Rillor

EXPLAINER

Preparing data for language model training: a buyer's map from source to delivery

Data preparation for LLM training, stage by stage: extraction, filters, deduplication, personal data, decontamination, mixing and what a buyer should receive.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

The published pipelines of open corpora and open model families show what data preparation for LLM training removes at each stage, and what record of each removal a buyer can ask for.

- Most of a web crawl is removed before training. RefinedWeb's stages removed nearly 90% of the documents originally in Common Crawl [2], and DCLM-Baseline kept 1.4% of its original pool's documents [6].
- Extraction changes results. Text pulled from raw crawl files with trafilatura trained a better model than Common Crawl's plain-text files in FineWeb's tests [1], and two extractors beat those files by at least 2.5 points on DCLM's core evaluation score [6].
- Thresholds are each pipeline's own settings. Language cut-offs in four published pipelines run from 0.5 in Dolma to 0.99 in C4 [1, 2, 4, 5], and DCLM's quality classifier kept the top 10% of documents [6].
- Filters have side effects. C4's bad-words blocklist removed African-American English text at 42%, against 6.2% for white-aligned English [9], and in one study quality filtering raised toxic generation along with task performance [11].
- Deduplication runs at the URL, document, paragraph or line level [5, 7]. Dolma removed 53.2% of documents as URL duplicates [5], and FineWeb's test models gained little from deduplicating across all 96 Common Crawl snapshots at once [1].
- Published handling of personal data is narrow. Dolma masked email addresses, IP addresses and phone numbers in 0.02% of documents and removed the 0.001% with denser personal data, and the choice between masking and removal had no effect on model performance [5].
- Decontamination can fail. The paper that introduced a 175-billion-parameter model reported a filtering bug that ignored some overlaps with its benchmarks [8], and DCLM asks every submission for a decontamination report [6].
- Disclosure rules ask about data preparation. California's AB 2013 asks developers whether datasets were cleaned or modified and why [27], and the EU AI Act lists cleaning among data governance practices for "high-risk AI systems" [25].

Background

What data preparation covers

Data preparation turns a source into the files a training run reads. The open web corpora in this report are built from Common Crawl [1, 2, 4, 5, 6]. As of February 2024 it held over 250 billion pages crawled since 2007, in 97 snapshots [5]. The crawl is published as WARC files, "the raw data from the crawl", and as WET files, which "only contain extracted plaintext" [29]. The paper that introduced C4 put the extracted text at around 20 TB a month, and its cleaned corpus from the April 2019 crawl came to about 750 GB [4].

The gap between those figures is the work this report describes. The MassiveText authors wrote that "The vast majority of text found on the web is of insufficient quality to be useful for language model training" [3]. The Dolma authors assumed "at least a 10x reduction from the sum of all data cleaning efforts", and measured their toolkit's filtering at 122 CPU hours per TB [5].

Why the record matters

Each stage removes data by a rule someone chose. The Dolma authors write that "commercial models rarely detail their data, and even open models are often released without accompanying training data or recipes to reproduce them" [5]. A 2021 audit of C4 argued "for more thorough documentation of what is not in the data" [9]. Datasheets for Datasets gives its preprocessing questions a purpose: to tell consumers "whether the 'raw' data has been processed in ways that are compatible with their chosen tasks" [12].

This report is for ML engineers, data leads and buyers who commission or accept training data and need to know what each preparation stage does and what to ask for at each one.

Method

What was examined. The stages between a raw source and training-ready files: the source record, text extraction, language identification, rule-based and model-based filters, deduplication, personal data, decontamination, mixing and repetition, and packaging, with a short look at post-training data. For each stage: the published rule, its settings, the share it removed where stated, and the record a buyer can ask for.

Sources. Thirty-five sources, each opened and read for the passages cited: papers on open corpora and their pipelines [1, 2, 4, 5, 6, 24], papers in which model developers describe their own data [3, 7, 8, 15], studies of filtering, deduplication, mixing, licensing and consent [9, 10, 11, 16, 17, 18, 19, 20], post-training recipes [13, 14], an OCR toolkit paper [23], documentation frameworks and file formats [12, 21, 31, 32, 34], tool documentation [22, 29, 30, 33], NIST guidance on de-identification [35], and legal texts and one court order [25, 26, 27, 28].

