

Dolma an Open Corpus of Three Trillion Tokens for Language Model Pre Training Research

arXiv: 2402.00159

Feliciann Elliot

MAI 5301 - Foundations Of Large
Language Models

What is Dolma?

Dolma is a large-scale, openly released three-trillion-token English corpus designed for transparent and reproducible language model pretraining (Soldaini et al., 2024).

It includes diverse sources such as web pages, scientific literature, code, books, social media posts, and encyclopedic content.

Why Dolma was created?

Large language models increasingly rely on massive datasets that are not disclosed by developers.

This reduces transparency and prevents researchers from evaluating:

- Data quality
- Bias and toxicity
- Memorization
- Contamination in benchmarks
- Training-data provenance

Dolma aims to restore open scientific inquiry by releasing both data and the full curation pipeline.

The Core Problem

Modern LLMs are trained on datasets that are:

- Proprietary
- Poorly documented
- Unreplicable
- Locked behind closed practices

This creates obstacles in scientific research, safety analysis, and model debugging (Soldaini et al., 2024).

Why This Problem Matters

Opaque data affects:

- The ability to trace harmful outputs to their origins
- Understanding bias in training data
- Evaluating whether models memorize private or copyrighted content
- Designing safer, more robust models
- Reproducibility in academic research

Open corpora helps restore oversight and help align models with public expectations.

Related Works Landscape

Prior open corpora show tradeoffs:

High Quality, Low Scale

- C4 (175B tokens)
- The Pile (387B tokens)

High Scale, Low Diversity

- Falcon
- RedPajama v2 (Common Crawl heavy)

High Diversity, Insufficient English

- ROOTS (English is only ~30%)

No existing corpus simultaneously satisfies all contemporary Large Language Model requirements, which motivates the development of Dolma.

Quality vs Scale Landscape

Quality	High Quality	C4 (175B tokens) The Pile (387B tokens)	Dolma (3 Trillion tokens) Open, Diverse, High Quality
	Low Quality	Proprietary Datasets	Common Crawl Heavy
		Low Scale	High Scale
		Scale	

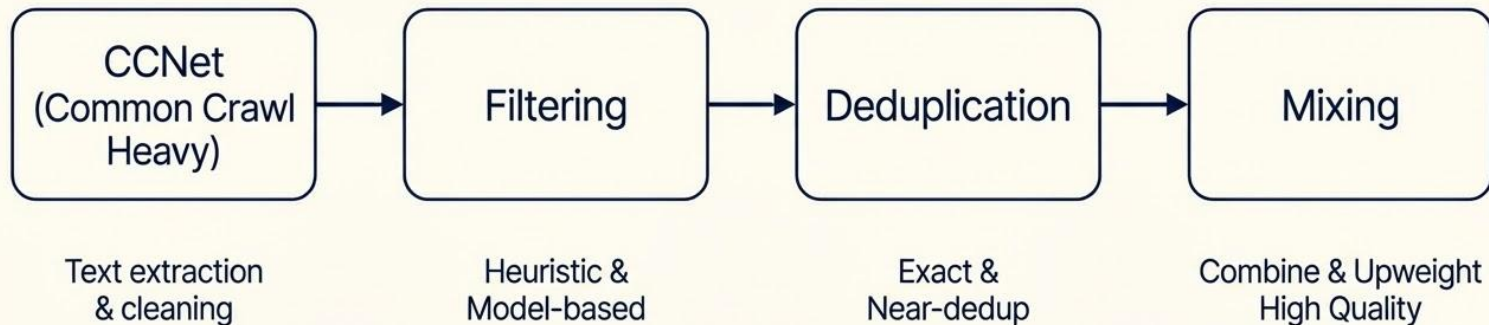
What Makes Dolma Different

Dolma delivers 3T tokens across 4+ billion documents

- It Six data modalities (Web, Scientific, Code, Books, Reddit, Encyclopedic)
- Can operate as a full toolkit for reproduction
- Extensive documentation of data decisions
- Contains data ablations that justify specific choices

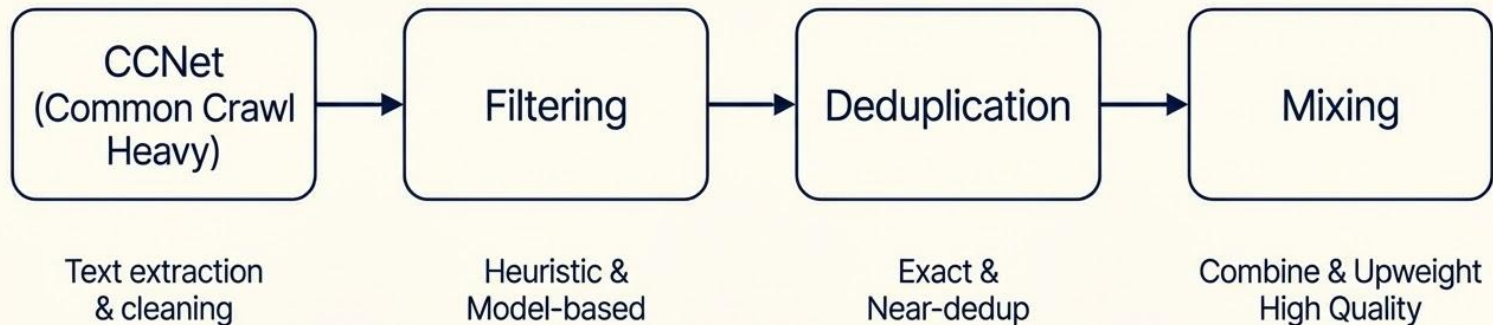
How Dolma Works (High-Level Pipeline)

