

Pre-training Data Curation

Sewon Min
09/01/2026

Acknowledgement: Some slides adapted from Kyle Lo (Ai2)

Today's topic: Pre-training data

- Pre-training: The very first step in LLM training

Today's topic: Pre-training data

- Pre-training: The very first step in LLM training
- Goals for data:
 - Quantity: Bigger is better – no labels needed, raw text only
 - Llama 3: >15 trillion tokens
 - Llama 4: >30 trillion tokens
 - (Reference: Wikipedia is just 3 billion tokens)
 - Diversity: Broad enough to reflect nearly all possible NLP tasks
 - Quality: Gibberish text won't be helpful

Today's topic: Pre-training data

- Pre-training: The very first step in LLM training
- Goals for data:
 - Quantity: Bigger is better – no labels needed, raw text only
 - Llama 3: >15 trillion tokens
 - Llama 4: >30 trillion tokens
 - (Reference: Wikipedia is just 3 billion tokens)
 - Diversity: Broad enough to reflect nearly all possible NLP tasks
 - Quality: Gibberish text won't be helpful
- Today's goal: Not only understand the data curation process but also build intuition for dataset scales across model generations (“what data scale are we talking about?”) and be familiar with common concepts.

Agenda

1. History
2. Overview of pre-training data curation process
3. Deep dive
 - a. DCLM paper
 - b. FineWeb paper
 - c. DCLM vs. FineWeb
4. Summary/Discussion

History of pre-training data

Pre-training (as it is used today) in 2018

- English Wikipedia (2.5B words) and books (~1B words) as popular choices for large text corpora
- ELMo [Peters et al. 2018] “is pretrained on a large text corpus” [...] on the 1B Word Benchmark

Pre-training (as it is used today) in 2018

- English Wikipedia (2.5B words) and books (~1B words) as popular choices for large text corpora
- ELMo [Peters et al. 2018] “is pretrained on a large text corpus” [...] on the 1B Word Benchmark
 - “[T]he 1B Word Benchmark, which is used by [ELMo] is approximately the same size but is shuffled at a sentence level - destroying long-range structure.” – GPT-1 [Radford et al. 2018]
 - “It is critical to use a document-level corpus rather than a shuffled sentence-level corpus such as the [1B Word Benchmark] to extract long contiguous sequences.” – BERT [Devlin et al. 2018]

Pre-training (as it is used today) in 2018

- English Wikipedia (2.5B words) and books (~1B words) as popular choices for large text corpora
- ELMo [Peters et al. 2018] “is pretrained on a large text corpus” [...] on the 1B Word Benchmark
 - “[T]he 1B Word Benchmark, which is used by [ELMo] is approximately the same size but is shuffled at a sentence level - destroying long-range structure.” – GPT-1 [Radford et al. 2018]
 - “It is critical to use a document-level corpus rather than a shuffled sentence-level corpus such as the [1B Word Benchmark] to extract long contiguous sequences.” – BERT [Devlin et al. 2018]
- GPT-1 [Radford et al. 2018]: pretrained on the BooksCorpus (800M words)
- BERT [Devlin et al. 2018]: pretrained on the BooksCorpus (800M words) and English Wikipedia (2,500M words).

GPT-2 [Radford et al. 2019] – WebText

- “Most prior work trained language models on a single domain of text [...] Our approach motivates building as large and diverse a dataset as possible in order to collect natural language demonstrations of tasks in as varied of domains and contexts as possible.”
- Web pages!
 - 1) Scraped web pages that are outbound links from Reddit w/ at least 3 karma.
 - 2) Extract text from HTML (Dragnet and Newspaper content extractors)
 - 3) De-duplication and some heuristic based cleaning → 8M+ docs
- Didn't release the raw data, but open-source efforts replicated it – [OpenWebText \(2019\)](#)

Common Crawl

Common Crawl

- A publicly-available web archive w/o markup and other non-text content from the scraped HTML files (~20TB/month)
 - “A promising source of diverse and nearly unlimited text is web scrapes such as Common Crawl.” – GPT-2 [Radford et al. 2019]

Common Crawl

- A publicly-available web archive w/o markup and other non-text content from the scraped HTML files (~20TB/month)
 - “A promising source of diverse and nearly unlimited text is web scrapes such as Common Crawl.” – GPT-2 [Radford et al. 2019]
- The majority of text is not even natural language text
 - “[Common Crawl’s] content are mostly unintelligible” – Trinh & Le (2018)
 - “It largely comprises gibberish or boiler-plate text like menus, error messages, or duplicate text”, “[Most] is unlikely to be helpful for any of the tasks we consider (offensive language, placeholder text, etc.). “ – T5 [Raffel et al. 2019]

T5 [Raffel et al. 2019] – C4 (Colossal Clean Crawled Corpus)

- Lots of heuristic filters:

- Only retained lines that ended in a terminal punctuation mark (i.e. a period, question mark, etc)
- Discarded any page with <3 sentences and only retained lines that contained >=5 words.
- Removed any page that contained any “List of Dirty, Naughty, Obscene or Otherwise Bad Words”
- Many scraped pages contained warnings stating that Javascript should be enabled → removed any line with the word Javascript.
- Removed any page where the phrase “lorem ipsum” appeared
- The curly bracket “{” appears in many programming languages (such as Javascript, widely used on the web) but not in natural text → removed any pages that contained a curly bracket.
- Some were sourced from Wikipedia and had citation markers → removed any such markers.
- Many pages had boilerplate policy notices → removed any lines containing the strings “terms of use”, “privacy policy”, “cookie policy”, “uses cookies”, “use of cookies”, or “use cookies”.
- Deduplication: discarded all but one of any three-sentence span occurring more than once.
- Used langdetect to filter out any pages that were not classified as English.

T5 [Raffel et al. 2019] – C4 (Colossal Clean Crawled Corpus)

- Lots of heuristic filters:

- Only retained lines that ended in a terminal punctuation mark (i.e. a period, question mark, etc)
- Discarded any page with <3 sentences and only retained lines that contained >=5 words.
- Removed any page that contained any “List of Dirty, Naughty, Obscene or Otherwise Bad Words”
- Many scraped pages contained warnings stating that Javascript should be enabled → removed any line with the word Javascript.
- Removed any page where the phrase “lorem ipsum” appeared
- The curly bracket “{” appears in many programming languages (such as Javascript, widely used on the web) but not in natural text → removed any pages that contained a curly bracket.
- Some were sourced from Wikipedia and had citation markers → removed any such markers.
- Many pages had boilerplate policy notices → removed any lines containing the strings “terms of use”, “privacy policy”, “cookie policy”, “uses cookies”, “use of cookies”, or “use cookies”.
- Deduplication: discarded all but one of any three-sentence span occurring more than once.
- Used langdetect to filter out any pages that were not classified as English.

T5 [Raffel et al. 2019] – C4 (Colossal Clean Crawled Corpus)

- Lots of heuristic filters:

- Only retained lines that ended in a terminal punctuation mark (i.e. a period, question mark, etc)
- Discarded any page with <3 sentences and only retained lines that contained >=5 words.
- Removed any page that contained any “List of Dirty, Naughty, Obscene or Otherwise Bad Words”
- Many scraped pages contained warnings stating that Javascript should be enabled → removed any line with the word Javascript.
- Removed any page where the phrase “lorem ipsum” appeared
- The curly bracket “{” appears in many programming languages (such as Javascript, widely used on the web) but not in natural text → removed any pages that contained a curly bracket.
- Some were sourced from Wikipedia and had citation markers → removed any such markers.
- Many pages had boilerplate policy notices → removed any lines containing the strings “terms of use”, “privacy policy”, “cookie policy”, “uses cookies”, “use of cookies”, or “use cookies”.
- Deduplication: discarded all but one of any three-sentence span occurring more than once.
- Used langdetect to filter out any pages that were not classified as English.

T5 [Raffel et al. 2019] – C4 (Colossal Clean Crawled Corpus)

- Lots of heuristic filters:

- Only retained lines that ended in a terminal punctuation mark (i.e. a period, question mark, etc)
- Discarded any page with <3 sentences and only retained lines that contained >=5 words.
- Removed any page that contained any “List of Dirty, Naughty, Obscene or Otherwise Bad Words”
- Many scraped pages contained warnings stating that Javascript should be enabled → removed any line with the word Javascript.
- Removed any page where the phrase “lorem ipsum” appeared
- The curly bracket “{” appears in many programming languages (such as Javascript, widely used on the web) but not in natural text → removed any pages that contained a curly bracket.
- Some were sourced from Wikipedia and had citation markers → removed any such markers.
- Many pages had boilerplate policy notices → removed any lines containing the strings “terms of use”, “privacy policy”, “cookie policy”, “uses cookies”, “use of cookies”, or “use cookies”.

Filtered 90% of the data, leading to 175B tokens

“Not only orders of magnitude larger than most data sets used for pre-training, but also comprises reasonably clean and natural English text.”

GPT-3 [Brown et al. 2020]

- Filter from Common Crawls
 - Trained a **logistic regression classifier** to distinguish high-quality data from low-quality Common Crawl

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

What Other Public Training Datasets Exist Today?

Dataset	Example LMs	Tokens	Sources	Dataset	Example LMs	Tokens	Sources
C4 (Oct 2019)	T5, FLAN-T5	175B	Common Crawl	OpenWebMath (Oct 2023)	Llama	15B	Common Crawl
Pile (Dec 2020)	GPT-J, GPT-NeoX, Pythia	387B	Common Crawl, arXiv, PubMed, Books3, Gutenberg, Wikipedia, etc...	RedPajama v2 (Oct 2023)	-	30T	Common Crawl
The Stack v1 (Nov 2022)	StarCoder	200B	Software Heritage	Amber (Dec 2023)	Amber	1.3T	C4, RefinedWeb, the Stack, RedPajama v1
RedPajama v1 (Apr 2023)	INCITE	1.2T	Common Crawl, C4, Github, arXiv, Gutenberg, Books3, Wikipedia, Internet Archive (Stack Exchange)	Dolma 1.7 (Apr 2024)	OLMo 0424	2.3T	Dolma, RefinedWeb, RP's StackExchange, Flan, OpenWebMath, ...
RefinedWeb (Jun 2023)	Falcon	580B*	Common Crawl	FineWeb (May 2024)	-	15T	Common Crawl
Dolma (Aug 2023)	OLMo	3.1T	Common Crawl, C4, Semantic Scholar, Pushshift Reddit, Gutenberg, the Stack, Wikipedia, Wikibooks	Matrix (May 2024)	MAP-Neo	4.7T	RedPajama v2, Dolma, CulturaX, Amber, SlimPajama, Falcon, crawled Chinese web
				DCLM (Jun 2024)	DCLM-Baseline	4T	Common Crawl

What Other Public Training Datasets Exist Today?