How. Figures are given as each source states them, with the corpus they describe. Token counts depend on the tokenizer each source used, so counts from different sources do not compare exactly. Papers by model developers are described by what they are and named by title in Sources. One open-weight model family's technical report [7] is cited often and is called the open-weight report below. Legal texts are quoted and stated as what they say on the dates given. Web pages were viewed on 8 November 2026.

Findings

1. Most of a web crawl is removed, at rates that differ by pipeline

The RefinedWeb authors write that "Subsequent stages of Macrodata Refinement remove nearly 90% of the documents originally in CommonCrawl" [2]. About 50% of documents were discarded for not being English, 24% of the remainder for insufficient quality and 12% for being duplicates [2]. After language identification 48% of the original documents were left, and after filtering 23% [2]. The pipeline produced five trillion tokens, and the authors released an extract of 600 billion [2].

Pipeline	Starting point	What remained	Source
RefinedWeb	Common Crawl documents	Nearly 90% of documents removed, five trillion tokens	[2]
Dolma, CCNet stage	175.1 TB of Common Crawl	27.7 TB, after 84.2% was filtered out	[5]
DCLM-Baseline	DCLM-Pool, 240 trillion tokens	1.4% of the original documents, 3.8 trillion tokens	[6]
FineWeb	96 Common Crawl snapshots	About 36 trillion tokens after base filtering, 15 trillion released	[1]

Table 1. Removal across four published pipelines, as each source counts it.

In DCLM, the English filter alone removed 50.8% of the original documents [6]. Each figure is a share of an English-heavy web crawl after one pipeline's rules, and none says what a buyer's licensed or enterprise sources lose.

2. Extraction choices change model results

The FineWeb authors compared Common Crawl's WET text with text they extracted from the WARC files. WET files "retained too much boilerplate and menu text", and "Using trafilatura-extracted text from WARC files clearly results in a more performant model", measured on 28 billion tokens of training [1]. In DCLM, resiliparse and trafilatura improved its core evaluation score "by at least 2.5 points over the WET extraction", and "resiliparse is 8× faster to run" [6]. The trafilatura paper states the task: "retaining the desired content while discarding the rest", and the software extracts main text, comments and metadata [30].

The open-weight report used its own HTML parser, kept structure for math and code, kept image alt text and removed markdown markers, finding markdown "harmful to the performance of a model that is primarily trained on web data compared to plain text" [7].

PDFs need layout handling and often OCR. The olmOCR authors write that PDFs "have the potential to provide trillions of novel, high-quality tokens for training language models" [23]. Their toolkit converts PDFs into "clean, linearized plain text in natural reading order while preserving structured content like sections, tables, lists, equations" [23]. It was trained on 260,000 pages from over 100,000 crawled PDFs and tested on a benchmark of 1,400 PDFs [23]. The authors put their cost at 176 US dollars per million PDF pages, against over 6,240 dollars per million through a proprietary vision model's API [23]. Those are the authors' own 2025 figures for their setup.

At this stage a buyer can ask for the extractor and its version, sample pages before and after extraction, and an accuracy test on documents like the buyer's own.

3. Language identification thresholds are pipeline settings

fastText distributes two language identification models "which can recognize 176 languages", trained on Wikipedia, Tatoeba and SETimes [22]. The published pipelines set their own cut-offs.

Pipeline	Rule	Effect where stated	Source
C4	langdetect, English with probability of at least 0.99	Pages below it discarded	[4]
RefinedWeb	Top language score of at least 0.65	48% of documents left, mostly from this filter	[2]
FineWeb	fastText, English score of at least 0.65	Part of base filtering	[1]
Dolma	English document score of at least 0.5	61.7% of the data removed, by bytes	[5]
DCLM	English filter	50.8% of original documents removed	[6]

Table 2. Language identification in five pipelines.

RefinedWeb's authors note that a score below 0.65 "usually corresponds to pages without any natural text" [2]. For its multilingual data, the open-weight report used "a fasttext-based language identification model to categorize documents into 176 languages" [7]. FineWeb2 built a pipeline "that can be automatically adapted to support any language" and scaled it to over 1,000 languages from almost 100 Common Crawl snapshots, producing a 20 TB dataset of 5 billion documents [24]. A buyer working outside English asks how the threshold was set for each language.

