Rillor

EXPLAINER

What data is used to train language models

What data is used to train LLMs, stage by stage: filtered web crawls, code, papers and books, then demonstrations, preference pairs, synthetic and checkable tasks, and what must be disclosed.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

Language models are trained in stages, and the open reports and dataset cards show what data each stage uses.

- Pretraining uses trillions of tokens, the units a tokenizer splits text into, mostly from filtered web text. Llama 3, one open-weight model family, pretrained on about 15 trillion multilingual tokens [1], and the OLMo 2 pretraining mix of about 3.9 trillion tokens took over 95% from web data [2].
- The open web corpora C4, RefinedWeb, FineWeb, DCLM and RedPajama-V2 are all built from Common Crawl, a free, open repository of web crawl data [5, 9, 11, 12, 15, 17]. Its October 2026 crawl held 2.16 billion pages [7].
- Curation keeps a fraction of the raw text. After heuristic filters and deduplication, a classifier kept the top 10% of DCLM's remaining documents, the 3.8 trillion token DCLM-Baseline [15], and Dolma was curated from about 200 TB of raw text down to 11 TB [19].
- Curated sources add code, scientific papers, books, encyclopedias, forums and math. The Pile combined 22 subsets [18], and the final Llama 3 mix was roughly 50% general knowledge, 25% math and reasoning, 17% code and 8% multilingual tokens [1].
- Post-training uses far smaller sets of written demonstrations, ranked or rated answers and tasks with checkable answers. A 2022 instruction-following study fine-tuned on about 13,000 prompts with labeler demonstrations [3], and the Tulu 3 supervised mixture holds 939,344 samples [25].
- Synthetic data appears in open recipes, from generated instructions [27] to the 25 billion tokens of the Cosmopedia dataset [28]. A 2024 Nature study found that indiscriminate use of model-generated content in training made the tails of the original distribution disappear [29].
- Rights records are often incomplete. An audit of more than 1,800 text datasets found license omission rates above 70% and error rates above 50% on widely used hosting sites [31].
- The EU AI Act requires providers of general-purpose AI models to publish a summary of their training content, and the AI Office's template of 24 July 2025 sorts its sources into six types [35, 36].

Background

Why the question has several answers

A language model learns from a sequence of datasets, each built for a stage of training. The authors of RefinedWeb describe the common pretraining recipe as "a mixture of filtered web data and curated 'high-quality' corpora, such as social media conversations, books, or technical papers" [11]. Post-training then uses prompts with written answers, preferences between answers and tasks with checkable answers [4].

Much of this is undocumented for closed models. 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" [19]. The Tulu 3 authors call post-training data and recipes "simultaneously the most important pieces of the puzzle and the portion with the least transparency" [4]. Open corpora and open model families fill part of that gap, and since 2 August 2025 the EU AI Act requires providers of general-purpose AI models to publish a summary of their training content [35, 36].

Tokens as the unit of size

Pretraining data is measured in tokens, the units a tokenizer splits text into. The count depends on the tokenizer. Dolma counts with the Llama tokenizer [19], DCLM with the GPT-NeoX tokenizer [15], and the FineWeb2 card notes that the count for a language "heavily depends on whether the tokenizer was trained with that language" [14]. Sizes in this report carry the source's own count, and counts from different sources do not compare exactly.

This report is for ML engineers, data leads and product owners who build, buy or document training data and need to know what the published corpora and recipes contain.

Method

What was examined. The data used at each stage of training, studies of scale, duplication and contamination, audits of licensing and consent, and the EU's disclosure template.

Sources. Dataset papers [9, 11, 12, 15, 16, 18, 19, 20, 33, 34], dataset cards [13, 14, 17, 21, 25, 26, 28], open model family reports [1, 2, 4], method and scaling studies [3, 22, 23, 24, 27, 29, 30], audits [10, 31, 32], Common Crawl's own pages [5, 6, 7, 8] and the EU regulation and template [35, 36].