Dataset	Example LMs	Tokens	Sources	Dataset	Example LMs	Tokens	Sources
C4 (Oct 2019)	T5, FLAN-T5	175B	Common Crawl	OpenWebMath (Oct 2023)	Llama	15B	Common Crawl
Pile (Dec 2020)	GPT-J, GPT-NeoX, Pythia	387B	Common Crawl, arXiv, PubMed, Books3, Gutenberg, Wikipedia, etc...	RedPajama v2 (Oct 2023)	-	30T	Common Crawl
The Stack v1 (Nov 2022)	StarCoder	200B	Software Heritage	Amber (Dec 2023)	Amber	1.3T	C4, RefinedWeb, the Stack, RedPajama v1
RedPajama v1 (Apr 2023)	INCITE	1.2T	Common Crawl, C4, Github, arXiv, Gutenberg, Books3, Wikipedia, Internet Archive (Stack Exchange)	Dolma 1.7 (Apr 2024)	OLMo 0424	2.3T	Dolma, RefinedWeb, RP's StackExchange, Flan, OpenWebMath, ...
RefinedWeb (Jun 2023)	Falcon	580B*	Common Crawl	FineWeb (May 2024)	-	15T	Common Crawl
Dolma (Aug 2023)	OLMo	3.1T	Common Crawl, C4, Semantic Scholar, Pushshift Reddit, Gutenberg, the Stack, Wikipedia, Wikibooks	Matrix (May 2024)	MAP-Neo	4.7T	RedPajama v2, Dolma, CulturaX, Amber, SlimPajama, Falcon, crawled Chinese web
				DCLM (Jun 2024)	DCLM-Baseline	4T	Common Crawl

“First open-source CC-based pre-training dataset”

What Other Public Training Datasets Exist Today?

Dataset	Example LMs	Tokens	Sources	Dataset	Example LMs	Tokens	Sources
C4 (Oct 2019)	T5, FLAN-T5	175B	Common Crawl	OpenWebMath (Oct 2023)	Llama	15B	Common Crawl
Pile (Dec 2020)	GPT-J, GPT-NeoX, Pythia	387B	Common Crawl, arXiv, PubMed, Books3, Gutenberg, Wikipedia, etc...	RedPajama v2 (Oct 2023)	-	30T	Common Crawl
The Stack v1 (Nov 2022)	StarCoder	200B	Software Heritage	Amber (Dec 2023)	Amber	1.3T	C4, RefinedWeb, the Stack, RedPajama v1
RedPajama v1 (Apr 2023)	INCITE	1.2T	Common Crawl, C4, Github, arXiv, Gutenberg, Books3, Wikipedia, Internet Archive (Stack Exchange)	Dolma 1.7 (Apr 2024)	OLMo 0424	2.3T	Dolma, RefinedWeb, RP's StackExchange, Flan, OpenWebMath, ...
RefinedWeb (Jun 2023)	Falcon	580B*	Common Crawl	FineWeb (May 2024)	-	15T	Common Crawl
Dolma (Aug 2023)	OLMo	3.1T	Common Crawl, C4, Semantic Scholar, Pushshift Reddit, Gutenberg, the Stack, Wikipedia, Wikibooks	Matrix (May 2024)	MAP-Neo	4.7T	RedPajama v2, Dolma, CulturaX, Amber, SlimPajama, Falcon, crawled Chinese web
				DCLM (Jun 2024)	DCLM-Baseline	4T	Common Crawl

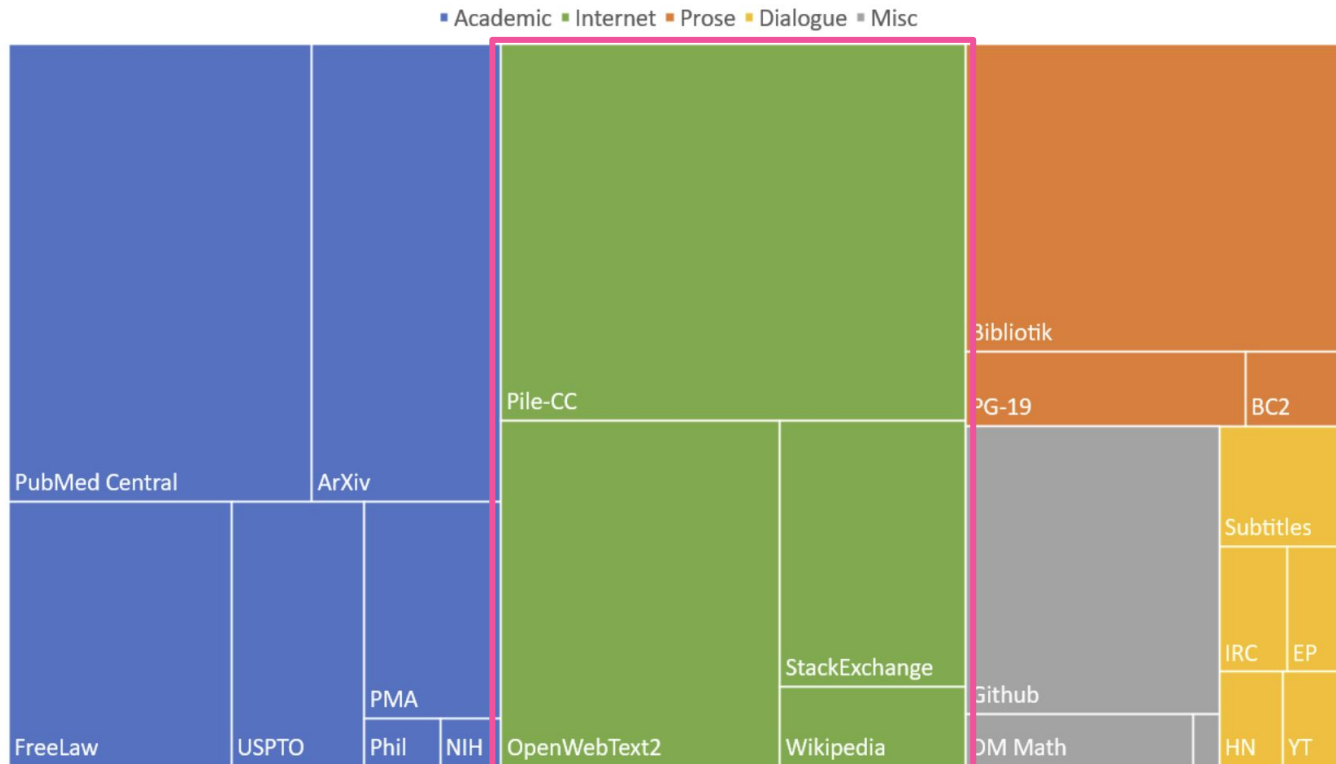
“First open-source pre-training dataset matching GPT-3’s training data size”

What Other Public Training Datasets Exist Today?

Dataset	Example LMs	Tokens	Sources	Dataset	Example LMs	Tokens	Sources
C4 (Oct 2019)	T5, FLAN-T5	175B	Common Crawl	OpenWebMath (Oct 2023)	Llama	15B	Common Crawl
Pile (Dec 2020)	GPT-J, GPT-NeoX, Pythia	387B	Common Crawl, arXiv, PubMed, Books3, Gutenberg, Wikipedia, etc...	RedPajama v2 (Oct 2023)	-	30T	Common Crawl
The Stack v1 (Nov 2022)	StarCoder	200B	Software Heritage	Amber (Dec 2023)	Amber	1.3T	C4, RefinedWeb, the Stack, RedPajama v1
RedPajama v1 (Apr 2023)	INCITE	1.2T	Common Crawl, C4, Github, arXiv, Gutenberg, Books3, Wikipedia, Internet Archive (Stack Exchange)	Dolma 1.7 (Apr 2024)	OLMo 0424	2.3T	Dolma, RefinedWeb, RP's StackExchange, Flan, OpenWebMath, ...
RefinedWeb (Jun 2023)	Falcon	580B*	Common Crawl	FineWeb (May 2024)	-	15T	Common Crawl
Dolma (Aug 2023)	OLMo	3.1T	Common Crawl, C4, Semantic Scholar, Pushshift Reddit, Gutenberg, the Stack, Wikipedia, Wikibooks	Matrix (May 2024)	MAP-Neo	4.7T	RedPajama v2, Dolma, CulturaX, Amber, SlimPajama, Falcon, crawled Chinese web
				DCLM (Jun 2024)	DCLM-Baseline	4T	Common Crawl

“Open-source pre-training datasets w/ >1T tokens”

Example 1: The Pile [Gao et al. 2020]



Example 2: Llama 1 [Touvron et al. 2023]

“For the most part, we reuse data sources that have been leveraged to train other LLMs, with the restriction of only using data that is publicly available, and compatible with open sourcing.”

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

Table 1: **Pre-training data.** Data mixtures used for pre-training, for each subset we list the sampling proportion, number of epochs performed on the subset when training on 1.4T tokens, and disk size. The pre-training runs on 1T tokens have the same sampling proportion.

Example 3: Dolma [Soldaini et al. 2024]

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

Today's Datasets/Models

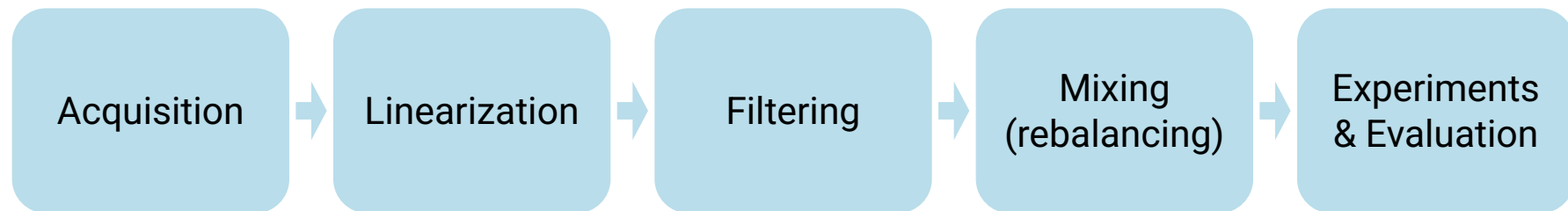
- Open source
 - C4 – leading to T5 (fully open, many sizes)
 - The Pile – leading to Pythia (fully open, many sizes/checkpoints)
 - RefinedWeb – leading to Falcon LLM (partially open, many sizes/checkpoints)
 - Dolma – leading to OLMo 1/2 (fully open, many sizes/checkpoints)
- Open Weights with data info
 - Llama 1
- Open Weights with little data info
 - Llama 2/3/4
 - Qwen series
 - DeepSeek series
 - GLM, Kimi, etc...

“Our training corpus includes a new mix of data from publicly available sources, which does not include data from Meta’s products or services. We made an effort to remove data from certain sites known to contain a high volume of personal information about private individuals. We trained on 2 trillion tokens of data as this provides a good performance–cost trade-off, up-sampling the most factual sources in an effort to increase knowledge and dampen hallucinations.”