4. Rule-based filters are simple and written down

The C4 rules fit in a paragraph. They retained only lines ending in a terminal punctuation mark, discarded pages with fewer than 5 sentences, kept lines of at least 3 words, removed pages with any word on a list of bad words, removed lines containing "Javascript" and pages containing "lorem ipsum" or a curly bracket, and kept one copy of any repeated three-sentence span [4].

The MassiveText authors set out their rules in an appendix headed "Quality Filtering (MassiveWeb only)". They removed documents outside 50 to 100,000 words, with a mean word length outside 3 to 10 characters, with a symbol-to-word ratio above 0.1 for the hash symbol or the ellipsis, or with more than 90% of lines starting with a bullet or more than 30% ending with an ellipsis [3]. They required 80% of words to contain a letter, and at least two of eight English stop words [3]. Their repetition table sets thresholds of its own.

Repetition measure	Threshold
Duplicate line fraction	0.30
Duplicate paragraph fraction	0.30
Duplicate line character fraction	0.20
Top 2-gram character fraction	0.20
Duplicate 10-gram character fraction	0.10

Table 3. Part of the MassiveText repetition thresholds, from Table A1 of [3].

FineWeb applied these quality and repetition filters "using the original thresholds", then a selection of C4's filters and three custom filters of its own [1]. The custom filters removed about 22% of tokens and raised the aggregate score by about 1% in its 28-billion-token tests [1]. The open-weight report added a duplicated n-gram coverage ratio, a count of "dirty" words and a Kullback-Leibler divergence over token distributions to catch documents full of outlier tokens [7].

Because these rules are short, a supplier can hand them over in full, with the share each one removed.

5. Model-based filters select hard, and their reference data sets what quality means

The DCLM authors conclude that "model-based filtering is key to assembling a high-quality training set" [6]. Their fastText classifier used instruction data and high-scoring posts from an explain-it-simply forum as positive examples, gave "a 3.5 percentage point lift" on the core score, and the authors' takeaway is to keep

"the top 10% of documents" [6]. Trained on 2.6 trillion tokens of DCLM-Baseline mixed with code and math data, a 7-billion-parameter model reached 64% five-shot accuracy on MMLU [6].

FineWeb-Edu took a different reference. A 70-billion-parameter instruction-tuned model scored 460,000 random web pages "for their educational quality on a scale from 0 to 5" [1]. A classifier trained on those scores, with a threshold of 3, reached an F1 score of 82% on validation, and the result was a 1.3-trillion-token collection [1]. The open-weight report trained fastText classifiers to recognize text "referenced by Wikipedia", and other classifiers on labels from an earlier model [7].

Each reference set, whether instruction data, educational value or pages cited by an encyclopedia, decides what the filter keeps. The Dolma authors write that "The term 'quality filter,' while widely used in literature, does not appropriately describe the outcome of filtering a dataset" [5]. A buyer asks what the positive examples were, where the cut-off sits and what documents just above and below it look like.

6. Filters have side effects

The 2021 audit of C4 found that "blocklist filtering disproportionately removes text from and about minority individuals" [9]. African-American English and Hispanic-aligned English were removed at 42% and 32%, against 6.2% for white-aligned English and 7.2% for other English [9]. The MassiveText authors used a search engine's safe-search signal "rather than manual word-list filters", citing that finding [3]. The same audit found "a significant amount of text from unexpected sources like patents and US military websites", and machine-generated text [9].

A Pretrainer's Guide to Training Data trained 28 decoder-only models of 1.5 billion parameters [11]. It found that "Quality filtering, removing low-quality text, substantially increases both toxic generation and downstream performance", and that removing toxic data "trades-off fewer toxic generations for reduced generalization performance" [11]. Its authors conclude that "there does not exist a one-size-fits-all solution to filtering training data" [11]. For a buyer, the sample of removed text is the check that a filter removed what was meant.