How. Figures come from each paper's full text or each card, with the version and date of the source. Token counts name the tokenizer where the source states it. Tables of corpus components list selected components, so their shares do not sum to 100%. Model families are named by family name. Web pages and cards were viewed on 8 November 2026. The report describes what the sources say about licenses, robots.txt and opt-outs, and draws no legal conclusions.

Findings

1. Each training stage uses its own data

The open model family reports describe training as a sequence of stages, each with its own mix.

Stage	Data	Published example
Pretraining	Filtered web text with code, papers, books, encyclopedias and math	About 15 trillion multilingual tokens for Llama 3 [1], about 3.9 trillion for OLMo 2 [2]
Mid-training or annealing	Up-sampled high-quality web documents, curated sources, synthetic data	OLMo 2 mixes of 50, 100 and 300 billion tokens [2]
Supervised fine-tuning	Prompts with written completions	About 13,000 training prompts in a 2022 study [3], 939,344 samples in the Tulu 3 mixture [25]
Preference tuning	Rankings or chosen and rejected pairs	33,000 reward model prompts [3], around 1 million possible pairs from UltraFeedback [26]
Reinforcement learning	Prompts, or tasks with verifiable answers	31,000 prompts [3], math and precise instruction following [4]

OLMo 2's base models "are trained in two stages each with its corresponding data mix" [2]. The first stage "uses mostly web-sourced data". In the second, which the authors call mid-training, they "up-sample the highest-quality web documents and curated non-web sources" and "employ synthetic data crafted to patch math capabilities" [2]. The mid-training mix, Dolmino Mix 1124, was sampled at 50, 100 and 300 billion tokens [2].

Llama 3's 2024 report describes the same pattern. Its pretraining data contains knowledge until the end of 2023 [1]. Annealing on small amounts of high-quality code and math data can boost performance on key benchmarks [1]. When the GSM8k and MATH training sets were added to annealing, an 8B model's scores on those validation sets rose by 24.0% and 6.4%, and the 405B model's gains were negligible [1]. Post-training ran in six rounds, each supervised fine-tuning followed by Direct Preference Optimization, on examples "collected either via human annotations or generated synthetically" [1].

Tulu 3, an open post-training recipe, names one data type per stage: "prompt-completion instances for supervised finetuning, preferences for preference tuning, or verifiable rewards for reinforcement learning" [4].

2. Web text comes from Common Crawl

Common Crawl "maintains a free, open repository of web crawl data that can be used by anyone" [5]. Its home page gives "Over 300 billion pages spanning 15 years", "Free and open corpus since 2007" and 3 to 5 billion new pages added each month [5]. Its overview page says the corpus "contains petabytes of data, regularly collected since 2008" [6].

The October 2026 crawl, published on 21 October 2026, was collected between 5 and 18 October [7]. It holds 2.16 billion web pages, or 371.19 TiB of uncompressed content, from 41.4 million hosts or 34.3 million registered domains, and includes 830.1 million URLs not visited in any earlier crawl [7]. It ships as 100,000 each of WARC, WAT and WET files [7]. WARC files hold raw HTTP responses and WET files hold extracted text only [12].

Each open web corpus starts from these files.

- **C4** used Common Crawl's extracted text from April 2019 [9]. Its authors note that Common Crawl produces "around 20TB of scraped text data each month" [9].
- **RefinedWeb** obtained five trillion tokens from Common Crawl and released an extract of 600 billion [11].
- **FineWeb** used 96 Common Crawl snapshots and re-extracted text from WARC files with trafilatura, because "WET files retained too much boilerplate and menu text" [12].
- **FineWeb2** used 96 snapshots "spanning the summer of 2013 to April 2024" and provides filtered data for 1,868 language-script pairs [14].
- **DCLM** re-extracted text from HTML with resiliparse. DCLM-Pool "contains 200B documents (370TB after gzip compression), resulting in 240T GPT-NeoX tokens" [15].
- **RedPajama-V2** holds "over 100B text documents coming from 84 CommonCrawl snapshots" [17].