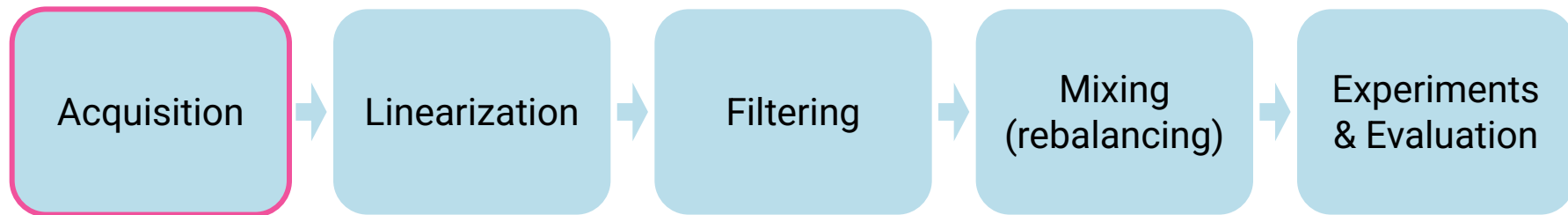
– Llama 2 paper

Overview of pre-training data curation

Overview



Overview



Common Crawl

+

additional sources obtained by developers

1. Acquisition – Why not just Common Crawl?

- Big crawlers have coverage issues
- For high-quality data, it's better to duplicate (e.g., Wikipedia)
- Some data may not be available on Common Crawl (e.g., non-HTML, PDFs)

December 11, 2024

Expanding the Language and Cultural Coverage of Common Crawl

We aim to enhance linguistic diversity in our dataset by inviting community contributions of non-English URLs and collaborating with MLCommons on a Language Identification campaign.



Pedro Ortiz Suarez

Pedro is a French-Colombian mathematician, computer scientist, and researcher. He holds a PhD in computer science and Natural Language Processing from Sorbonne Université.

OPENWEBMATH: AN OPEN DATASET OF HIGH-QUALITY MATHEMATICAL WEB TEXT

✦Keiran Paster,[‡] ✦Marco Dos Santos,[‡] ✦Zhangir Azerbayev, ✦Jimmy Ba

✦University of Toronto; Vector Institute for Artificial Intelligence

[‡]University of Cambridge, [‡]Princeton University
keirp@cs.toronto.edu, mjad3@cam.ac.uk

ABSTRACT

There is growing evidence that pretraining on high quality, carefully thought-out tokens such as code or mathematics plays an important role in improving the reasoning abilities of large language models. For example, Minerva, a PaLM model finetuned on billions of tokens of mathematical documents from arXiv and the web, reported dramatically improved performance on problems that require quantitative reasoning. However, because all known publicly released web datasets employ preprocessing that does not faithfully preserve mathematical notation, the benefits of large scale training on quantitative web documents are unavailable to the research community. We introduce OpenWebMath, an open dataset inspired by these works containing 14.7B tokens of mathematical webpages from Common Crawl. We describe in detail our method for extracting text and $\text{L}^{\text{a}}\text{T}_{\text{E}}\text{X}$ content and removing boilerplate from HTML documents, as well as our methods for quality filtering and deduplication. Additionally, we run small-scale experiments by training 1.4B parameter language models on OpenWebMath, showing that models trained on 14.7B tokens of our dataset surpass the performance of models trained on over 20x the amount of general language data. We hope that our dataset, openly released on the Hugging Face Hub, will help spur advances in the reasoning abilities of large language models.

..does not faithfully
preserve
mathematical
notation...

1. Acquisition – Example sources

	Why?	Examples
Encyclopedia	Open-source, free, multilingual, educational, written and edited by experts or community contributors, providing a certain level of authority and reliability, easily accessible	Wikipedia
Books	Very high quality, longer textual content, breadth (covers a wide range of subjects and topics), educational (biographies, textbooks)	Project Gutenberg (earliest digital library) Toronto Book Corpus, BookCorpus, Books1, Books2, Books3 – no longer accessible
Academic contents	Papers, journal articles, etc. Highly professional and has academic rigor. Educational. Professional information	arXiv (in LATEX format), S2ORC, PubMed
Code	Critical for coding tasks	The Stack, Github, BIG-QUERY, StarCoder
Math	Critical for math tasks	OpenMathText, FineMath
Forums	QA formatted	StackExchange, Reddit

1. Acquisition

Wikipedia

We use the Wikipedia dataset available on Huggingface, which is based on the Wikipedia dump from 2023-03-20 and contains text in 20 different languages. The dataset comes in preprocessed format, so that hyperlinks, comments and other formatting boilerplate has been removed.

Gutenberg and Books3

Defunct: The 'book' config is defunct and no longer accessible due to reported copyright infringement for the Book3 dataset contained in this config.

ArXiv

ArXiv data is downloaded from Amazon S3 in the `arxiv` requester pays bucket. We only keep latex source files and remove preambles, comments, macros and bibliographies.

huggingface.co/datasets/togethercomputer/RedPajama-Data-1

More discussion in the later class

Meta Secretly Trained Its AI on a Notorious Piracy Database, Newly Unredacted Court Docs Reveal

One of the most important AI copyright legal battles just took a major turn.

In his order, Chhabria referenced an internal quote from a Meta employee, included in the documents, in which they speculated, “If there is media coverage suggesting we have used a dataset we know to be pirated, such as LibGen, this may undermine our negotiating position with regulators on these issues.” Meta declined to comment.

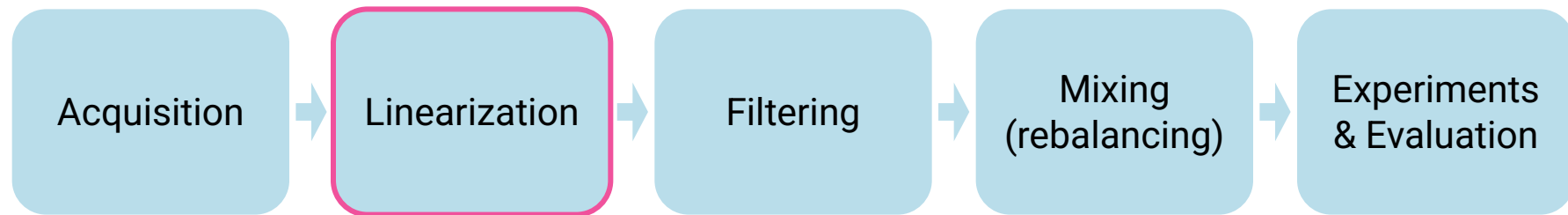
wired.com/story/new-documents-unredacted-meta-copyright-ai-lawsuit/

Stack Overflow Will Charge AI Giants for Training Data

The programmer Q&A site joins Reddit in demanding compensation when its data is used to train algorithms and ChatGPT-style bots

wired.com/story/stack-overflow-will-charge-ai-giants-for-training-data/

Overview



2. Linearization

- HTML -> Plain Text
- PDF -> Plain Text
- ...

The screenshot shows the Clizbeats website interface. At the top, there's a navigation bar with links like 'Press', 'Music', 'New Music Now', 'Throwback Thursday', '#ICYMI', 'Contact Us', and 'About Us'. Below this is a 'Breaking News' banner with a date of '30 Apr, 2019' and a headline about Disney Music Group's 'Avengers: Endgame' playlist. The main content area features the article 'Warner Music Group Strengthens Global Technology & Data Expertise' by Clizbeats.Com, dated '15 Dec, 2016'. The article includes a large blue circular logo with a white 'W' and text describing the new senior management structure at Warner Music Group (WMG), mentioning executives Ralph Munsen and Winnie Freda. A sidebar on the right contains sections for 'HIGHLIGHTS', 'POPULAR', 'COMMENTS', 'Check Out Our Newest Hot List Pick', 'CATEGORIES' (listing #ICYMI, Articles, Artist of the Month, Featured, Indie Artist Spotlight, Interviews, Music, New Music Now, Throwback Thursday, Video), 'SEARCH CLIZBEATS', and 'ARCHIVE'.

"text": "New Music Now\nThrowback Thursday\n#ICYMI\n30 Apr, 2019\nLIFETIME TO DEBUT FOLLOW UP SPECIAL SURVIVING R. KELLY: THE IMPACT WITH HOST SOLEDAD O'BRIEN ON MAY 4 Following the debut of the record-breaking premiere of Lifetime's Surviving R. Kelly this past January, the network will debut... Read More\n30 Apr, 2019\nListen To Disney Music Group's \"Avengers: Endgame\" Playlist On Streaming Services As Marvel Studios' Avengers Endgame continues to break box office records around the world, we are happy... Read More\n24 Apr, 2019\nSONY MUSIC ENTERTAINMENT NAMES SYLVIA RHONE CHAIRMAN AND CEO OF EPIC RECORDS Sony Music Entertainment recently announced the promotion of Sylvia Rhone to Chairman and CEO of Epic Records. In... Read More\n24 Apr, 2019\nANDERSON .PAAK ANNOUNCES FIRST EVER HEADLINING SHOW AT THE FORUM IN LOS ANGELES JUNE 29 TH +UPCOMING TOUR On the heels of a #4 debut on the Billboard Top 200 for his latest album Ventura... Read More\nArtist of the Month\nIndie Artist Spotlight\nClizbeats.com\nArticles\nWarner Music Group Strengthens Global Technology & Data Expertise\n15 Dec, 2016\nArticles\nWarner Music Group Logo\nWarner Music Group (WMG) today announced a new senior management structure for the company's technology and data areas. Veteran media and music executives Ralph Munsen and Winnie Freda have been named Chief Information Officer (CIO) and Chief Data Officer (CDO), respectively, reporting to CEO Steve Cooper. Munsen and Freda bring a wealth of diverse and complementary knowledge and experience to their new posts and will work closely together to ensure WMG is providing its artists, songwriters and labels with cutting-edge, best-in-class technology architecture and data analytics capabilities. As WMG's CIO, Munsen will lead the company's global technology strategy, infrastructure, and operations. His responsibilities will include developing innovative solutions, systems, and platforms that serve the creative and commercial needs of WMG's divisions, artists, and business partners. The newly created post of Chief Data Officer recognizes data's role as a critical asset in informing all aspects of the modern music business. Freda will be responsible for developing and executing a comprehensive vision and strategy for data management throughout the company, championing the use of data to provide useful insight. The establishment of these two interdependent senior management posts reflects the importance of providing advanced tech solutions and

2. Linearization *matters*

[WET file]

A Guide To Markets - The New York Times
NYTimes.com no longer supports Internet Explorer 9 or earlier. Please upgrade your browser. [LEARN !](#)
Sections
Home
Search
Skip to content Skip to navigation View mobile version
The New York Times
Archives | A Guide To Markets
Search
Subscribe Now
Log In
0
Settings
Close search
Site Search Navigation
Search NYTimes.com
Clear this text input
Go
<https://nyti.ms/29nVV3Q>
Loading...
See next articles
See previous articles
Site Navigation
Site Mobile Navigation
Advertisement
Archives | 1990
A Guide To Markets
MAY 10, 1990
Continue reading the main story [Share This Page](#)
Continue reading the main story

← Choppy. Sentences split to many newlines. A lot of undesirable website content.