7. Deduplication: levels, settings and a counterexample

Level	Published example	Share or setting	Source
URL	Dolma, exact URL	53.2% of documents removed	[5]
URL	The open-weight report	Keeps "the most recent version" of pages for each URL	[7]

Level	Published example	Share or setting	Source
Document	Dolma, exact document	14.9% of URL-deduplicated documents	[5]
Near-duplicate document	FineWeb, MinHash	5-grams, 112 hash functions in 14 buckets of 8, at least 75% similar	[1]
Paragraph	Dolma, exact paragraph	18.7% of paragraphs	[5]
Three-sentence span	C4	One copy kept	[4]
Line	The open-weight report	Lines seen more than 6 times in each bucket of 30 million documents removed	[7]
Substring	Exact-substring method	7.18% of C4's tokens	[10]

Table 4. Deduplication levels and settings.

Order matters. Dolma ran paragraph deduplication last because it "risks disrupting content analysis" [5]. The open-weight report found that line deduplication removed boilerplate and also frequent high-quality text, and its evaluations still showed "strong improvements" [7].

The effects of duplicates are measured. Lee and colleagues found that "over 1% of the unprompted output of language models trained on these datasets is copied verbatim from the training data" [10]. One 61-word English sentence was repeated over 60,000 times in C4, and models trained on deduplicated data "emit memorized text ten times less frequently and require fewer train steps" [10]. Near duplicates made up 3.04% of C4 and 13.63% of RealNews, and one cluster held 250,933 copies [10]. An analysis of ten corpora found "about 50% of the documents in RedPajama and LAION-2B-en are duplicates" [18].

More deduplication did not always help. FineWeb's authors ran MinHash across all 96 snapshots at once [1]. In the oldest snapshots "this process removed as much as 90% of the original base filtered data", the whole deduplicated dataset came to 4 trillion tokens, and "our ablation models showed little improvement over a model trained on the non-deduplicated data" [1]. They deduplicated each snapshot on its own instead [1]. FineWeb2 rebalances its data using "both duplication count and quality" [24]. A buyer asks at which level duplicates were removed, with what settings, and whether the scope was one snapshot, one source or all of them.

8. Personal data handling in published pipelines is narrow

Dolma focused on "three kinds of PII that are detectable with high precision: email addresses, IP addresses and phone numbers" [5]. Documents with 5 or fewer such spans had them replaced with a token such as `|||EMAIL_ADDRESS|||`, which affected 0.02% of documents [5]. Documents with more were removed, 0.001% of documents [5]. In the authors' ablations, the choice between masking and removal had no effect on model performance [5]. FineWeb anonymized email and public IP addresses for its public release [1]. The open-weight report removed whole domains "that contain large amounts of personally identifiable information (PII)", along with domains with known adult content [7]. The ten-corpus analysis listed personally identifiable information among its findings [18].

NIST SP 800-188 defines de-identification as "a general term for any process of removing the association between a set of identifying data and the data subject" [35]. A buyer whose sources hold names, addresses, account numbers or free-text notes names the types to handle, and the method for each, in the specification.

The EU summary template asks providers for "measures taken to avoid or remove illegal content under Union law from the training data (such as blacklists, keywords, and model-based classifiers)", and adds that "Such measures do not include data selection practices, for example to increase the capability of the model" [26].

9. Decontamination methods and how they fail

Benchmark text sits inside web crawls. The C4 audit found 1.87% to 24.88% of the target texts of the generation benchmarks it studied in C4 [9]. Train-test overlap "affects over 4% of the validation set of standard datasets" [10], and the ten-corpus analysis listed benchmark contamination among its findings [18].

Method	Rule	Source
13-gram overlap	Examples with "a 13-gram overlap with anything in the pretraining set", or with the whole example when shorter	[8]
8-gram prompt matching	A test item overlaps when more than 50% of its tokens have 8-gram matches with one training instance, and a set is contaminated when its instances overlap more than 2% of any evaluation	[13]
Bloom filter	Seeded with test examples, then flags duplicates while the pretraining data is mixed	[5]
Disclosure	Every submission discloses a decontamination report	[6]

Table 5. Decontamination methods.

The 175-billion-parameter model's paper reports the failure plainly: "Unfortunately, a bug in the filtering caused us to ignore some overlaps, and due to the cost of training it was not feasible to retrain the model" [8]. Tulu 3 checked prompts alone, because completions are often regenerated, and did not use embeddings because "we found it difficult to distinguish mere distributional similarity from actual paraphrasing" [13]. Its removals ran from 0.7% to 11.3% of a dataset, with 11.3% taken from a math set checked against a math benchmark [13].