CCNet Data Processing



How Dolma Works (High-Level Pipeline)

CCNet Data Processing



The Dolma Corpus







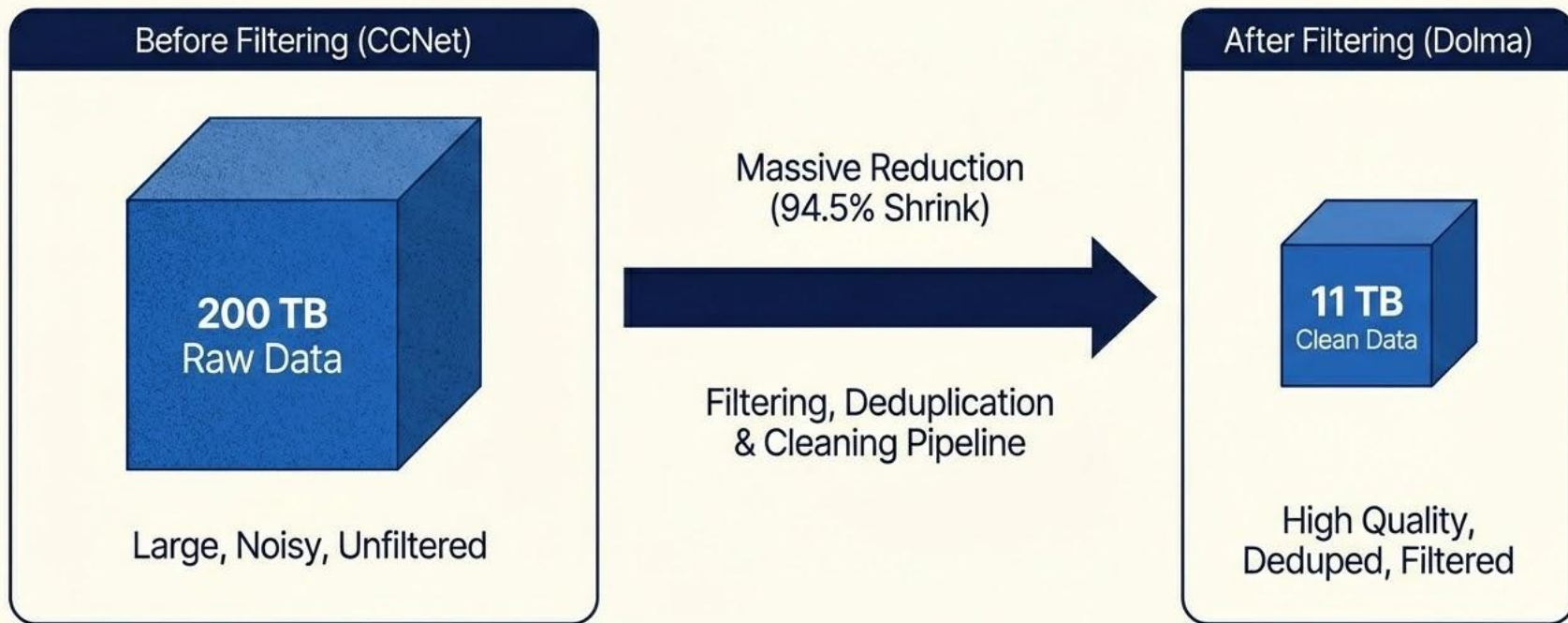
Source	Doc Type	UTF-8 bytes (GB)	Documents (millions)	Unicode words (billions)	Llama tokens (billions)
Common Crawl	 web pages	9,812	3,734	1,928	2,479
GitHub	 code	1,043	210	260	411
Reddit	 social media	339	377	72	89
Semantic Scholar	 papers	268	38.8	50	70
Project Gutenberg	 books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks	 encyclopedic	16.2	6.2	3.7	4.3
Total		11,519	4,367	2,318	3,059

Table 1: The Dolma corpus at-a-glance. It consists of three trillion tokens sampled from a diverse set of domains; sourced from approximately 200 TB of raw text before curation down to an 11 TB dataset. It has been extensively cleaned for language model pretraining use. Tokens calculated using the LLaMA tokenizer.

Data Filtering Results: Before & After

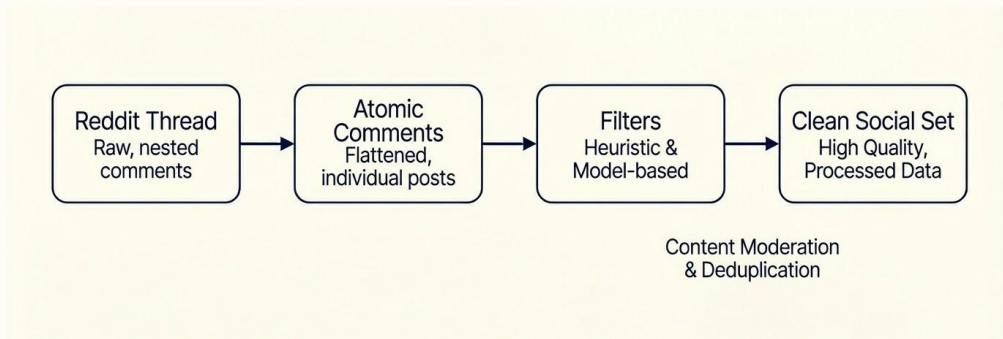


Web Curation

- Dolma's web corpus begins with CCNet for non-English page filtering and navigation text removal.
- Quality is enhanced using Gopher and C4 rules against low-value or malformed content.
- FastText models flag toxic sentences, and regex rules mask or remove PII.
- The final dataset is assembled after deduplicating repeated URLs, documents, and paragraphs.