[Resiliparse]

This is a digitized version of an article from The Times's print archive, before the start of online publication in 1996. To preserve these articles as they originally appeared, The Times does not alter, edit or update them.

Occasionally the digitization process introduces transcription errors or other problems. Please send reports of such problems to archive_feedback@nytimes.com.

May 10, 1990, Page 00006 The New York Times Archives

HERE is a sampling of some of the better antiques and flea markets around the United States.

Two or Three Times a Year

BRIMFIELD Route 20, Brimfield, Mass. 01010; 413-245-3436. Second weekend of May and July, and the second weekend after Labor Day.

RENNINGER'S OUTDOOR EXTRAVAGANZA Noble Street, Kutztown, Pa.; 717-385-0104. Thursday, Friday and Saturday of the last weekend of April, June, September.

FARMINGTON ANTIQUES WEEKEND Farmington Polo Grounds, Town Farm Road, Farmington, Conn. 06032; 508-839-9735. Starting Wednesday before shows open; 203-677-7862. June 9-10 and Sept. 1-2.

Monthly

ANN ARBOR ANTIQUES MARKET, P.O. Box 1512, Ann Arbor, Mich. 48106; 313-662-9453. May through October, third Sunday.

Continue reading the main story

KANE COUNTY FLEA MARKET, Kane County Fairgrounds, P.O. Box 549, St. Charles, Ill. 60174; 708-377-2252. Year-round, first weekend.

THE METROLINA EXPO, 7100 Statesville Road, Charlotte, N.C. 28213; 704-596-4643. Year-round, first weekend of every month.

SPRINGFIELD ANTIQUE SHOW AND FLEA MARKET, Clark County Fairgrounds, Route 41, Springfield, Ohio, 45501; 513-325-0053. Year-round, third weekend.

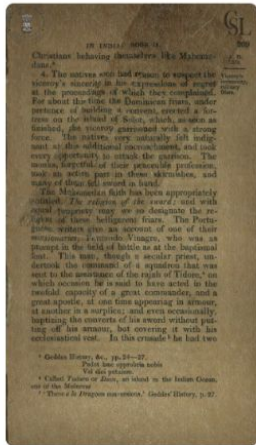
Weekly

BAKERSFIELD SWAP-O-RAMA, 4501 Wible Road, Bakersfield, Calif. 93313; 805-831-9342. Saturday and Sunday.

LAMBERTVILLE ANTIQUE MARKET, Route 29, Lambertville, N.J. 08530. Weekend number: 609-397-0456. Weekday: 215-752-4485, between 5 and 7 P.M. Market on Saturday and Sunday.

ATLANTA FLEA MARKET AND ANTIQUE CENTER, 5360 Peachtree Industrial Boulevard, Chamblee, Ga. 30341; 404-458-0456. Friday,

High variability in plain text output across different tools



IN INDIA * BOOK II

Christians behaving themselves like Mahomedans.

The natives soon had reason to suspect the viceroys's sincerity in his expressions of regret at the proceedings of which they complained. For about this time the Dominican friars, under pretence of building a convent, erected a fortress on the island of Solor, which, as soon as finished, the viceroys gararrisoned with a strong force. The natives, very naturally full indignant at this additional encroachment, and took every opportunity to attack the monks. The monks, forgetful of their peaceable profession, took an active part in these skirmishes, and many of them fell sword in hand.

The Mahomedan faith has been appropriately entitled, "The religion of the sword"; and with equal propriety may we so designate the religion of these belligerent friars. The Portuguese writers give an account of one of their missionaries, Fernando Vinagre, who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was sent to the assistance of the rajah of Tidore, on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, at another in a surplice, and even occasionally, baptizing the converts of his sword without putting off his armour, but covering it with his ecclesiastical vest. In this crusade he had two

3 a
of
t
uo

High quality

11 11 IN INDIA: BOOK

Christians behaving themselves like Mahomedans 3.4. The natives soon had reason to suspect the Viceroy's sincerity in his expressions of regret insincerely at the proceedings of which they complained.

For about this time the Dominican friars, under pretence of building a convent, erected a fortress on the island of Solor, which, as soon as finished, the viceroys gararrisoned with a strong force.

The natives very naturally felt indig S nant at this additional encroachment, and took every opportunity to attack the garrison. The monks, forgetful of their peaceable profession, took an active part in these skirmishes, and many of them fell sword in hand.

The Mchomedan faith has been appropriately rttitled. The religion of the sword; and with qual propriety may we so designate the religion of these belligerent friars. The Portuguese writers give an account of one of their missionaries, Fernando Vinagre, who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was sent to the assistance of the rajah of Tidore, on which occasion he is said to have acted in the twofold capacity of a great commander, and a great

3 a
of
t
uo

High quality

Pudet haec opprobria nobis Vel dici potuisse.

4 Called Tadura or Daco, an island in the Indian Ocean, one of the Moluccas 5. These a la Dragon conversions.' Geddes' History, p. 27.

Christians behaving themselves like Ma borne- a. >-. dans.3 "5/0-

4. The natives soon had reason to suspect the viceroys's sincerity in his expressions of regret

at the proceedings of which they complained. &"e"m For about this time the Dominican friars, under

pretence of building a convent, erected a for tress on the island of Sol or, which, as soon as finished, the viceroys gararrisoned with a strong

force. The natives' very naturally felt indig S nant at this additional encroachment, and took every opportunity to attack the garrison. The monks, forgetful of their peaceable profession, took an active part in these skirmishes, and many of tbg tr fell sword in hand.

The ffinomedan faith has been appropriately

en

Low quality

gls

missionaries, Fernando Vinagre, who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was

I sent to the assistance of the rajah of Tidore,4 on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, ; at another in a surplice; and even occasionally, baptizing the converts of his sword without put ting off his armour, but covering it with his ecclesiastical vest. In this crusade5 he had two

3 Geddesa History, &c., pp. 24—27.

Pudet haec opprobria nobis

Vel dici potuisse.

* Called T a d u r a or D a c o , an island in the Indian Ocean,

one of the Moluccas

5 'These a la D r a g o o n conversions.' Geddes' History, p. 27.

IN INDIA * BOOK TI. S69

IN INDIA * BOOK TI.

S

I

; dans.3

"5/0-

S69

Christians behaving themselves like Ma borne- a. >-. dans.3

"5/04. The natives soon had reason to suspect the viceroys's sincerity in his expressions of regret

at the proceedings of which they complained. &"e"m For about this time the Dominican friars, under

pretence of building a convent, erected a for

tress on the island of Sol or, which, as soon as

finished, the viceroys gararrisoned with a strong

force. The natives' very naturally felt indig S nant at this additional encroachment, and took

every opportunity to attack the garrison. The monks, forgetful of their peaceable profession, took an active part in these skirmishes, and

many of tbg tr fell sword in hand.

The ffinomedan faith has been appropriately

en

Low quality

gls

missionaries, Fernando Vinagre, who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was

I sent to the assistance of the rajah of Tidore,4 on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, ; at another in a surplice; and even occasionally, baptizing the converts of his sword without put ting off his armour, but covering it with his ecclesiastical vest. In this crusade5 he had two

3 Geddesa History, &c., pp. 24—27.

Pudet haec opprobria nobis

Vel dici potuisse.

* Called T a d u r a or D a c o , an island in the Indian Ocean,

one of the Moluccas

5 'These a la D r a g o o n conversions.' Geddes' History, p. 27.

IN INDIA * BOOK TI. S69

Christians behaving themselves like Ma borne- a. >-. dans.3

"5/0-

S69

4. The natives soon had reason to suspect the viceroys's sincerity in his expressions of regret

at the proceedings of which they complained. &"e"m For about this time the Dominican friars, under

pretence of building a convent, erected a for

tress on the island of Sol or, which, as soon as

finished, the viceroys gararrisoned with a strong

force. The natives' very naturally felt indig-

S

nant at this additional encroachment, and took

every opportunity to attack the garrison. The monks, forgetful of their peaceable profession, took an active part in these skirmishes, and

many of tbg tr fell sword in hand.

The ffinomedan faith has been appropriately

en

Low quality

gls

missionaries, Fernando Vinagre, who was as prompt in the field of battle as at the baptismal font. This man, though a secular priest, undertook the command of a squadron that was

I sent to the assistance of the rajah of Tidore,4 on which occasion he is said to have acted in the twofold capacity of a great commander, and a great apostle, at one time appearing in armour, ; at another in a surplice; and even occasionally, baptizing the converts of his sword without put ting off his armour, but covering it with his ecclesiastical vest. In this crusade5 he had two

3 Geddesa History, &c., pp. 24—27.

Pudet haec opprobria nobis

Vel dici potuisse.

* Called T a d u r a or D a c o , an island in the Indian Ocean,

one of the Moluccas

5 'These a la D r a g o o n conversions.' Geddes' History, p. 27.

IN INDIA * BOOK TI. S69

Overview



3. Filtering

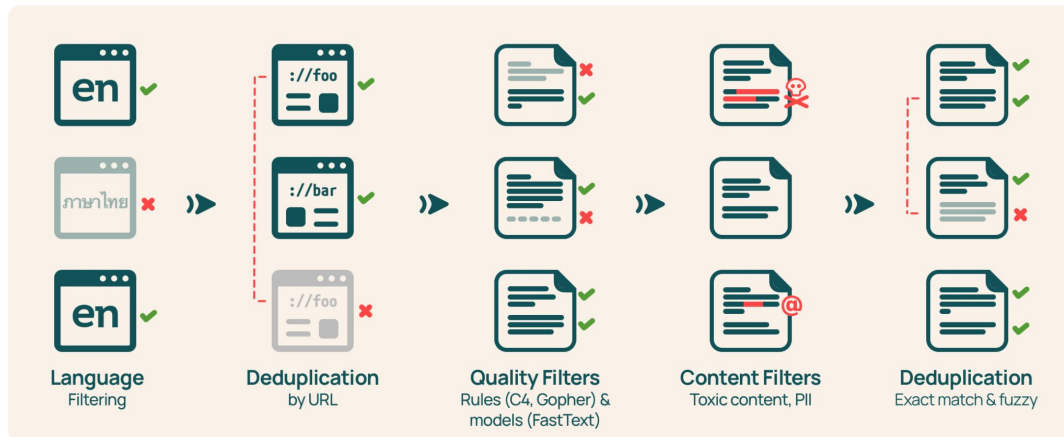
Considerations

- Language
- Junk
- Content
- Quality
- Duplicates
- etc...