DCLM asks all submissions "to disclose a decontamination report and avoid using highly-contaminated data" [6]. A buyer names the benchmarks and internal test sets in use, and asks for the overlap found and removed against each.

10. Mixing and repetition

The open-weight report states its final pretraining mix: "roughly 50% of tokens corresponding to general knowledge, 25% of mathematical and reasoning tokens, 17% code tokens, and 8% multilingual tokens", chosen through "scaling law experiments in which we train several small models on a data mix" [7]. Its corpus was about 15 trillion multilingual tokens, against 1.8 trillion for the previous generation [7].

The DoReMi authors write that "The mixture proportions of pretraining data domains (e.g., Wikipedia, books, web text) greatly affect language model (LM) performance". A 280-million-parameter proxy model set domain weights for an 8-billion-parameter model, which improved average few-shot accuracy by 6.5 points over The Pile's default weights and reached the baseline's accuracy with 2.6 times fewer training steps [19].

Repetition has a measured range. In experiments up to 900 billion tokens and 9 billion parameters, "training with up to 4 epochs of repeated data yields negligible changes to loss compared to having unique data", and with more repetition "the value of adding compute eventually decays to zero" [20]. A compute-optimal study found that model size and training tokens "should be scaled equally", with a 70-billion-parameter model trained on 1.4 trillion tokens [15]. Data age matters as well: the Pretrainer's Guide saw "performance degradation if evaluation data is either before or after pretraining data collection" [11].

A buyer asks for the share of each source in the delivered set, how many times each part repeats and the capture period it covers.

11. Published post-training sets are smaller and curated or scored

LIMA fine-tuned a 65-billion-parameter model "on only 1,000 carefully curated prompts and responses", about 750,000 tokens in all [14]. The open-weight report notes that "most of our training data is model-generated, it requires careful cleaning and quality control" [7]. Its post-training data went through rule-based cleaning of overused phrases, topic classification, quality scoring, difficulty scoring and semantic deduplication, keeping only examples whose maximum cosine similarity to the examples seen so far in a cluster was below a threshold [7]. Tulu 3's prompt decontamination, above, applies to post-training sets [13]. Under California's AB 2013, to train a system "includes testing, validating, or fine tuning by the developer" [27].

12. Documentation and disclosure ask for the processing record

Datasheets for Datasets asks three questions about preprocessing: whether any "preprocessing/cleaning/labeling of the data" was done, whether "the 'raw' data" was saved, and whether "the software that was used to preprocess/clean/label the data" is available [12]. A dataset card is the README of a dataset repository, with metadata on "license, language, and size" [34]. Croissant metadata "provides a vocabulary for dataset attributes, streamlining how data is loaded across ML frameworks", and an extension covers the data life cycle and labeling [21]. JSON Lines requires UTF-8, one valid JSON value per line and `\n` as the line terminator, with a terminator after the last value recommended but not required [32]. Parquet is "an open source, column-oriented data file format designed for efficient data storage and retrieval" [31].

Some pipelines release their software. FineWeb's authors released their curation codebase [1], and DataTrove is "a library to process, filter and deduplicate text data at a very large scale" [33]. Dolma released its toolkit [5]. Rights records lag behind. An audit of more than 1,800 text datasets found "license omission of 70%+ and error rates of 50%+" on widely used hosting sites [16]. Terms also change after capture. A 2024 audit of 14,000 web domains found that restrictions added in a single year, 2023 to 2024, left about 5% or more of all C4 tokens fully restricted, and that terms of service crawling restrictions covered 45% of C4 at the time of the audit [17].

California's AB 2013 was chaptered on 28 September 2024 as Chapter 817 [27]. It requires the developer of a "generative artificial intelligence system or service" released on or after 1 January 2022 to post documentation about its training data on or before 1 January 2026, and again each later time the system or a substantial modification of it is made publicly available to Californians [27]. Item 9 asks "Whether there was any cleaning, processing, or other modification to the datasets by the developer, including the intended purpose of those efforts in relation to the artificial intelligence system or service" [27]. In a challenge to the law, a developer alleged that "cleaning methods" were "valuable and non-public" [28]. On 4 March 2026 the court found that it "has not identified any dataset or approach to cleaning and using datasets that is distinct from its competitors in a manner warranting trade secret protection", and denied the preliminary injunction [28].