3. Filtering and deduplication decide what stays

The C4 authors found that "the majority of the resulting text is not natural language" in Common Crawl's extracted text [9]. Later corpora tested and reused some of its rules [12, 16]. C4 kept only lines ending in a terminal punctuation mark and lines with at least five words, discarded pages with fewer than three sentences, removed pages containing any word on a published list of bad words or a curly bracket, and "discarded all but one of any three-sentence span occurring more than once" [9]. A language detector kept pages with an English probability of at least 0.99 [9]. About 750 GB remained [9].

Filters remove more than noise. A 2021 audit of C4 found that "blocklist filtering disproportionately removes text from and about minority individuals" [10]. It also found text from unexpected sources, including machine-translated patents and 33,874,654 tokens from the .mil top-level domain [10].

FineWeb tested each step with ablation models [12]. Base filtering by URL blocklist, fastText English language identification and MassiveText quality and repetition filters left "roughly 36 trillion tokens" across 96 snapshots [12]. Deduplication used MinHash on 5-grams with 112 hash functions in 14 buckets of 8, targeting

documents at least 75% similar [12]. Deduplicating all snapshots together gave only a modest gain, while "individually deduplicating each snapshot" improved the average score. Among C4's filters, terminal punctuation "gives the biggest individual boost, but removes around 30% of all tokens" [12]. The paper reports 15 trillion tokens [12]. The card, last modified on 11 July 2025, describes "more than 18.5T tokens (originally 15T tokens)", after version 1.4.0 of that date added six snapshots from January to June 2025 [13].

Model-based classifiers sit in several pipelines.

- **FineWeb-Edu.** An instruction model of the Llama 3 family scored 460,000 randomly sampled pages for educational quality on a scale from 0 to 5 [12]. A classifier trained on 410,000 of those labels, with a threshold of 3, reached an F1 of 82% on the held-out 50,000 [12]. The result is a 1.3 trillion token set [12].
- **DCLM.** After RefinedWeb's heuristic filters and deduplication, a fastText classifier trained on instruction-formatted data (OpenHermes 2.5) and high-scoring posts from the r/ExplainLikeImFive subreddit kept the top 10% of documents, giving DCLM-Baseline, 3.8 trillion tokens [15]. A 7B model trained on 2.6 trillion tokens reached 64% 5-shot accuracy on MMLU, and the authors conclude that "model-based filtering is key to assembling a high-quality training set" [15].
- **Llama 3.** Its pipeline uses a custom HTML parser, removal of markdown markers, deduplication by URL, by document with global MinHash and by line (lines that "appeared more than 6 times in each bucket of 30M documents"), heuristic filters such as duplicated n-gram coverage and "dirty word" counts, and model-based quality classifiers [1]. It also removed domains with large amounts of personally identifiable information and domains with known adult content [1].

RedPajama-V2 takes the other route. It is "raw, unfiltered text data together with quality signals and metadata", with signals drawn from the C4, Gopher and RefinedWeb rules and others, so users choose their own filter [16]. Dolma shows the overall effect of curation: its sources held about 200 TB of raw text, curated down to an 11 TB dataset [19].

4. Curated sources add code, papers, books, encyclopedias, forums and math

Pretraining mixes add sources chosen by domain. The Pile, released on 31 December 2020, is "an 825 GiB English text corpus" built from "22 diverse high-quality subsets", "many of which derive from academic or professional sources" [18]. Selected components:

The Pile component	Size	Share of bytes	Passes per epoch
Pile-CC (Common Crawl)	227.12 GiB	18.11%	1.0
PubMed Central	90.27 GiB	14.40%	2.0
Books3	100.96 GiB	12.07%	1.5

The Pile component	Size	Share of bytes	Passes per epoch
OpenWebText2	62.77 GiB	10.01%	2.0
ArXiv	56.21 GiB	8.96%	2.0
GitHub	95.16 GiB	7.59%	1.0
Stack Exchange	32.20 GiB	5.13%	2.0
USPTO Backgrounds	22.90 GiB	3.65%	2.0
Wikipedia (en)	6.38 GiB	1.53%	3.0

Source: The Pile, Table 1 [18]. The raw total is 825.18 GiB and the effective size, after repeats, 1,254.20 GiB [18]. The authors "increase the weights of higher quality components", so Wikipedia is seen up to three times [18].