Methods

- Heuristic based
- Model based

More in DCLM & Fineweb papers



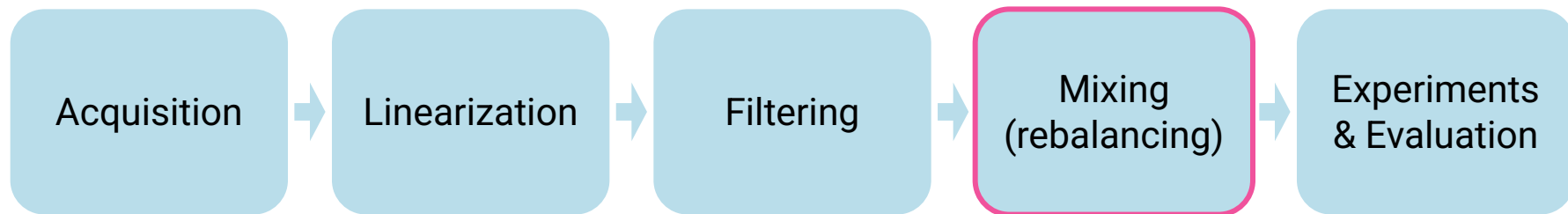
➤ *Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research* (Soldaini et al; ACL 2024)

➤ *The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale* (Penedo et al 2024)

➤ *DataComp-LM: In search of the next generation of training sets for language models* (Li et al; Neurips 2024)

.. and more

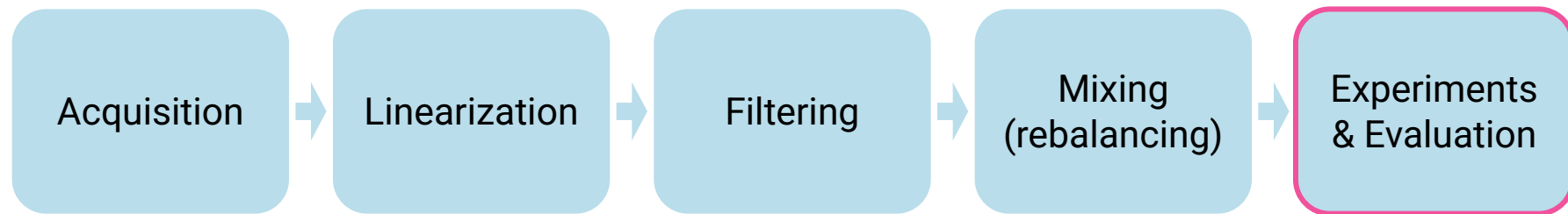
Overview



Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

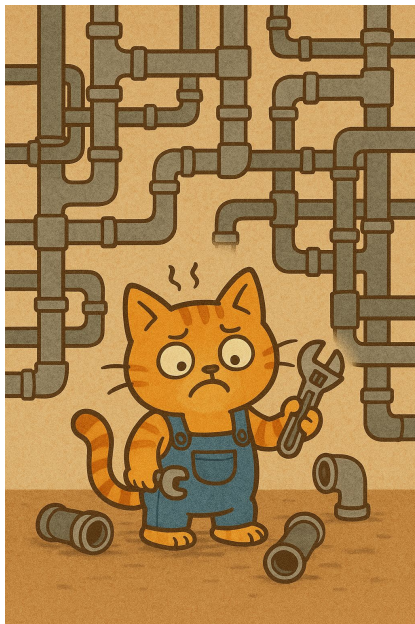
[from GPT-3 paper]

Overview

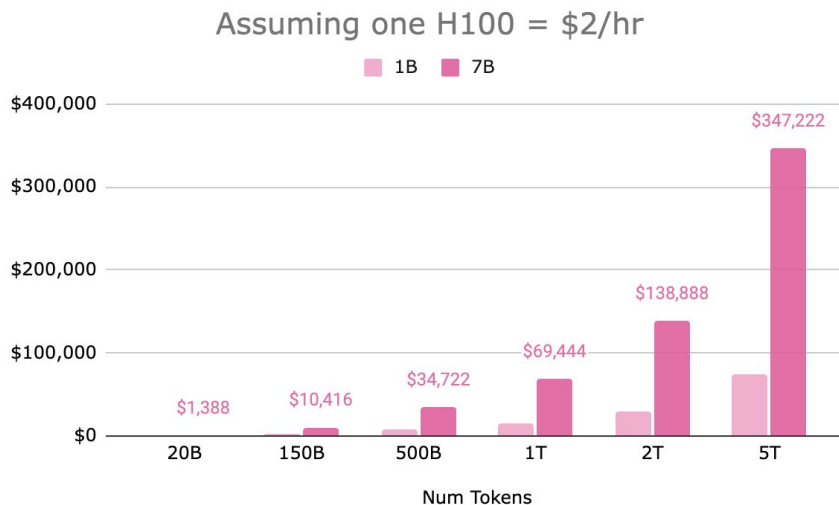


5. Experimentation – Why challenging?

Pipelines steps aren't independent



Pre-training runs are expensive



5. Experimentation – Rules of thumbs

- Fix scale
 - Quantity and quality have inherent tradeoffs → complicated!
 - Solve backwards from your compute budget → maximize quality
- Fix evaluation sets
 - Choose a subset that is a good representation of all NLP tasks
 - Choose benchmarks suitable for pre-training (different from fully trained models)
 - “How to get early, stable signals” rather than “most challenging”
- Invest in smallest experimental design that generalizes
 - “Scaling laws”
 - “Annealing”-style ablations, appear in many papers from Meta (Llama), Databricks, Ai2 (OLMo)

Paper Discussion

Overview

Today, we'll look at two papers:

- “DataComp-LM: In search of the next generation of training sets for language models”
- “FineWeb: decanting the web for the finest text data at scale”



Both used Common Crawls as (almost) only source and deep-dived into linearization & filtering

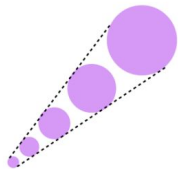
DCLM & FineWeb Motivation

- Supporting open research on training data
 - “Details about training sets are becoming increasingly rare, even for open-weight models such as the Llama, Mistral, or Gemma models” – DCLM paper
 - “[T]he pretraining datasets for state-of-the-art open LLMs like Llama 3 and Mixtral are not publicly available and very little is known about how they were created.” – FineWeb paper

DCLM & FineWeb Motivation

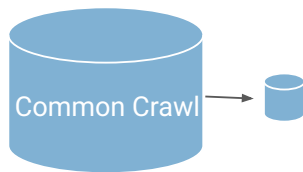
- Supporting open research on training data
 - “Details about training sets are becoming increasingly rare, even for open-weight models such as the Llama, Mistral, or Gemma models” – DCLM paper
 - “[T]he pretraining datasets for state-of-the-art open LLMs like Llama 3 and Mixtral are not publicly available and very little is known about how they were created.” – FineWeb paper
- Goal: Fix everything other than the training data (e.g. fix compute for scale, architecture, etc.)

A. Select a scale

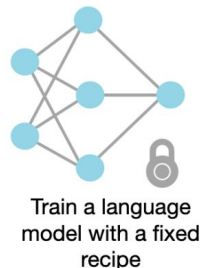


Pick a scale: 400M-1x,
1B-1x, 3B-1x, 7B-1x,
or 7B-2x

B. Build a dataset



C. Train a model



Train a language
model with a fixed
recipe

D. Evaluate



53 downstream
zero-shot and
few-shot tasks

DCLM & FineWeb: Settings

[DCLM]

Scale	Model parameters	Train tokens	Train FLOPs	Train H100 hours	Pool size
400M-1x	412M	8.2B	2.0e19	26	469B
1B-1x	1.4B	28.8B	2.4e20	240	1.64T
3B-1x	2.8B	55.9B	9.4e20	740	3.18T
7B-1x	6.9B	138B	5.7e21	3,700	7.85T
7B-2x	6.9B	276B	1.1e22	7,300	15.7T

Model size: # of model parameters

Data size: 1x or 2x of Chinchilla – 1x or 2x of the size that is optimal according to the Chinchilla scaling laws (next Tuesday)

[FineWeb]

Always trained a 1.8B model

- For most ablations, ~28B tokens (1x Chinchilla)
- After each step of filtering, ~350B tokens (12x Chinchilla)

DCLM & FineWeb: Evaluation

- Criteria (in both papers) – meaningful signals at small scale (e.g., can't really do complex tasks like reasoning tasks)
 - Small variance
 - Trends predictable
 - Above random baseline even at small scale
- DCLM:
 - MMLU 5-shot accuracy
 - CORE: Avg of 22 tasks (e.g., HellaSwag and ARC) that provide a low-variance signal
 - EXTENDED: Avg of 53 tasks
- FineWeb:
 - Avg. of 8 datasets (CSQA, HellaSwag, OBQA, PIQA, SIQA, WinoGrande, ARC, MMLU)

Data Curation: Overview

1. Acquisition: Unfiltered web-text corpus from all Common Crawl data prior to 2023
2. Linearization: Extract text from HTML → Called DCLM-Pool (240T tokens)
3. Heuristic-based filtering → 19.9% of DCLM-Pool retained
4. Deduplication → 13.7% of DCLM-Pool retained
5. Model-based filtering → 1.4% of DCLM-Pool retained → called DCLM-Baseline (3.8T tokens)

-
1. Acquisition: Unfiltered web-text corpus from all Common Crawl data (96 dumps)
 2. Linearization: Extract text from HTML
 3. Heuristic filtering I → 36T tokens
 4. Deduplication → 20T tokens
 5. Heuristic filtering II → 15T tokens → Called FineWeb
 6. Model-based filtering → 1.3T tokens → Called FineWeb-Edu

Data Curation: Overview

1. Acquisition: Unfiltered web-text corpus from all Common Crawl data prior to 2023
2. Linearization: Extract text from HTML → Called DCLM-Pool (240T tokens)
3. Heuristic-based filtering → 19.9% of DCLM-Pool retained
4. Deduplication → 13.7% of DCLM-Pool retained
5. Model-based filtering → 1.4% of DCLM-Pool retained → called DCLM-Baseline (3.8T tokens)

-
1. Acquisition: Unfiltered web-text corpus from all Common Crawl data (96 dumps)
 2. Linearization: Extract text from HTML
 3. Heuristic filtering I → 36T tokens
 4. Deduplication → 20T tokens
 5. Heuristic filtering II → 15T tokens → Called FineWeb
 6. Model-based filtering → 1.3T tokens → Called FineWeb-Edu

Linearization (1/2)

Both papers are some of the first to ablate linearization!