The EU AI Act's Article 10 covers "high-risk AI systems" that train models [25]. Their data governance practices "shall concern in particular" the "data collection processes and the origin of data", "relevant data-preparation processing operations, such as annotation, labelling, cleaning, updating, enrichment and aggregation" and "examination in view of possible biases" [25]. Article 10(3) asks that data sets be "relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose" [25]. Article 53(1)(d) asks providers of "general-purpose AI models" for a public summary of training content on a template from the AI Office [25]. The template of 24 July 2025 has a section on "Relevant data processing aspects", covering reservations of rights under the text and data mining exception and the removal of illegal content [26]. Its list of sources asks for the domain names "in the top 10 % of all domain names determined by the size of the content", and for SMEs the top 5% of domain names or 1,000 domains, "whichever is lower" [26].

Implications

A checklist by stage. Each row lists what a buyer can ask a supplier to deliver, and the finding behind it.

Stage	Ask for	Finding
Source record	Source list, capture dates, terms at capture, where the raw copy is kept	12
Extraction	Extractor and version, sample pages before and after, an accuracy test	2
Language	Classifier, threshold per language, share removed	3
Rule-based filters	Every rule with its numbers, share removed per rule	4
Model-based filters	Classifier, positive examples, cut-off, samples above and below it	5
Side effects	A sample of removed text, read for what was not meant to go	6
Deduplication	Level, method, settings, scope, share removed at each level	7
Personal data	Types handled, method for each, share affected	8
Decontamination	Overlap found and removed against each named benchmark and test set	9
Mix	Share of each source, repetition, capture period	10
Post-training	Scoring rules, deduplication method, decontamination of prompts	11
Delivery	Files, datasheet, removal log per stage, version, change log	12

Table 6. What to ask for at each stage.

For buyers. Ask for the record before the data. A share removed means little without the rule and a sample of what it removed [9, 11]. Ask for thresholds as settings with a reason, since the published ones differ by pipeline and were tuned on small models [1, 2, 5, 11]. Name the benchmarks and internal test sets that must stay out [8, 13].

For teams preparing data. Keep the raw copy and the processing software, as the datasheet questions ask [12]. Log each stage's removals, because AB 2013 asks a developer whether any cleaning took place and for what purpose [27], and Article 10 names data preparation among governance practices for "high-risk AI systems" [25]. Test deduplication and filtering scope on held-out tasks, since one wide deduplication gave little improvement [1].

Limits of this analysis

- The published pipelines process web text, mostly English. Removal shares from them do not describe licensed, enterprise or non-web sources.
- Thresholds were tuned with small test models: FineWeb's had 1.71 billion parameters and trained on about 28 billion tokens [1], and the Pretrainer's Guide used 1.5 billion [11]. They are those papers' settings, and none is a standard.
- Closed developers rarely detail their pipelines [5]. Only one developer's position is quoted, from a court order [28].
- Token counts depend on each source's tokenizer, so sizes from different sources do not compare exactly.
- The olmOCR cost figures are the authors' own, for their 2025 setup [23].
- The legal texts are stated as they read on the dates given. Only the order of 4 March 2026 in the AB 2013 case was read [28], and this report states no date on which Article 10 applies. This summary of the texts is research, never legal advice.
- Several sources are preprints.

Sources

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

2. 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

3. 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

4. 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 (2020), viewed 8 November 2026. https://jmlr.org/papers/v21/20-074.html

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

6. 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

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

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

9. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. Jesse Dodge, Maarten Sap, Ana Marasović and others, EMNLP 2021, Association for Computational Linguistics, November 2021, viewed 8 November 2026. https://aclanthology.org/2021.emnlp-main.98/

10. Deduplicating Training Data Makes Language Models Better. Katherine Lee, Daphne Ippolito, Andrew Nystrom and others, ACL 2022, Association for Computational Linguistics, May 2022, viewed 8 November 2026. https://aclanthology.org/2022.acl-long.577/

11. 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

12. Datasheets for Datasets. Timnit Gebru, Jamie Morgenstern, Briana Vecchione and others, arXiv:1803.09010, version 8, 1 December 2021, viewed 8 November 2026. https://arxiv.org/abs/1803.09010

