-Training Data Processing to Every Language. Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec and others, arXiv:2506.20920, 26 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2506.20920

21. The Llama 3 Herd of Models. Aaron Grattafiori, Abhimanyu Dubey and others, arXiv:2407.21783, version 3, 23 November 2024, viewed 8 November 2026. https://arxiv.org/abs/2407.21783

22. DataDecide: How to Predict Best Pretraining Data with Small Experiments. Ian Magnusson, Nguyen Tai, Ben Bogin and others, ICML 2025, arXiv:2504.11393, version 2, 13 July 2025, viewed 8 November 2026. https://arxiv.org/abs/2504.11393

23. Scaling Data-Constrained Language Models. Niklas Muennighoff, Alexander M. Rush, Boaz Barak and others, NeurIPS 2023, arXiv:2305.16264, version 5, 28 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2305.16264

24. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. Leo Gao, Stella Biderman, Sid Black and others, arXiv:2101.00027, 31 December 2020, viewed 8 November 2026. https://arxiv.org/abs/2101.00027

25. Challenges in Detoxifying Language Models. Johannes Welbl, Amelia Glaese, Jonathan Uesato and others, Findings of EMNLP 2021, arXiv:2109.07445, 15 September 2021, viewed 8 November 2026. https://arxiv.org/abs/2109.07445

26. Bag of Tricks for Efficient Text Classification. Armand Joulin, Edouard Grave, Piotr Bojanowski and Tomas Mikolov, arXiv:1607.01759, version 3, 9 August 2016, viewed 8 November 2026. https://arxiv.org/abs/1607.01759

27. A Survey on Data Selection for Language Models. Alon Albalak, Yanai Elazar, Sang Michael Xie and others, Transactions on Machine Learning Research, arXiv:2402.16827, version 3, 2 August 2024, viewed 8 November 2026. https://arxiv.org/abs/2402.16827

28. FineWeb-Edu dataset card. Hugging Face, HuggingFaceFW/fineweb-edu, last modified 11 July 2025, viewed 8 November 2026. https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu

29. FineWeb-Edu classifier model card. Hugging Face, HuggingFaceFW/fineweb-edu-classifier, last modified 17 November 2024, viewed 8 November 2026. https://huggingface.co/HuggingFaceFW/fineweb-edu-classifier

30. DCLM-baseline-1.0 dataset card. DCLM Team, Hugging Face, mlfoundations/dclm-baseline-1.0, last modified 22 July 2024, viewed 8 November 2026. https://huggingface.co/datasets/mlfoundations/dclm-baseline-1.0

31. gopher_quality_filter.py. Hugging Face, datatrove source code on GitHub, file last changed 8 August 2025, viewed 8 November 2026. https://github.com/huggingface/datatrove/blob/main/src/datatrove/pipeline/filters/gopher_quality_filter.py

32. base_filter.py. Hugging Face, datatrove source code on GitHub, file last changed 16 March 2026, viewed 8 November 2026. https://github.com/huggingface/datatrove/blob/main/src/datatrove/pipeline/filters/base_filter.py

33. Language identification. fastText documentation, Facebook Inc., undated, viewed 8 November 2026. https://fasttext.cc/docs/en/language-identification.html

34. Heuristic Filtering. NeMo Curator documentation, NVIDIA, version v1.4.0, viewed 8 November 2026. https://docs.nvidia.com/nemo/curator/latest/curate-text/process-data/quality-assessment/heuristic.html

35. Distributed Data Classification. NeMo Curator documentation, NVIDIA, version v1.4.0, viewed 8 November 2026. https://docs.nvidia.com/nemo/curator/latest/curate-text/process-data/quality-assessment/distributed-classifier.html

About this report

Published by Rillor on 13 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Quality filtering for language model training data: heuristics and classifiers. 13 November 2026. https://rillor.com/insights/quality-filtering-for-language-model-training-data

Notices

NVIDIA, AMD, Ceph, Kubernetes and related product names are trademarks of their owners and are used only to identify products. No affiliation or endorsement is implied.

Work with Rillor

Send the question the data should answer, the sources you know of, the coverage you need and how often it should update. Contact us at rillor.com/contact?type=custom-data or contact@rillor.com. More on data services: rillor.com/data