Dolma v1.6, described in its June 2024 paper, is "a three-trillion-token English corpus" of web content, scientific papers, code, public-domain books, social media and encyclopedic material [19].

Dolma source	UTF-8 GB	Documents (millions)	Llama tokens (billions)
Common Crawl web pages	9,812	3,734	2,479
GitHub code	1,043	210	411
Reddit social media	339	377	89
Semantic Scholar papers (peS2o)	268	38.8	70
Project Gutenberg books	20.4	0.056	6.0
Wikipedia and Wikibooks	16.2	6.2	4.3
Total	11,519	4,367	3,059

Source: Dolma, Table 1 [19]. peS2o is about 40 million open-access academic papers [19].

Specialist corpora cover one domain. The Stack v2 is built on Software Heritage's source code archive, with repositories "spanning 619 programming languages", plus pull requests, notebooks and code documentation [20]. It is "4x larger than the first StarCoder dataset", its models were trained on 3.3 to 4.3 trillion tokens, and the archive identifiers of the source code are released [20]. OpenWebMath was "filtered and extracted from over 200B HTML files on Common Crawl down to a set of 6.3 million documents containing a total of 14.7B tokens" [21].

OLMo 2's pretraining mix, OLMo 2 Mix 1124, shows how these pieces combine [2]:

OLMo 2 Mix 1124 source	Tokens
DCLM-Baseline web pages	3.71 trillion
StarCoder code	83.0 billion
peS2o academic papers	58.6 billion
arXiv STEM papers	20.8 billion
OpenWebMath	12.2 billion
Algebraic Stack	11.8 billion
Wikipedia and Wikibooks	3.7 billion
Total	3.90 trillion

Source: 2 OLMo 2 Furious, Table 4 [2]. Over 95% of the tokens are web data, and code from repositories with fewer than two GitHub stars was removed [2]. Llama 3 reports its final mix as shares: roughly 50% general knowledge, 25% mathematical and reasoning tokens, 17% code and 8% multilingual [1]. Its language identifier covers 176 languages [1]. Coverage across languages is uneven. Of FineWeb2's 1,868 language-script pairs, 474 have more than a thousand documents and 203 more than ten thousand [14].

5. How much data: token counts and the stock of text

A 2022 study of compute-optimal training trained over 400 models from 70 million to over 16 billion parameters on 5 to 500 billion tokens [22]. It found that "for every doubling of model size the number of training tokens should also be doubled" [22]. Its 70B model was trained on 1.4 trillion tokens, four times the data of its 280B predecessor [22].

Published token counts since then:

Model or dataset	Training tokens	Source
Llama 2	1.8 trillion	[1]
Llama 3	About 15 trillion (405B model: 15.6 trillion, 8K context)	[1]
OLMo 2 pretraining mix	3.90 trillion	[2]
DCLM-Baseline 7B model	2.6 trillion	[15]
StarCoder2 models	3.3 to 4.3 trillion	[20]
Comma v0.1 (Common Pile)	1 and 2 trillion	[33]

Repetition stretches a fixed corpus up to a point. In experiments up to 900 billion training tokens and 9 billion parameters, "training with up to 4 epochs of repeated data yields negligible changes to loss compared to having unique data", but "with more repetition, the value of adding compute eventually decays to zero" [23].