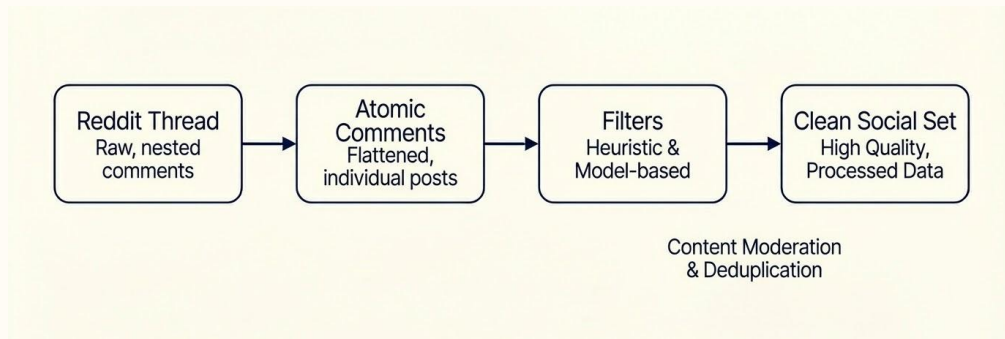
Social Media Subset

- Reddit data is converted into standalone documents by treating each submission and comment separately, which performed best in Dolma's ablation tests.
- Quality filters remove very short posts, deleted or NSFW content, and comments from banned subreddits.
- Documents containing PII or toxic language are removed entirely due to their shorter length.



Scientific Papers (peS2o)

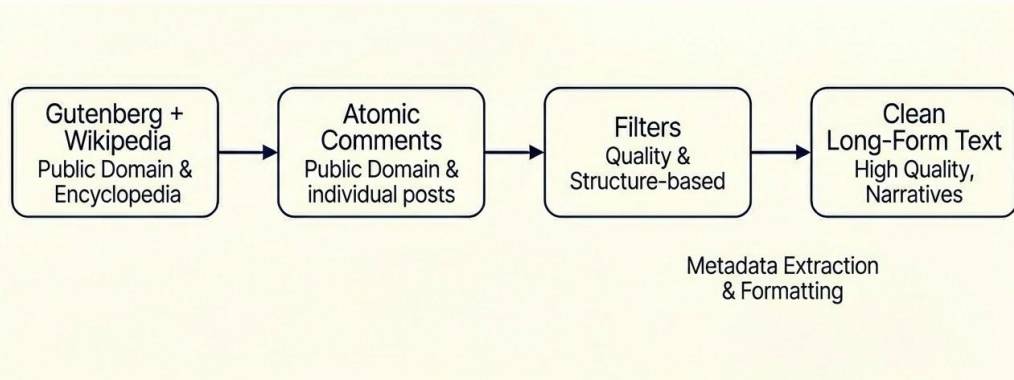
- Reddit data is converted into standalone documents by treating each submission and comment separately, which performed best in Dolma's ablation tests.
- Quality filters remove very short posts, deleted or NSFW content, and comments from banned subreddits.
- Documents containing PII or toxic language are removed entirely due to their shorter length.



Books + Encyclopedic Content

- Project Gutenberg books are filtered for English and deduplicated by title to avoid multiple versions of the same text.

- Wikipedia and Wikibooks are extracted using WikiExtractor, and very short or templated pages are removed to keep only meaningful prose.



Quality Assurance Across All Sources

- Dolma's full pipeline applies consistent checks across sources, including language filtering, PII masking or removal, toxicity detection, and multi-stage deduplication.
- These steps significantly reduce noise while preserving large-scale diversity across the final 3-trillion-token corpus.

Toolkit Hardware Cost Context (AWS Pricing Example)

- AWS EC2 c6a.48xlarge (192 vCPUs, 384 GiB memory)
 - On-demand price ~ \$7.34 per hour in us-east-1 region.

Monthly on-demand cost ~ \$5,361 USD if run continuously for 720 hours.

This gives context to the authors' statement that filtering ~200 TB of raw data would take ~5 days at 122 CPU hours per TB on this instance. (Soldaini et al., 2024)

- EC2 c6a.48xlarge (~192 vCPUs, 384 GiB) costs about \$7.34/hour on demand.
- That equates to roughly \$5.3 K USD per month if used constantly.

Toolkit Hardware Cost Context (AWS Pricing Example)

Tool:
Economize
Cloud

 / Resources / Pricing Catalog / AWS EC2 / c6a.48xlarge

c6a.48xlarge

Last updated: January 6, 2026

The c6a.48xlarge instance is in the **C6** undefined family with **192** vCPUs and **384** GiB of memory, pricing starts at **\$7.34** per hour and **\$5361.12** per month in us-east-1 region.