13. Tulu 3: Pushing Frontiers in Open Language Model Post-Training. Nathan Lambert, Jacob Morrison, Valentina Pyatkin and

others, arXiv:2411.15124, version 5, 14 April 2025, viewed 8 November 2026. https://arxiv.org/abs/2411.15124

14. LIMA: Less Is More for Alignment. Chunting Zhou, Pengfei Liu, Puxin Xu and others, arXiv:2305.11206, 18 May 2023, viewed 8 November 2026. https://arxiv.org/abs/2305.11206

15. Training Compute-Optimal Large Language Models. Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch and others, arXiv:2203.15556, 29 March 2022, viewed 8 November 2026. https://arxiv.org/abs/2203.15556

16. The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI. Shayne Longpre, Robert Mahari, Anthony Chen and others, arXiv:2310.16787, version 3, 4 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2310.16787

17. Consent in Crisis: The Rapid Decline of the AI Data Commons. Shayne Longpre, Robert Mahari, Ariel Lee and others, arXiv:2407.14933, version 2, 24 July 2024, viewed 8 November 2026. https://arxiv.org/abs/2407.14933

18. What's In My Big Data? Yanai Elazar, Akshita Bhagia, Ian Magnusson and others, arXiv:2310.20707, version 2, 5 March 2024, viewed 8 November 2026. https://arxiv.org/abs/2310.20707

19. DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining. Sang Michael Xie, Hieu Pham, Xuanyi Dong and others, arXiv:2305.10429, version 4, 21 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2305.10429

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

21. Croissant Format Specification, Version 1.0. MLCommons, 1 March 2024, viewed 8 November 2026. https://docs.mlcommons.org/croissant/docs/croissant-spec.html

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

23. olmOCR: Unlocking Trillions of Tokens in PDFs with Vision Language Models. Jake Poznanski, Aman Rangapur, Jon Borchardt and others, arXiv:2502.18443, version 3, 2 July 2025, viewed 8 November 2026. https://arxiv.org/abs/2502.18443

24. 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

25. Regulation (EU) 2024/1689 (Artificial Intelligence Act). European Parliament and Council of the European Union, Official Journal of the European Union, 12 July 2024, viewed 8 November 2026. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

26. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. European Commission, AI Office, 24 July 2025, viewed 8 November 2026. https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models

27. AB 2013, Generative artificial intelligence: training data transparency (Chapter 817, Statutes of 2024). California Legislature, chaptered text as published by CalMatters Digital Democracy, 28 September 2024, viewed 8 November 2026. https://calmatters.digitaldemocracy.org/bills/ca_202320240ab2013

28. X.AI LLC v. Rob Bonta, Case No. CV 25-12295 JGB (SSCx), Civil Minutes, Order Denying Preliminary Injunction. United States District Court, Central District of California, 4 March 2026, filed copy hosted by Ars Technica, viewed 8 November 2026. https://cdn.arstechnica.net/wp-content/uploads/2026/03/xAI-v-Bonta-Order-Denying-Preliminary-Injunction-3-4-26.pdf

29. Get Started. Common Crawl, undated, viewed 8 November 2026. https://commoncrawl.org/get-started

30. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. Adrien Barbaresi, ACL-IJCNLP 2021 System Demonstrations, Association for Computational Linguistics, August 2021, viewed 8 November 2026. https://aclanthology.org/2021.acl-demo.15/

31. Overview. Apache Parquet documentation, The Apache Software Foundation, undated, viewed 8 November 2026. https://parquet.apache.org/docs/overview/

32. JSON Lines. jsonlines.org, undated, viewed 8 November 2026. https://jsonlines.org/

33. DataTrove (README). Hugging Face, GitHub repository huggingface/datatrove, undated, viewed 8 November 2026. https://github.com/huggingface/datatrove

34. Dataset Cards. Hugging Face Hub documentation, undated, viewed 8 November 2026. https://huggingface.co/docs/hub/datasets-cards

35. De-Identifying Government Datasets: Techniques and Governance, NIST SP 800-188. Simson Garfinkel, Barbara Guttman, Joseph Near and others, National Institute of Standards and Technology, September 2023, viewed 8 November 2026. https://csrc.nist.gov/pubs/sp/800/188/final

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. Preparing data for language model training: a buyer's map from source to delivery. 13 November 2026. https://rillor.com/insights/preparing-data-for-language-model-training

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