The stock of public text is finite. A study revised on 4 June 2024 estimated the stock of web data, as medians with 95% intervals: Common Crawl 130 trillion tokens [100T, 260T], the indexed web 510 trillion [130T, 2,100T] and the whole web 3,100 trillion [1,900T, 5,200T] [24]. It put the effective stock of indexed web text at around 4e14 tokens [24]. It projected that, if current trends continued, models would be trained on datasets roughly equal in size to the stock of public human text between 2026 and 2032, with a median year of 2028, or slightly earlier if models are overtrained [24]. It named synthetic data generation, transfer learning from data-rich domains and data efficiency improvements as ways forward [24].

6. Post-training data: demonstrations, preferences and checkable tasks

The 2022 instruction-following study shows the basic shape of post-training data [3]. It started from labeler-written prompts and prompts submitted through a production API, collected "a dataset of labeler demonstrations of the desired model behavior" for supervised fine-tuning, then "a dataset of rankings of model outputs" for reinforcement learning from human feedback [3]. Outputs of the tuned 1.3B model were preferred to those of the 175B base model "despite having 100x fewer parameters" [3].

Later recipes refine each part.

- **Preferences with strength.** Llama 3's preference annotators rated the strength of their preference on four levels and could edit the chosen response [1].

- **Rejection sampling.** Llama 3 drew "typically between 10 and 30" outputs per prompt and kept the one its reward model picked [1].
- **On-policy preferences.** Tulu 3 built preference data from its own model's completions against other models', adapting the UltraFeedback pipeline [4].
- **Verifiable rewards.** Tulu 3's reinforcement learning stage "only gets a reward if the model's completions are verified to be correct", aimed at skills with verifiable answers such as mathematics and precise instruction following [4].

Post-training sets are mixtures. Tulu 3's prompts were curated "from publicly available datasets with clear provenance" and synthesized "to fill any gaps" [4]. The released supervised mixture "contains 939,344 samples", among them FLAN v2 (89,982 prompts), No Robots (9,500, under CC-BY-NC-4.0), OpenAssistant (7,132) and the synthetic Tulu 3 Persona MATH (149,960), Persona GSM (49,980) and Persona Python (34,999) sets [25]. UltraFeedback "consists of 64k prompts, 256k responses and 380k high-quality feedback", from which researchers "could further construct around 1 million comparison pairs" [26]. Its ratings were written by a model acting as the annotator, which the card says "also makes mistakes and provides inaccurate feedbacks" [26].

7. Synthetic data: what the studies show

Synthetic data enters at several stages. Self-Instruct, published at ACL 2023, "generates instructions, input, and output samples from a language model, then filters invalid or similar ones", and its authors reported "a 33% absolute improvement over the original model on Super-NaturalInstructions" [27].

Cosmopedia is "a dataset of synthetic textbooks, blogposts, stories, posts and WikiHow articles" generated by an open-weight mixture-of-experts instruction model, with "over 30 million files and 25 billion tokens" [28]. Each prompt uses a seed sample, such as an extract from a web page, and the two web-sample splits make up about 75% of the dataset [28]. Tulu 3 conditioned synthetic prompts on about 250,000 personas [4], OLMo 2 used synthetic data to patch math in mid-training [2], and Llama 3's post-training mixed human-annotated and synthetic examples [1]. Models also label data: FineWeb-Edu's quality labels and UltraFeedback's ratings were written by models [12, 26].

The risk is recursion. A Nature study published on 24 July 2024 found that "indiscriminate use of model-generated content in training causes irreversible defects in the resulting models, in which tails of the original content distribution disappear", an effect it called model collapse [29]. In its experiment with a language model fine-tuned on wikitext2, performance degraded across generations when no original data was kept [29]. With "a random 10% of the original data points" kept in each generation, the result was "only minor degradation of performance" [29]. The study covers recursive training on generated data in controlled settings. Self-Instruct, by contrast, reported a gain from filtered synthetic data [27].