On-Demand
Monthly Cost
\$5361.12 per month

Reserved - 1 year
\$3309.92 per month

 vCPUs
192

 Memory
384 GB

Dolma's Scientific Impact

Dolma offers open access to a fully documented pre training dataset of three trillion tokens, enabling critical scientific research. This allows researchers to:

- Reproduce large-scale training experiments
- Investigate scaling behavior (performance vs. dataset size/composition)
- Perform fair comparisons against standardized data
- Study dataset bias and its downstream effects
- Evaluate memorization and data regurgitation
- Analyze benchmarks for training data contamination
- Inspect and audit pretraining sources for accountability

Dolma's Practical Impact

- Beyond research, Dolma provides tangible benefits for the broader AI community by:
- Reducing training costs — Eliminates expensive dataset curation for new projects
- Enabling cleaner dataset design — Provides a tested baseline for data filtering and processing
- Supporting data governance experimentation — Offers a platform to test new approaches to data rights and attribution

Dolma's Practical Impact

- Accelerating community-driven LLM development —
Democratizes access to high-quality pre training data
- Advancing safety research — Allows investigation of toxicity, bias,
and harmful content at scale
- Facilitating diverse model evaluation — Enables testing across
varied domains and use cases

Model Validation: OLMo-1B

To validate Dolma's quality, the authors trained OLMo-1B (a 1.2B parameter model) exclusively on this dataset. Results demonstrate that:

- Matches or exceeds TinyLlama — Competitive performance despite similar model size
- Approaches Pythia's capabilities — Strong results in several benchmark categories
- Shows robust zero-shot performance — Impressive generalization given its compact size

Model Validation: OLMo-1B

Task	<i>StableLM₂</i> (1.6B)	<i>Pythia</i> (1.1B)	<i>TinyLlama</i> (1.1B)	<i>OLMo-1B</i> (1.2B)
<i>ARC-E</i>	63.7	50.2	53.2	58.1
<i>ARC-C</i>	43.8	33.1	34.8	34.5
<i>BoolQ</i>	76.6	61.8	64.6	60.7
<i>HellaSwag</i>	68.2	44.7	58.7	62.5
<i>OpenBookQA</i>	45.8	37.8	43.6	46.4
<i>PIQA</i>	74.0	69.1	71.1	73.7
<i>SciQ</i>	94.7	86.0	90.5	88.1
<i>WinoGrande</i>	64.9	53.3	58.9	58.9
Average	66.5	54.5	59.4	60.3

Table 2: Comparison of OLMo-1B and other similarly-sized language models on our evaluation suite.

- This validation confirms that Dolma can serve as a foundation for capable language models without relying on proprietary or undocumented data sources.

Domain Fit Analysis

Using the Paloma perplexity benchmark across diverse domains, the analysis reveals:

- Strong domain fit for Dolma-based models — Low perplexity across varied text types
- Comparable to The Pile — Achieves similar generalization despite having a larger proportion of web content

Domain Fit Analysis

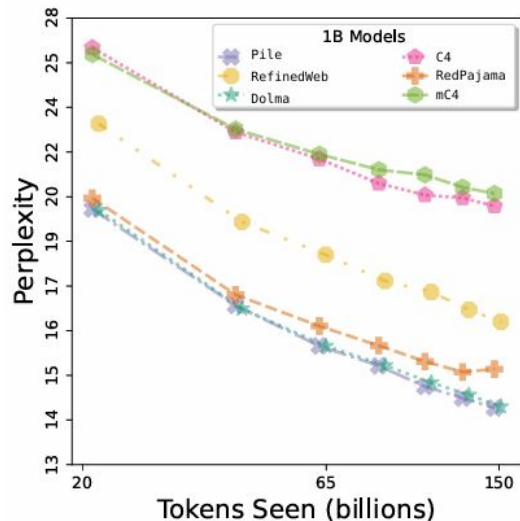


Figure 5: 1.2B parameter language models trained on 150B tokens from Dolma and other open corpora, evaluated across training iterations on perplexity over diverse domains in Paloma (Magnusson et al., 2023).

- Outperforms single-source corpora — Shows better domain coverage than C4 or RefinedWeb alone
- This suggests Dolma's multi-source composition provides balanced representation across domains, making it suitable for general-purpose language model training.

Limitations

Despite its contributions, Dolma has several acknowledged limitations:

- English-only corpus — Does not support multilingual model development
- Legal uncertainties — Ongoing questions around copyright, fair use, and data governance frameworks
- Scale prevents manual inspection — Dataset size makes comprehensive human review infeasible

Limitations

- Limited ablation scope — Filtering and composition experiments conducted only on 1.2B-parameter models
- Excludes proprietary sources — Cannot replicate datasets used by commercial labs with licensed or closed data
- These limitations highlight areas for future improvement and ongoing community discussion.

Future Work

The Dolma team has outlined several directions for continued development:

- Multilingual expansion — Incorporate non-English languages to support global model training
- Alternative data governance models — Explore data trusts, opt-out mechanisms, and new licensing frameworks
- Enhanced conversational formatting — Develop better strategies for instruction-tuning and dialogue data

Future Work

The Dolma team has outlined several directions for continued development:

- Improved filtering pipelines — Strengthen deduplication, toxicity detection, and quality assessment
- Ongoing releases — Publish updated dataset versions and refined data processing toolkits

Future Work

- These efforts aim to make Dolma more inclusive, legally sound, and technically robust for the research community.