CommonCrawl data is available in two main formats:

- WET (WARC Encapsulated Text) files provide a plain text only version.
 - Most prior work takes the WET files as their starting point.
- WARC (Web ARChive format) files contain the raw data from the crawl (HTML and metadata)

Both papers found that own linearization from WARC better than using Common Crawl provided WET files

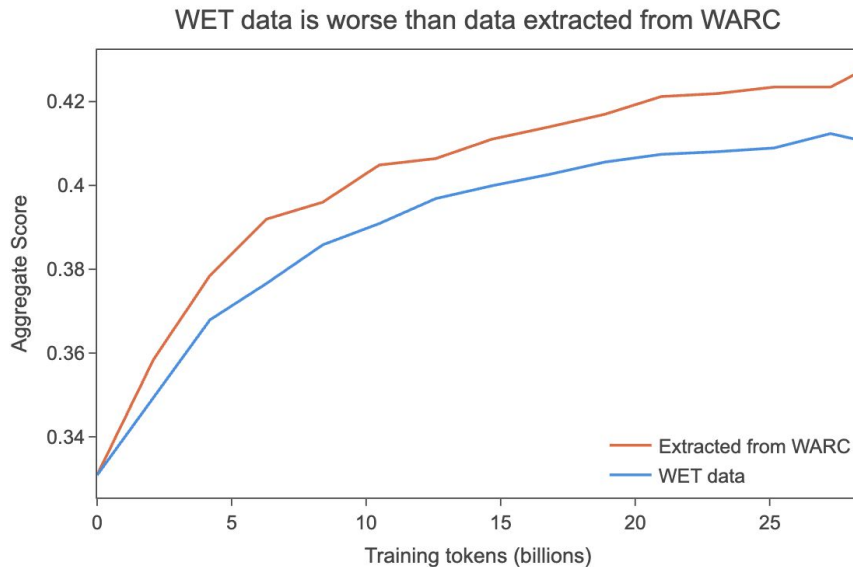
Linearization (2/2)

DCLM: Resiliparse chosen because of efficiency

Table 3: **Comparison of text extractors** (1B-1x scale).

Text Extraction	CORE	EXTENDED
resiliparse	24.1	13.4
trafilatura	24.5	12.5
WET files	20.7	12.2

FineWeb: trafilatura chosen



[WET file]

A Guide To Markets - The New York Times
NYTimes.com no longer supports Internet Explorer 9 or earlier. Please upgrade your browser. [LEARN MORE](#) »
Sections
Home
Search
Skip to content Skip to navigation View mobile version
The New York Times
Archives|A Guide To Markets
Search
Subscribe Now
Log In
0
Settings
Close search
Site Search Navigation
Search NYTimes.com
Clear this text input
Go
<https://nyti.ms/29nV3Q>
Loading...
See next articles
See previous articles
Site Navigation
Site Mobile Navigation
Advertisement
Archives | 1990
A Guide To Markets
MAY 10, 1990
Continue reading the main story [Share This Page](#)
Continue reading the main story
About the Archive
This is a digitized version of an article from The Times's print archive, before the start of online publication in 1996. To preserve these articles as they originally appeared, The Times does not alter, edit or update them.
Occasionally the digitization process introduces transcription errors or other problems. Please send reports of such problems to archive_feedback@nytimes.com.
May 10, 1990, Page 00006 The New York Times Archives
HERE is a sampling of some of the better antiques and flea markets around the United States.
Two or Three Times a Year
BRIMFIELD Route 20, Brimfield, Mass. 01010; 413-245-3436. Second weekend of May and July, and the second weekend after Labor Day.
Continue reading the main story
What's Next
Loading...
Go to Home Page »
Site Index The New York Times
Site Index Navigation
News
World
U.S.
Politics
N.Y.
Business
Tech
Science
Health
Sports
Education
Obituaries
Today's Paper
Corrections
Opinion
Today's Opinion
Op-Ed Columnists
Editorials
Op-Ed Contributors
Letters
Sunday Review
Video: Opinion

[Trafilatura]

HERE is a sampling of some of the better antiques and flea markets around the United States.
Two or Three Times a Year
BRIMFIELD Route 20, Brimfield, Mass. 01010; 413-245-3436. Second weekend of May and July, and the second weekend after Labor Day.
RENNINGER'S OUTDOOR EXTRAVAGANZA Noble Street, Kutztown, Pa.; 717-385-0104. Thursday, Friday and Saturday of the last weekend of April, June, September.
FARMINGTON ANTIQUES WEEKEND Farmington Polo Grounds, Town Farm Road, Farmington, Conn. 06032; 508-839-9735. Starting Wednesday before shows open; 203-677-7862. June 9-10 and Sept. 1-2.
Monthly
ANN ARBOR ANTIQUES MARKET, P.O. Box 1512, Ann Arbor, Mich. 48106; 313-662-9453. May through October, third Sunday.
Continue reading the main storyKANE COUNTY FLEA MARKET, Kane County Fairgrounds, P.O. Box 549, St. Charles, Ill. 60174; 708-377-2252. Year-round, first weekend.
THE METROLINA EXPO, 7100 Statesville Road, Charlotte, N.C. 28213; 704-596-4643. Year-round, first weekend of every month.
SPRINGFIELD ANTIQUE SHOW AND FLEA MARKET, Clark County Fairgrounds, Route 41, Springfield, Ohio, 45501; 513-325-0053. Year-round, third weekend.
Weekly
BAKERSFIELD SWAP-O-RAMA, 4501 Wible Road, Bakersfield, Calif. 93313; 805-831-9342. Saturday and Sunday.
LAMBERTVILLE ANTIQUE MARKET, Route 29, Lambertville, N.J. 08530. Weekend number: 609-397-0456. Weekday: 215-752-4485, between 5 and 7 P.M. Market on Saturday and Sunday.
ATLANTA FLEA MARKET AND ANTIQUE CENTER, 5360 Peachtree Industrial Boulevard, Chamblee, Ga. 30341; 404-458-0456. Friday, Saturday and Sunday.
Continue reading the main story

[Resiliparse]

This is a digitized version of an article from The Times's print archive, before the start of online publication in 1996. To preserve these articles as they originally appeared, The Times does not alter, edit or update them.
Occasionally the digitization process introduces transcription errors or other problems. Please send reports of such problems to archive_feedback@nytimes.com.

May 10, 1990, Page 00006 The New York Times Archives

HERE is a sampling of some of the better antiques and flea markets around the United States.

Two or Three Times a Year

BRIMFIELD Route 20, Brimfield, Mass. 01010; 413-245-3436. Second weekend of May and July, and the second weekend after Labor Day.

RENNINGER'S OUTDOOR EXTRAVAGANZA Noble Street, Kutztown, Pa.; 717-385-0104. Thursday, Friday and Saturday of the last weekend of April, June, September.

FARMINGTON ANTIQUES WEEKEND Farmington Polo Grounds, Town Farm Road, Farmington, Conn. 06032; 508-839-9735. Starting Wednesday before shows open; 203-677-7862. June 9-10 and Sept. 1-2.

Monthly

ANN ARBOR ANTIQUES MARKET, P.O. Box 1512, Ann Arbor, Mich. 48106; 313-662-9453. May through October, third Sunday.

Continue reading the main story

KANE COUNTY FLEA MARKET, Kane County Fairgrounds, P.O. Box 549, St. Charles, Ill. 60174; 708-377-2252. Year-round, first weekend.

THE METROLINA EXPO, 7100 Statesville Road, Charlotte, N.C. 28213; 704-596-4643. Year-round, first weekend of every month.

Data Curation: Overview

1. Acquisition: Unfiltered web-text corpus from all Common Crawl data prior to 2023
2. Linearization: Extract text from HTML → Called DCLM-Pool (240T tokens)
3. Heuristic-based filtering → 19.9% of DCLM-Pool retained
4. Deduplication → 13.7% of DCLM-Pool retained
5. Model-based filtering → 1.4% of DCLM-Pool retained → called DCLM-Baseline (3.8T tokens)

-
1. Acquisition: Unfiltered web-text corpus from all Common Crawl data (96 dumps)
 2. Linearization: Extract text from HTML
 3. Heuristic filtering I → 36T tokens
 4. Deduplication → 20T tokens
 5. Heuristic filtering II → 15T tokens → Called FineWeb
 6. Model-based filtering → 1.3T tokens → Called FineWeb-Edu

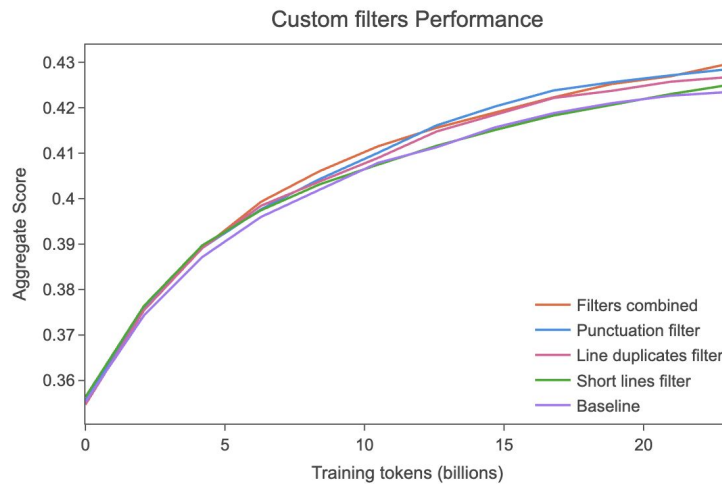
Heuristic Filtering: DCLM vs. FineWeb

- Many different prior datasets with mostly overlapping but slightly different filters
- DCLM: Compared heuristic filtering from prior work, and ended up with RefinedWeb
- FineWeb: First apply RefinedWeb's filter, then ablate different individual filters (length filter, punctuation, filter, etc) in C4

Table 2: **Comparison to existing datasets** (7B–1x scale).

Dataset	CORE	EXTENDED
C4	34.2	18.0
Dolma-V1	35.0	18.4
RedPajama	35.3	18.2
RefinedWeb	36.9	19.8

[DCLM]



[FineWeb]

Data Curation: Overview

1. Acquisition: Unfiltered web-text corpus from all Common Crawl data prior to 2023
2. Linearization: Extract text from HTML → Called DCLM-Pool (240T tokens)
3. Heuristic-based filtering → 19.9% of DCLM-Pool retained
4. Deduplication → 13.7% of DCLM-Pool retained
5. Model-based filtering → 1.4% of DCLM-Pool retained → called DCLM-Baseline (3.8T tokens)

-
1. Acquisition: Unfiltered web-text corpus from all Common Crawl data (96 dumps)
 2. Linearization: Extract text from HTML
 3. Heuristic filtering I → 36T tokens
 4. Deduplication → 20T tokens
 5. Heuristic filtering II → 15T tokens → Called FineWeb
 6. Model-based filtering → 1.3T tokens → Called FineWeb-Edu

Deduplication

- Common Crawls often contain many duplicate or near-duplicate data strings. Removing these duplicates serves the dual purpose of improving performance by reducing memorization and increasing data diversity.
- Often do fuzzy deduplication rather than exact deduplication
 - e.g., remove documents with 50% overlap in 13-grams

Deduplication

- Common Crawls often contain many duplicate or near-duplicate data strings. Removing these duplicates serves the dual purpose of improving performance by reducing memorization and increasing data diversity.
- Often do fuzzy deduplication rather than exact deduplication
 - e.g., remove documents with 50% overlap in 13-grams
- Really not trivial!!
 - DCLM has 8 pages of ablations
 - FineWeb has 7 pages of ablations
 - Their ablations mostly don't overlap!