8. Rights, consent and licenses

License records are often missing or wrong. The Data Provenance Initiative audited "1800+ text datasets" and found "frequent miscategorization of licenses on widely used dataset hosting sites, with license omission of 70%+ and error rates of 50%+" [31]. Cards themselves carry layered terms. FineWeb is released under ODC-By 1.0 [13], OpenWebMath under ODC-By 1.0 with the note that users "should also abide by the CommonCrawl ToU" [21], and the Tulu 3 mixture is ODC-BY-1.0 while "different licenses apply to subsets of the data. Some portions of the dataset are non-commercial" [25].

Web consent signals changed quickly. An audit of 14,000 web domains behind corpora such as C4, RefinedWeb and Dolma found that "in a single year (2023-2024)" restrictions rendered "~5%+ of all tokens in C4, or 28%+ of the most actively maintained, critical sources in C4, fully restricted from use" [32]. Under Terms of Service crawling restrictions, "a full 45% of C4 is now restricted", and the authors found inconsistencies between sites' Terms of Service and their robots.txt [32]. Common Crawl's crawler identifies itself as CCBot/2.0, and its page tells site owners to block it with a robots.txt rule for CCBot [8].

Code corpora set their own rules. The Stack v2 detected licenses file by file where no repository-level license was available, for 96.93% of repositories, included permissively licensed and unlicensed files and excluded commercial and copyleft licenses [20]. Developers could opt out until 20 November 2023, which removed 1,561 repositories of 91 users and organizations and 22,066 files [20].

Two corpora are built only from open text.

- **The Common Pile v0.1** is "an eight terabyte collection of openly licensed text" from 30 sources, among them research papers, code, books, encyclopedias, educational materials and audio transcripts [33]. "Openly licensed" follows the Open Definition 2.1, with CC BY, CC BY-SA and the MIT license among the examples [33]. The authors warn of license laundering, where "a copyrighted work is redistributed (typically by a non-rights holder) with an incorrect license", and left out OpenAlex, YouTube Commons and a Hacker News dataset [33]. Its two 7B models, trained on 1 and 2 trillion tokens, "attain competitive performance to LLMs trained on unlicensed text with similar computational budgets" [33].
- **Common Corpus**, presented at ICLR 2026, holds data that "are either uncopyrighted or under open licenses, totaling about two trillion tokens", from high-resource European languages to some low-resource languages, plus a large amount of code [34].

This finding reports what the papers, cards and audits say. It draws no conclusion about the legality of any corpus.

9. What the EU template asks providers to disclose

Article 53(1)(d) of the EU AI Act requires providers of general-purpose AI models to "draw up and make publicly available a sufficiently detailed summary about the content used for training", on a template from the AI Office [35]. Article 53(1)(c) requires a policy to comply with Union copyright law, including reservations of

rights under Article 4(3) of Directive (EU) 2019/790 [35]. Recital 107 says the summary should be "generally comprehensive in its scope instead of technically detailed" [35].

The Commission published the explanatory notice and template on 24 July 2025 as "a common minimal baseline for the information to be made publicly available" [36]. Its source types:

Template section	Source type	Detail from the template
2.1	Publicly available datasets	Compiled by a third party, free and readily downloadable, such as snapshots of Common Crawl. "Large" means over 3% of all public datasets for that modality
2.2	Private datasets from third parties	Not publicly available, obtained from third parties
2.3	Data crawled and scraped from online sources	The top 10% of domain names by size of content scraped (for SMEs, the top 5% or 1,000 domain names, whichever is lower)
2.4	User data	Yes or no, for data from user interactions with the model and from the provider's other services
2.5	Synthetic data	Created from the outputs of another model, "in particular through model distillation or model alignment"
2.6	Other sources	For example datasets labelled by humans commissioned by the provider

Source: the template, Section 2 [36]. Section 1.3 gives text size as a range, "Less than 1 billion tokens", "1 billion to 10 trillions tokens" or "More than 10 trillions tokens" [36]. The template notes that free availability "does not mean that the content at issue is necessarily free of rights" [36]. Section 3 covers measures to respect text and data mining opt-outs and the removal of illegal content [36]. The notice says the obligation "becomes applicable as of 2 August 2025", and that providers of models placed on the market before that date should publish their summary no later than 2 August 2027 [36].