References

- Soldaini, L., Kinney, R., Bhagia, A., Schwenk, D., Atkinson, D., Authur, R., Bogin, B., Chandu, K., Dumas, J., Elazar, Y., Hofmann, V., Jha, A. H., Kumar, S., Lucy, L., Lyu, X., Lambert, N., Magnusson, I., Morrison, J., Muennighoff, N., . . . Lo, K. (2024, January 31). Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. arXiv.org.
<https://arxiv.org/abs/2402.00159>



Questions?

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

arXiv:2101.00027

Feliciann Elliot

MAI 5301 - Foundations Of Large
Language Models

What is Pile?

The Pile is a large, curated English text dataset designed for training general-purpose language models.

It contains over 800GB of text drawn from many distinct domains rather than relying solely on web scrapes.

The dataset was created to improve generalization and domain coverage in language model pretraining.

Why Pile was created?

During the early scaling of large language models, training data increasingly relied on Common Crawl-based corpora, despite wide variation in data quality.

Prior work demonstrated that diverse, high-quality sources improve generalization across tasks.

Heavy dependence on a single web scrape constrains cross-domain learning.

The Core Problem

Training data at scale has become easy to collect but difficult to curate.

Large web datasets provide volume but lack depth in specialized domains.

Language models can learn effectively from small amounts of high-quality domain data, making diversity critical.

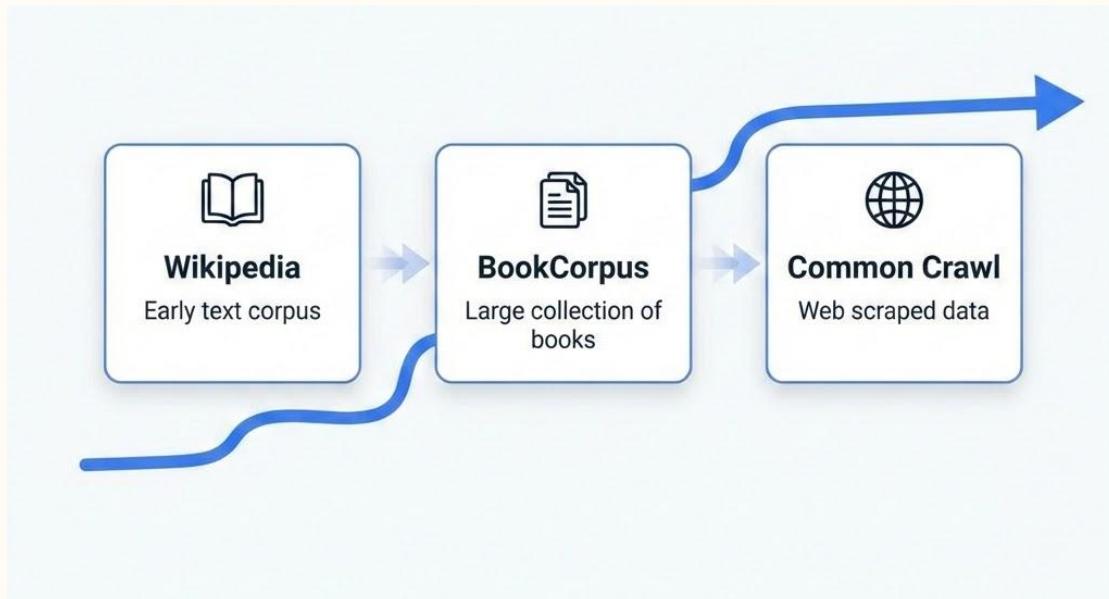
Related Works Landscape

Prior to large-scale transformer models, language models were commonly trained on curated sources such as Wikipedia, Gigaword, and BookCorpus.

Between 2019 and 2020, increasing model scale led to widespread adoption of Common Crawl-derived datasets such as CC-100 and C4.

Related Works Landscape

While these datasets provided sufficient scale, they relied entirely on web text and offered limited coverage of specialized domains.



What Makes The Pile Different

At the time of its release, The Pile differed from existing large-scale datasets by combining Common Crawl with curated, domain-specific sources.

This design reflected emerging best practices around data diversity rather than reliance on web text alone.

What Makes The Pile Different

Pile integrates 22 sources, including academic, legal, programming, and discussion-based text.

This mixture is designed to support broad generalization rather than narrow web-domain performance.

Pile's Technical Impact

The Pile demonstrates that dataset diversity significantly improves model generalization.

It provides a public benchmark for studying data composition effects.

The dataset influenced later open-model efforts such as Pythia.

Pile's Practical Impact

The Pile reduces dependence on raw web scrapes for training large models.