Deduplication

- Common Crawls often contain many duplicate or near-duplicate data strings. Removing these duplicates serves the dual purpose of improving performance by reducing memorization and increasing data diversity.
- Often do fuzzy deduplication rather than exact deduplication
 - e.g., remove documents with 50% overlap in 13-grams
- Really not trivial!!
 - DCLM has 8 pages of ablations
 - FineWeb has 7 pages of ablations
 - Their ablations mostly don't overlap!
- Factors:
 - Hyperparams: $p\%$ overlap in n -gram
 - Data structure: MinHash vs. Suffix array vs. near-duplicate Bloom filtering (BFF)
 - Paragraph-level? Doc-level? A doc vs. doc level or doc vs. corpus level?
 - Sharded deduplication vs. global deduplication

Deduplication

- Common Crawls often contain many duplicate or near-duplicate data strings. Removing these duplicates serves the dual purpose of improving performance by reducing memorization and increasing data diversity.
- Often do fuzzy deduplication rather than exact deduplication
 - e.g., remove documents with 50% overlap in 13-grams
- Really not trivial!!
 - DCLM has 8 pages of ablations
 - FineWeb has 7 pages of ablations
 - Their ablations mostly do not include deduplication
- Factors:
 - Hyperparams: p% overlap in n-gram
 - Data structure: MinHash vs. Suffix array vs. near-duplicate Bloom filtering (BFF)
 - Paragraph-level? Doc-level? A doc vs. doc level or doc vs. corpus level?
 - Sharded deduplication vs. global deduplication

DCLM used “sharded deduplication” and ablated these

Conclusions:

- MinHash+SuffixArray+Exact $\approx$ BFF
- Former at competition scale; latter at larger scale

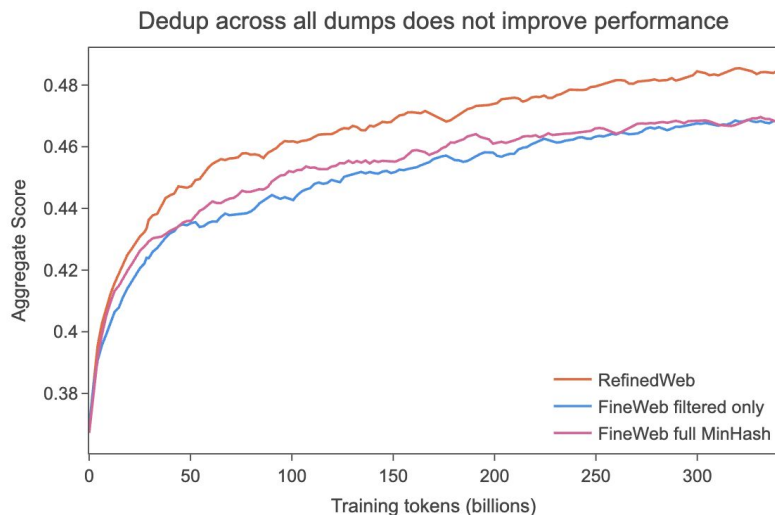
Deduplication

- Common Crawls often contain many duplicate or near-duplicate data strings. Removing these duplicates serves the dual purpose of improving performance by reducing memorization and increasing data diversity.
- Often do fuzzy deduplication rather than exact deduplication
 - e.g., remove documents with 50% overlap in 13-grams
- Really not trivial!!
 - DCLM has 8 pages of ablations
 - FineWeb has 7 pages of ablations
 - Their ablations mostly do
- Factors:
 - Hyperparams: $p\%$ overlap in n -gram
 - Data structure: MinHash vs. Suffix array vs. near-duplicate Bloom filtering (BFF)
 - Paragraph-level? Doc-level? A doc vs. doc level or doc vs. corpus level?
 - Sharded deduplication vs. global deduplication

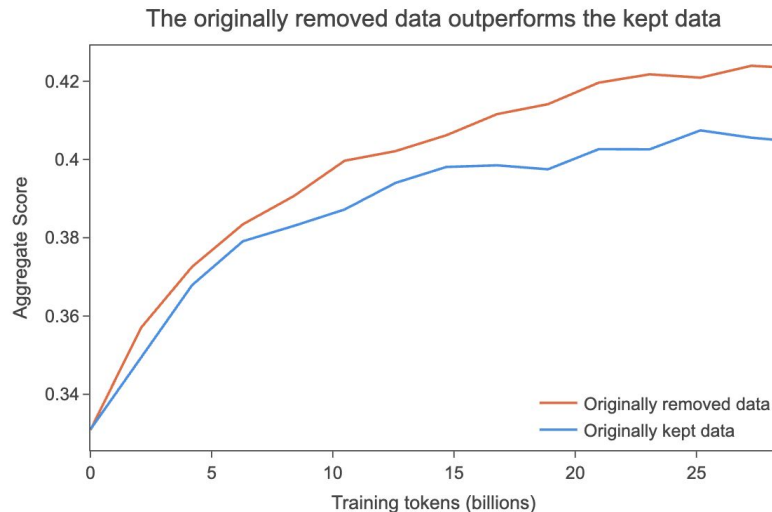
FineWeb used MinHash (removing 75+% similar docs in 5-grams) and ablated sharded vs. global [next slide]

Deduplication – FineWeb (1/2)

“Initially, we were operating under the assumption that more deduplication is always better, so [we took] the entire dataset (all 90+ dumps) and deduplicate them together”

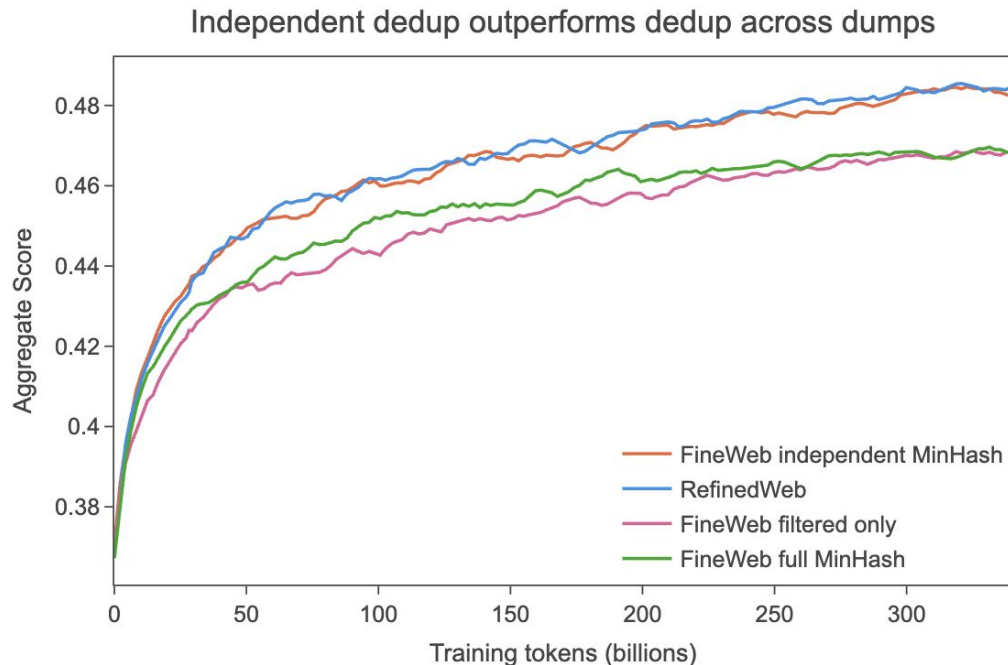


“This challenged our assumption that more deduplication would inevitably result in higher benchmark scores”



“Data that does not find a duplicate match in any other dump might actually be worse quality/more out of distribution”

Deduplication – FineWeb (2/2)



“We deduplicated each dump with MinHash individually (independently of the other dumps).”

Data Curation: Overview

1. Acquisition: Unfiltered web-text corpus from all Common Crawl data prior to 2023
2. Linearization: Extract text from HTML → Called DCLM-Pool (240T tokens)
3. Heuristic-based filtering → 19.9% of DCLM-Pool retained
4. Deduplication → 13.7% of DCLM-Pool retained
5. Model-based filtering → 1.4% of DCLM-Pool retained → called DCLM-Baseline (3.8T tokens)

Involves training a model, from an n-gram, logistic regression to a Transformer-based model

1. Acquisition
2. Linearization: Extract text from HTML
3. Heuristic filtering I → 36T tokens
4. Deduplication → 20T tokens
5. Heuristic filtering II → 15T tokens → Called FineWeb
6. Model-based filtering → 1.3T tokens → Called FineWeb-Edu

DCLM's Model-based filtering (1/2)

“Train a model” to serve as quality filters

- PageRank: scores based on how likely they're linked to other docs
- SemDedupe: dedupe based on semantic similarity
- Linear classifier: based on pretrained BGE embeddings
- AskLLM: prompt an LLM whether a doc is helpful
- Perplexity filtering: retain low PPL docs
- Top-k avg logits: avg. the top-k model logits over all words in a doc, and see how confident a model is
- fastText: a bigram classifier trained on positive and negative examples [details in next slide]

Table 4: **Quality filtering comparison** (1B-1x scale).

Filter	CORE	EXTENDED
RefinedWeb reproduction	27.5	14.6
Top 20% by Pagerank	26.1	12.9
SemDedup [1]	27.1	13.8
Classifier on BGE features [185]	27.2	14.0
AskLLM [146]	28.6	14.3
Perplexity filtering	29.0	15.0
Top-k average logits	29.2	14.7
fastText [87] OH-2.5 +ELI5	30.2	15.4

DCLM's Model-based filtering (1/2)

“Train a model” to serve as quality filters

- PageRank: scores based on how likely they're linked to other docs
- SemDedupe: dedupe based on semantic similarity
- Linear classifier: based on pretrained BGE embeddings
- AskLLM: prompt an LLM whether a doc is helpful
- Perplexity filtering: retain low PPL docs
- Top-k avg logits: avg. the top-k model logits over all words in a doc, and see how confident a model is
- fastText: a bigram classifier trained on positive and negative examples [details in next slide]

Table 4: **Quality filtering comparison** (1B-1x scale).