10. Duplicates and benchmark contamination

Web corpora contain many copies of the same text. A 2022 study found that datasets "contain many near-duplicate examples and long repetitive substrings", and that "over 1% of the unprompted output of language models trained on these datasets is copied verbatim from the training data" [30]. It removed from C4 "a

single 61 word English sentence that is repeated over 60,000 times" [30]. Models trained on the deduplicated data "emit memorized text ten times less frequently", and train-test overlap "affects over 4% of the validation set of standard datasets" [30].

Benchmarks leak into corpora. The 2021 C4 audit found that C4 contains "evaluation examples from other benchmark NLP datasets", and that the share of GLUE benchmark inputs found in C4.EN "varies widely, from less than 2% to over 50%" [10]. Tulu 3 checked its training prompts against its evaluations with 8-gram matching [4]. A test instance overlapped when "more than 50% of the test tokens have 8-gram matches with the same training instance", a training set counted as contaminated if it overlapped "more than 2% of the instances in any of the evaluations", and training sets contaminated with its held-out evaluations were removed [4]. Llama 3 states: "We do not include any training sets from commonly used benchmarks in our annealing data" [1].

Implications

For teams that build, buy or document training data, the sources point to seven checks.

1. **Ask what each part of a dataset is, by stage and source type.** The EU template's six source types [36] and the stage table in Finding 1 give a shared vocabulary for describing a dataset.
2. **Record version and date with every size.** FineWeb grew from 15 trillion tokens in its paper to more than 18.5 trillion on its card [12, 13].
3. **Keep the license with each subset.** Mixtures carry several licenses, some non-commercial [25], and hosting sites often omit or misstate licenses [31].
4. **Check consent signals when data is captured.** Restrictions on C4's sources rose within a single year [32], and The Stack v2 removed files after developers opted out [20].
5. **Deduplicate, then decontaminate against the evaluations that matter.** Deduplicated training data cut memorized output tenfold [30], and benchmark inputs appear in web corpora [10]. Tulu 3 publishes its thresholds [4].
6. **Record how synthetic data was made.** Cosmopedia's card names its seed samples and the kind of model that generated it [28], and in the Nature study the share of original data kept changed the outcome [29].
7. **Compare sizes in one tokenizer.** Counts depend on the tokenizer and the language [14, 15, 19].

Limits of this analysis

- The report relies on open corpora and open model family reports. Closed models' data mixes are not public, so it makes no claim about what any closed model was trained on.
- Token counts come from different tokenizers. Figures from dataset cards are as viewed on 8 November 2026.
- It did not check which providers have published EU training content summaries.
- License, robots.txt and opt-out findings report what the sources say. They are research, not legal advice.
- The model collapse result comes from controlled experiments on recursive training.
- Disclosure rules outside the EU template are outside its scope.

Sources

1. 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

2. 2 OLMo 2 Furious. Team OLMo, Pete Walsh, Luca Soldaini and others, COLM 2025, arXiv:2501.00656, version 3, 8 October 2025, viewed 8 November 2026. https://arxiv.org/abs/2501.00656

3. Training language models to follow instructions with human feedback. Long Ouyang, Jeff Wu, Xu Jiang and others, arXiv:2203.02155, 4 March 2022, viewed 8 November 2026. https://arxiv.org/abs/2203.02155

4. 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

5. Common Crawl. Common Crawl Foundation, home page, undated, viewed 8 November 2026. https://commoncrawl.org/

6. Overview. Common Crawl Foundation, undated, viewed 8 November 2026. https://commoncrawl.org/overview

7. October 2026 Crawl Archive Now Available. Hande Çelikkanat, Common Crawl Foundation, 21 October 2026, viewed 8 November 2026. https://commoncrawl.org/blog/october-2026-crawl-archive-now-available