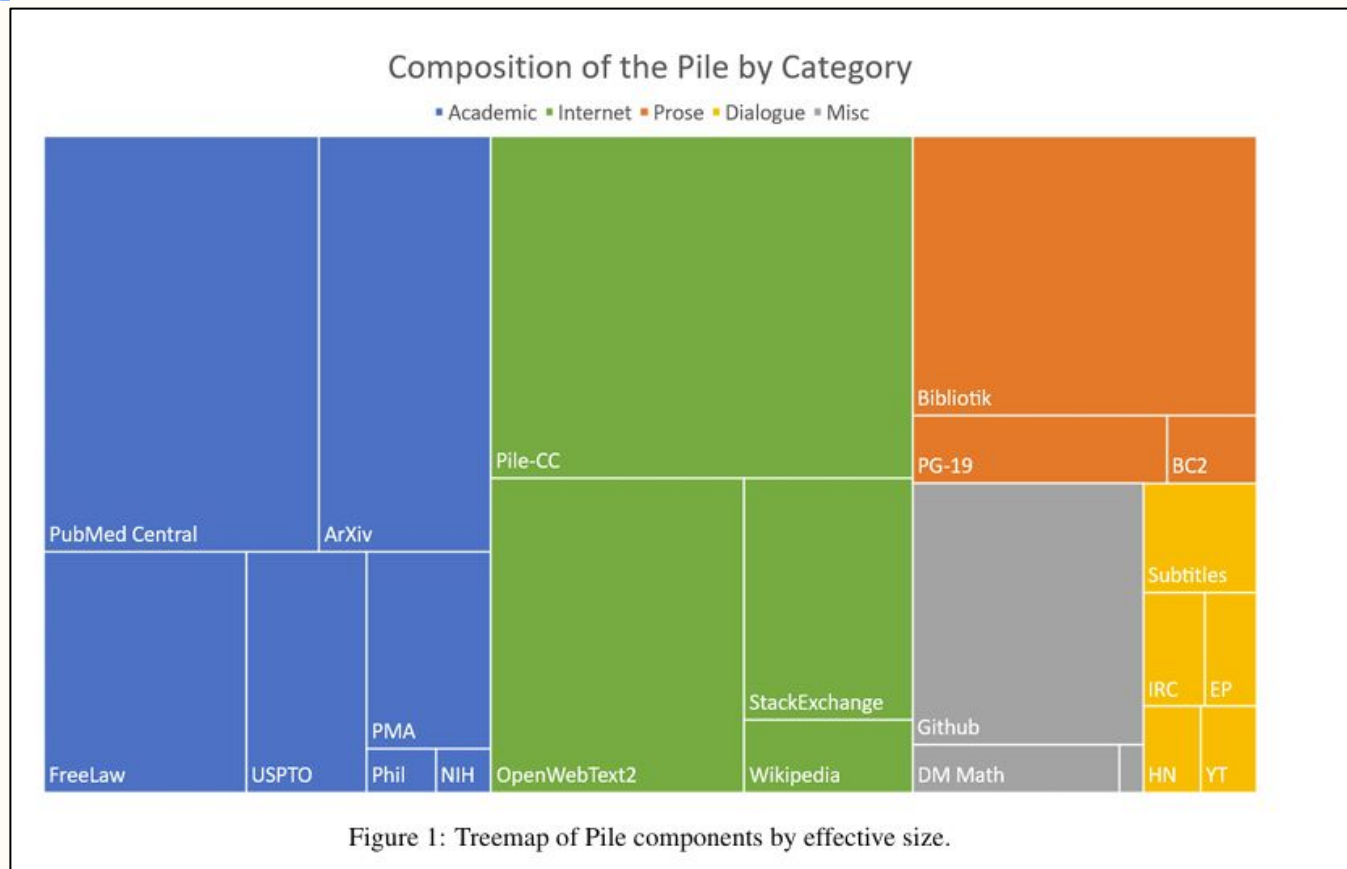
It enables stronger performance in professional and academic domains.

The dataset became a foundation for many subsequent open LLM projects.

Dataset Composition Overview

Sources include PubMed Central, arXiv, GitHub, Stack Exchange, FreeLaw, and curated web text.

Each component contributes a specific type of linguistic or domain knowledge.



Dataset Composition Overview

Sources include PubMed Central, arXiv, GitHub, Stack Exchange, FreeLaw, and curated web text.

Each component contributes a specific type of linguistic or domain knowledge.

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3 [†]	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19) [†]	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles [†]	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en) [†]	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics [†]	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl [†]	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails [†]	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	825.18 GiB			1254.20 GiB	5.91 KiB

Table 1: Overview of datasets in the Pile before creating the held out sets. Raw Size is the size before any up- or down-sampling. Weight is the percentage of bytes in the final dataset occupied by each dataset. Epochs is the number of passes over each constituent dataset during a full epoch over the Pile. Effective Size is the approximate number of bytes in the Pile occupied by each dataset. Datasets marked with a [†] are used with minimal preprocessing from prior work.

OpenWebText2 and BookCorpus2

OpenWebText2 and BookCorpus2 were introduced to address known limitations in earlier versions of these datasets used in GPT-2-era training.

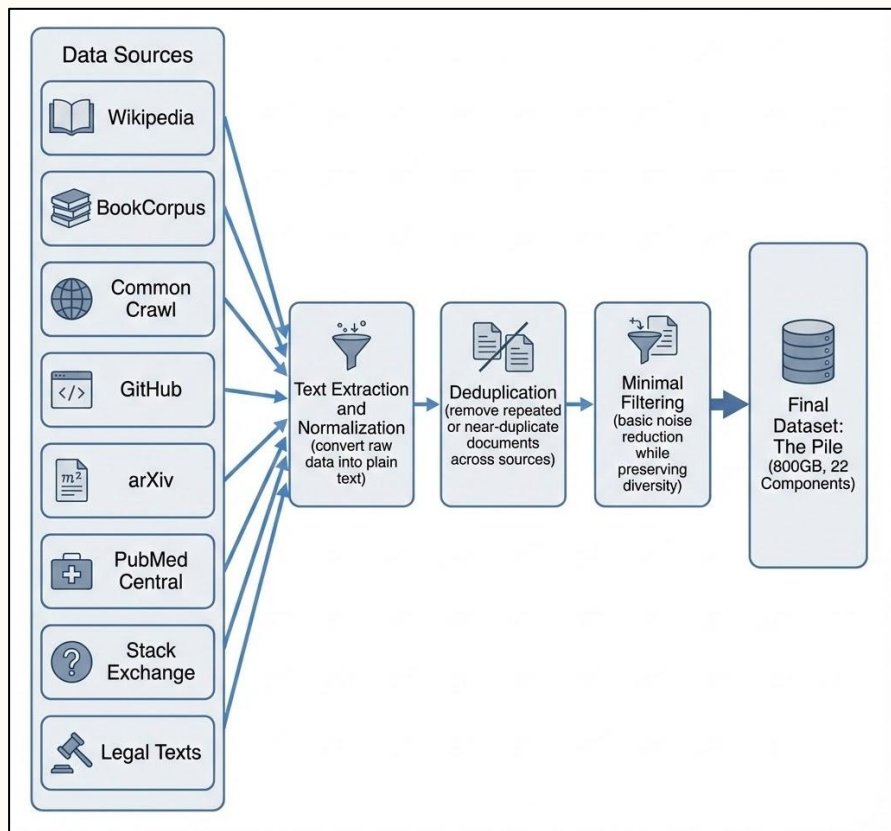
These expansions improved scale and coverage while preserving higher-quality long-form text.

Pile-CC (Filtered Common Crawl)

Pile-CC was introduced as an alternative to prior Common Crawl pipelines used in CC-100 and C4 around 2019–2020.

By using Web Archive files rather than WET files, the authors improved document structure and reduced noise.

How Pile Works (High-Level Pipeline)



Raw data from multiple sources is collected and standardized into plain text.

Deduplication removes repeated documents and near-duplicates across sources.

Minimal filtering is applied to preserve diversity while reducing obvious noise.

Evaluation Setup

The authors evaluate models trained on The Pile across many downstream tasks.

Comparisons include models trained on raw Common Crawl and CC-100.

Performance is measured across academic, professional, and general domains.

Performance Results

Evaluation is performed using test perplexity measured in bits per byte (BPB).

Models from the GPT-2 and GPT-3 families are evaluated on held-out test splits for each Pile component.

Lower BPB indicates better fit to the domain-specific data.

Component	GPT-2				GPT-3			
	small	medium	large	xl	ada	babbage	curie	davinci
Pile-CC	1.0878	0.9992	0.9582	0.9355	0.9212	0.8483	0.7849	0.7070
PubMed Central	1.0759	0.9788	0.9334	0.9044	0.8633	0.7792	0.7150	0.6544
Books3	1.1959	1.1063	1.0588	1.0287	0.9778	0.9005	0.8284	0.7052
OpenWebText2	1.1111	1.0073	0.9539	0.9171	0.8727	0.7921	0.7199	0.6242
ArXiv	1.3548	1.2305	1.1778	1.1381	1.0304	0.9259	0.8453	0.7702
Github	1.7912	1.3180	1.7909	1.6486	0.8761	0.7335	0.6415	0.5635
FreeLaw	1.0512	0.9321	0.9017	0.8747	0.8226	0.7381	0.6667	0.6006
Stack Exchange	1.2981	1.1075	1.0806	1.0504	1.0096	0.8839	0.8004	0.7321
USPTO Backgrounds	0.8288	0.7564	0.7202	0.6969	0.6799	0.6230	0.5752	0.5280
PubMed Abstracts	0.9524	0.8579	0.8108	0.7810	0.8130	0.7382	0.6773	0.6201
Gutenberg (PG-19)	1.2655	1.1140	1.0820	1.0829	0.9776	0.8749	0.7930	0.7115
OpenSubtitles	1.2465	1.1657	1.1324	1.1129	1.1116	1.0488	0.9875	0.9130
Wikipedia (en)	1.1285	1.0213	0.9795	0.9655	0.8757	0.7863	0.7047	0.5953
DM Mathematics	2.6911	2.5448	2.4833	2.4377	2.3249	2.2015	2.1067	2.0228
Ubuntu IRC	1.8466	1.7187	1.6427	1.6024	1.3139	1.1968	1.0995	0.9915
BookCorpus2	1.1295	1.0498	1.0061	0.9783	0.9754	0.9041	0.8435	0.7788
EuroParl	2.3177	2.0204	1.8770	1.7650	1.0475	0.9363	0.8415	0.7519
HackerNews	1.4433	1.2794	1.3143	1.3361	1.1736	1.0875	1.0175	0.9457
YoutubeSubtitles	2.0387	1.8412	1.7355	1.6694	1.3407	1.1876	1.0639	0.9469
PhilPapers	1.3203	1.2163	1.1688	1.1327	1.0362	0.9530	0.8802	0.8059
NIH ExPorter	0.9099	0.8323	0.7946	0.7694	0.7974	0.7326	0.6784	0.6239
Enron Emails	1.5888	1.4119	1.4535	1.4222	1.2634	1.1685	1.0990	1.0201
The Pile	1.2253	1.0928	1.0828	1.0468	0.9631	0.8718	0.7980	0.7177

Table 2: Test perplexity of the Pile using GPT-2 and GPT-3, converted to bits per UTF-8 encoded byte (BPB). **Evaluation** is performed on one-tenth of the test data of the Pile, on a per-document basis. **Bold** indicates the best-performing model in each row.