8. CCBot. Common Crawl Foundation, undated, viewed 8 November 2026. https://commoncrawl.org/ccbot

9. 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), arXiv:1910.10683, version 4, 19 September 2023, viewed 8 November 2026. https://arxiv.org/abs/1910.10683

10. 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

11. 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

12. 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

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

14. FineWeb2 dataset card. Hugging Face, last modified 27 October 2025, viewed 8 November 2026. https://huggingface.co/datasets/HuggingFaceFW/fineweb-2

15. 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

16. 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

17. RedPajama-Data-V2 dataset card. Hugging Face, last modified 21 November 2024, viewed 8 November 2026. https://huggingface.co/datasets/togethercomputer/RedPajama-Data-V2

18. 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

19. 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

20. StarCoder 2 and The Stack v2: The Next Generation. Anton Lozhkov, Raymond Li, Loubna Ben Allal and others, arXiv:2402.19173, 29 February 2024, viewed 8 November 2026. https://arxiv.org/abs/2402.19173

21. OpenWebMath dataset card. Hugging Face, last modified 17 October 2023, viewed 8 November 2026. https://huggingface.co/datasets/open-web-math/open-web-math

22. 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

23. 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

24. Will we run out of data? Limits of LLM scaling based on human-generated data. Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim and Marius Hobbhahn, arXiv:2211.04325, version 2, 4 June 2024, viewed 8 November 2026. https://arxiv.org/abs/2211.04325

25. Tulu 3 SFT mixture dataset card. Hugging Face, last modified 2 December 2024, viewed 8 November 2026. https://huggingface.co/datasets/allenai/tulu-3-sft-mixture

26. UltraFeedback dataset card. Hugging Face, last modified 29 December 2023, viewed 8 November 2026. https://huggingface.co/datasets/openbmb/UltraFeedback

27. Self-Instruct: Aligning Language Models with Self-Generated Instructions. Yizhong Wang, Yeganeh Kordi, Swaroop Mishra and others, ACL 2023, arXiv:2212.10560, version 2, 25 May 2023, viewed 8 November 2026. https://arxiv.org/abs/2212.10560

28. Cosmopedia dataset card. Hugging Face, last modified 12 August 2024, viewed 8 November 2026. https://huggingface.co/datasets/HuggingFaceTB/cosmopedia

29. AI models collapse when trained on recursively generated data. Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot and others, Nature 631, 755, 24 July 2024, viewed 8 November 2026. https://www.nature.com/articles/s41586-024-07566-y

30. Deduplicating Training Data Makes Language Models Better. Katherine Lee, Daphne Ippolito, Andrew Nystrom and others, ACL 2022, arXiv:2107.06499, version 2, 24 March 2022, viewed 8 November 2026. https://arxiv.org/abs/2107.06499

31. 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

32. 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

33. The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text. Nikhil Kandpal, Brian Lester, Colin Raffel and others, arXiv:2506.05209, 5 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2506.05209

34. Common Corpus: The Largest Collection of Ethical Data for LLM Pre-Training. Pierre-Carl Langlais, Pavel Chizhov, Catherine Arnett and others, ICLR 2026, arXiv:2506.01732, version 4, 5 October 2026, viewed 8 November 2026. https://arxiv.org/abs/2506.01732

35. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Article 53 and recital 107. Official Journal of the European Union, OJ L, 12 July 2024 (regulation of 13 June 2024), via EUR-Lex, viewed 8 November 2026. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

36. Explanatory Notice and Template for the Public Summary of Training Content for general-purpose AI models. European Commission, AI Office, published 24 July 2025, English notice C(2025) 8311 final of 5 December 2025, page last updated 26 March 2026, viewed 8 November 2026. https://digital-strategy.ec.europa.eu/en/library/explanatory-notice-and-template-public-summary-training-content-general-purpose-ai-models

About this report

Published by Rillor on 11 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. What data is used to train language models. 11 November 2026. https://rillor.com/insights/what-data-is-used-to-train-language-models

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