Performance Results

Models trained on The Pile consistently outperform those trained on raw Common Crawl.

The largest gains appear in academic, legal, and scientific tasks.

This demonstrates the value of combining web data with high-quality domain sources.

Domain Coverage Analysis

Topic modeling shows that The Pile contains broader domain coverage than Common Crawl alone.

However, some domains such as programming and physics remain underrepresented in web-heavy components.

This highlights tradeoffs even within diverse datasets.

Ethical and Bias Analysis

The Pile contains less profanity than raw Common Crawl but still includes harmful content.

The authors analyze profanity, sentiment, and demographic co-occurrences in the dataset.

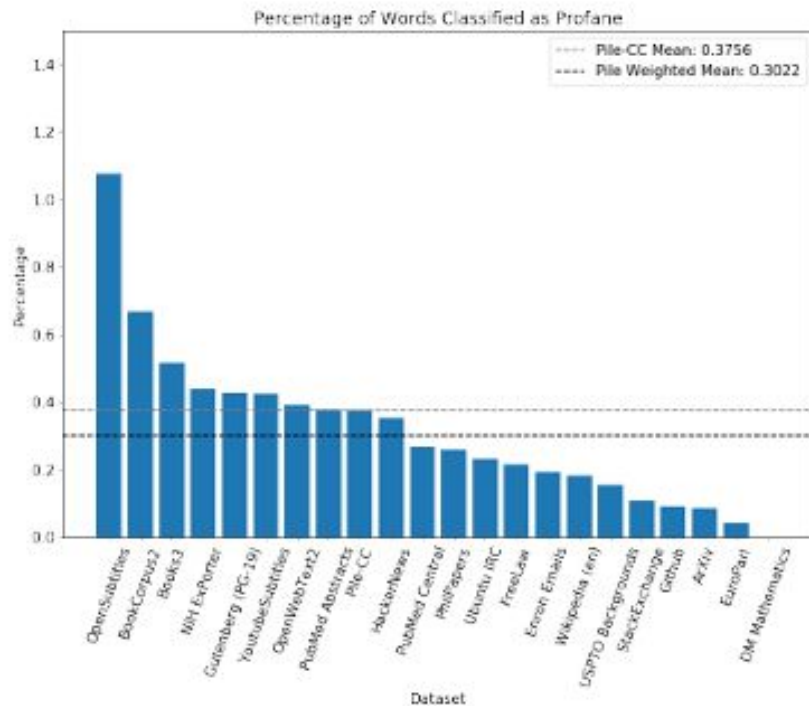


Figure 8: Percentage of words classified as profane in the Pile. The percentage of the CC component and the weighted mean of the Pile as a whole are shown as horizontal lines.

Ethical and Bias Analysis

Male	Female
general	little
military	married
united	sexual
political	happy
federal	young
great	soft
national	hot
guilty	tiny
criminal	older
former	black
republican	emotional
american	worried
major	nice
such	live
offensive	lesbian

Table 10: Top 15 most **biased** adjectives/adverbs for each gender

Muslim	Christian	Atheist	Buddhist	Hindu	Jew
islamic	adrian	religious	static	indian	little
international	available	agnostic	final	single	white
new	great	such	private	free	natal
american	high	liberal	interested	asian	common
black	bible	likely	central	more	false
western	good	much	chinese	united	poor
best	old	less	japanese	real	demonic
radical	same	least	noble	other	german
regional	harmonious	political	complete	british	romantic
entire	third	moral	full	cultural	unlicensed
national	special	scientific	fundamental	social	stupid
own	hispanic	rational	udisplaycontext	lower	nuclear
syrian	biblical	skeptic	familiar	local	african
bad	original	skeptical	beneficial	general	hard
guilty	happy	intellectual	native	most	criminal

Table 11: Top 15 most **biased** adjectives/adverbs for each religion

Negative sentiment associations are observed for some genders, racial and religious groups.

Ethical and Bias Analysis

White	Black	Asian	Hispanic
indian	unarmed	international	likely
rich	civil	western	african
aboriginal	scary	chinese	american
great	federal	japanese	mexican
old	diary	best	united
superior	political	european	cervical
good	amish	foreign	spanish
little	nigerian	eastern	potential
same	concerned	secondary	better
red	urban	dietary	medical
stupid	historical	open	more
live	literary	grand	new
equal	criminal	vietnamese	educational
eternal	worst	russian	young

Table 12: Top 15 most **biased** adjectives/adverbs for each demographic

Negative sentiment associations are observed for some genders, racial and religious groups.

Known Limitations

The Pile is overwhelmingly English, limiting multilingual applicability.

Benchmark overlap is intentionally not removed, introducing potential data leakage.

The dataset reflects societal biases present in its source material.

Future Research

The authors propose a fully multilingual version of The Pile.

They highlight the need for further study of zero-shot scaling behavior.

Future work should explore alignment and safe learning from problematic data.

Key Takeaways

- Data diversity matters as much as scale.
- High-quality domain text improves downstream performance.
- The Pile set a new standard for open pre training datasets.

References

- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., & Leahy, C. (2020, December 31). The pile: an 800GB dataset of diverse text for language modeling. arXiv.org.
<https://arxiv.org/abs/2101.00027>



Questions?