Filter	CORE	EXTENDED
RefinedWeb reproduction	27.5	14.6
Top 20% by Pagerank	26.1	12.9
SemDedup [1]	27.1	13.8
Classifier on BGE features [185]	27.2	14.0
AskLLM [146]	28.6	14.3
Perplexity filtering	29.0	15.0
Top-k average logits	29.2	14.7
fastText [87] OH-2.5 +ELI5	30.2	15.4

DCLM's Model-based filtering (1/2)

“Train a model” to serve as quality filters

- PageRank: scores based on how likely they're linked to other docs
- SemDedupe: dedupe based on semantic similarity
- Linear classifier: based on pretrained BGE embeddings
- AskLLM: prompt an LLM whether a doc is helpful
- Perplexity filtering: retain low PPL docs
- Top-k avg logits: avg. the top-k model logits over all words in a doc, and see how confident a model is
- fastText: a bigram classifier trained on positive and negative examples [details in next slide]

Table 4: **Quality filtering comparison** (1B-1x scale).

Filter	CORE	EXTENDED
RefinedWeb reproduction	27.5	14.6
Top 20% by Pagerank	26.1	12.9
SemDedup [1]	27.1	13.8
Classifier on BGE features [185]	27.2	14.0
AskLLM [146]	28.6	14.3
Perplexity filtering	29.0	15.0
Top-k average logits	29.2	14.7
fastText [87] OH-2.5 +ELI5	30.2	15.4

DCLM's Model-based filtering (1/2)

“Train a model” to serve as quality filters

- PageRank: scores based on how likely they're linked to other docs
- SemDedupe: dedupe based on semantic similarity
- Linear classifier: based on pretrained BGE embeddings
- AskLLM: prompt an LLM whether a doc is helpful
- Perplexity filtering: retain low PPL docs
- Top-k avg logits: avg. the top-k model logits over all words in a doc, and see how confident a model is
- fastText: a bigram classifier trained on positive and negative examples [details in next slide]

Table 4: **Quality filtering comparison** (1B-1x scale).

Filter	CORE	EXTENDED
RefinedWeb reproduction	27.5	14.6
Top 20% by Pagerank	26.1	12.9
SemDedup [1]	27.1	13.8
Classifier on BGE features [185]	27.2	14.0
AskLLM [146]	28.6	14.3
Perplexity filtering	29.0	15.0
Top-k average logits	29.2	14.7
fastText [87] OH-2.5 +ELI5	30.2	15.4

DCLM's Model-based filtering (1/2)

“Train a model” to serve as quality filters

- PageRank: scores based on how likely they're linked to other docs
- SemDedupe: dedupe based on semantic similarity
- Linear classifier: based on pretrained BGE embeddings
- AskLLM: prompt an LLM whether a doc is helpful
- Perplexity filtering: retain low PPL docs
- Top-k avg logits: avg. the top-k model logits over all words in a doc, and see how confident a model is
- fastText: a bigram classifier trained on positive and negative examples [details in next slide]

Table 4: **Quality filtering comparison** (1B-1x scale).

Filter	CORE	EXTENDED
RefinedWeb reproduction	27.5	14.6
Top 20% by Pagerank	26.1	12.9
SemDedup [1]	27.1	13.8
Classifier on BGE features [185]	27.2	14.0
AskLLM [146]	28.6	14.3
Perplexity filtering	29.0	15.0
Top-k average logits	29.2	14.7
fastText [87] OH-2.5 +ELI5	30.2	15.4

DCLM's Model-based filtering (2/2)

How to train a classifier?

- Positive examples
 - Wikipedia
 - OpenWebText2
 - GPT-3 Approx: Wikipedia, OpenWebText2, books
 - Instruction-formatted data from OpenHermes 2.5 (OH-2.5) and high-scoring posts from the ELI5 subreddit
- Negative examples: Random samples

Table 5: **fastText ablations** (7B-1x scale).

Dataset	Threshold	CORE	MMLU	EXTENDED
OH-2.5 + ELI5	10%	41.0	29.2	21.4
Wikipedia	10%	35.7	27.0	19.1
OpenWebText2	10%	34.7	25.0	18.7
GPT-3 Approx	10%	37.5	24.4	20.0
OH-2.5 + ELI5	15%	39.8	27.2	21.5
OH-2.5 + ELI5	20%	38.7	24.2	20.3

Using a fairly strict threshold (top-10%) helps over more permissive ones (top-15% and 20%)

FineWeb's Model-based filtering – Motivation

Uses synthetic data to develop classifiers for identifying “educational” content.

- Inspired by Llama 3 and Phi3
 - Phi3: “Our training data consists of heavily filtered publicly available web data (according to the 'educational level') from various open internet sources, as well as synthetic LLM-generated data.”
 - Llama 3: “We found that previous generations of Llama are good at identifying high-quality data, so we used Llama 2 to help build the text-quality classifiers that are powering Llama 3.”
- However, details as well as ablations has not been publicly available.

FineWeb's Model-based filtering – Method

- First use Llama-3-70B-Instruct to annotate the quality of the doc, based on its educational contents (on a scale of 0 to 5)
 - Found to be better than Mixtral-8x7B-Instruct and Mixtral-8x22B-Instruct

FineWeb's Model-based filtering – Method

- First use Llama-3-70B-Instruct to annotate the quality of the doc, based on its educational contents (on a scale of 0 to 5)
 - Found to be better than Mixtral-8x7B-Instruct and Mixtral-8x22B-Instruct
- Llama3-70B annotations were used to train a small classifier
 - Snowflake-arctic-embed embedding model (110M params)
 - Trained this model on the 450,000 Llama 3 annotations for 20 epochs
 - After training, rounded the scores to 0 to 5, and converted the problem to a binary classification task by using a fixed threshold to determine if a file is educational.

FineWeb's Model-based filtering – Method

- First use Llama-3-70B-Instruct to annotate the quality of the doc, based on its educational contents (on a scale of 0 to 5)
 - Found to be better than Mixtral-8x7B-Instruct and Mixtral-8x22B-Instruct
- Llama3-70B annotations were used to train a small classifier
 - Snowflake-arctic-embed embedding model (110M params)
 - Trained this model on the 450,000 Llama 3 annotations for 20 epochs
 - After training, rounded the scores to 0 to 5, and converted the problem to a binary classification task by using a fixed threshold to determine if a file is educational.
- Applied the classifier to the 15T tokens of FineWeb (required 6,000 H100 GPU hours).

Below is an extract from a web page. Evaluate whether the page has a high educational value and could be useful in an educational setting for teaching from primary school to grade school levels using the additive 5-point scoring system described below. Points are accumulated based on the satisfaction of each criterion:

- Add 1 point if the extract provides some basic information relevant to educational topics, even if it includes some irrelevant or non-academic content like advertisements and promotional material.
- Add another point if the extract addresses certain elements pertinent to education but does not align closely with educational standards. It might mix educational content with non-educational material, offering a superficial overview of potentially useful topics, or presenting information in a disorganized manner and incoherent writing style.
- Award a third point if the extract is appropriate for educational use and introduces key concepts relevant to school curricula. It is coherent though it may not be comprehensive or could include some extraneous information. It may resemble an introductory section of a textbook or a basic tutorial that is suitable for learning but has notable limitations like treating concepts that are too complex for grade school students.
- Grant a fourth point if the extract is highly relevant and beneficial for educational purposes for a level not higher than grade school, exhibiting a clear and consistent writing style. It could be similar to a chapter from a textbook or a tutorial, offering substantial educational content, including exercises and solutions, with minimal irrelevant information, and the concepts aren't too advanced for grade school students. The content is coherent, focused, and valuable for structured learning.
- Bestow a fifth point if the extract is outstanding in its educational value, perfectly suited for teaching either at primary school or grade school. It follows detailed reasoning, the writing style is easy to follow and offers profound and thorough insights into the subject matter, devoid of any non-educational or complex content.

The extract: <extract>.

After examining the extract:

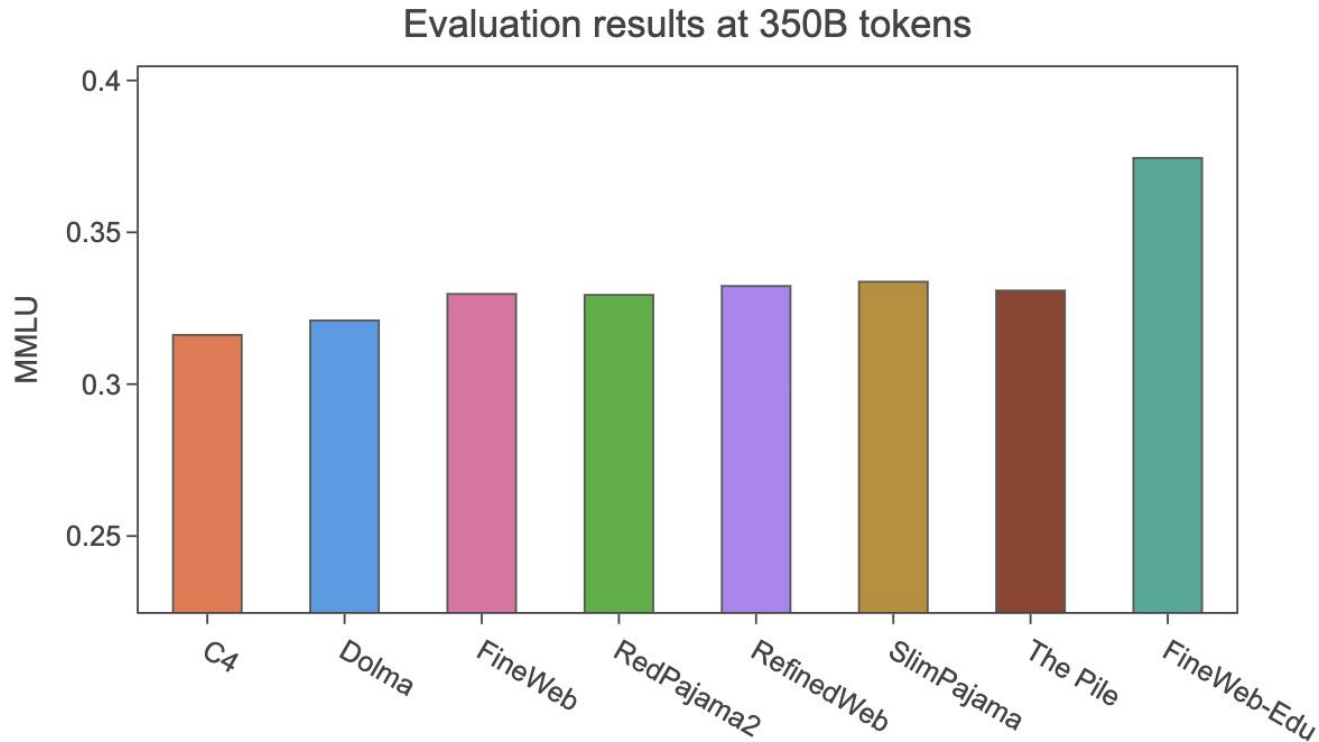
- Briefly justify your total score, up to 100 words.
- Conclude with the score using the format: "Educational score: <total points>"

Data Curation: Overview

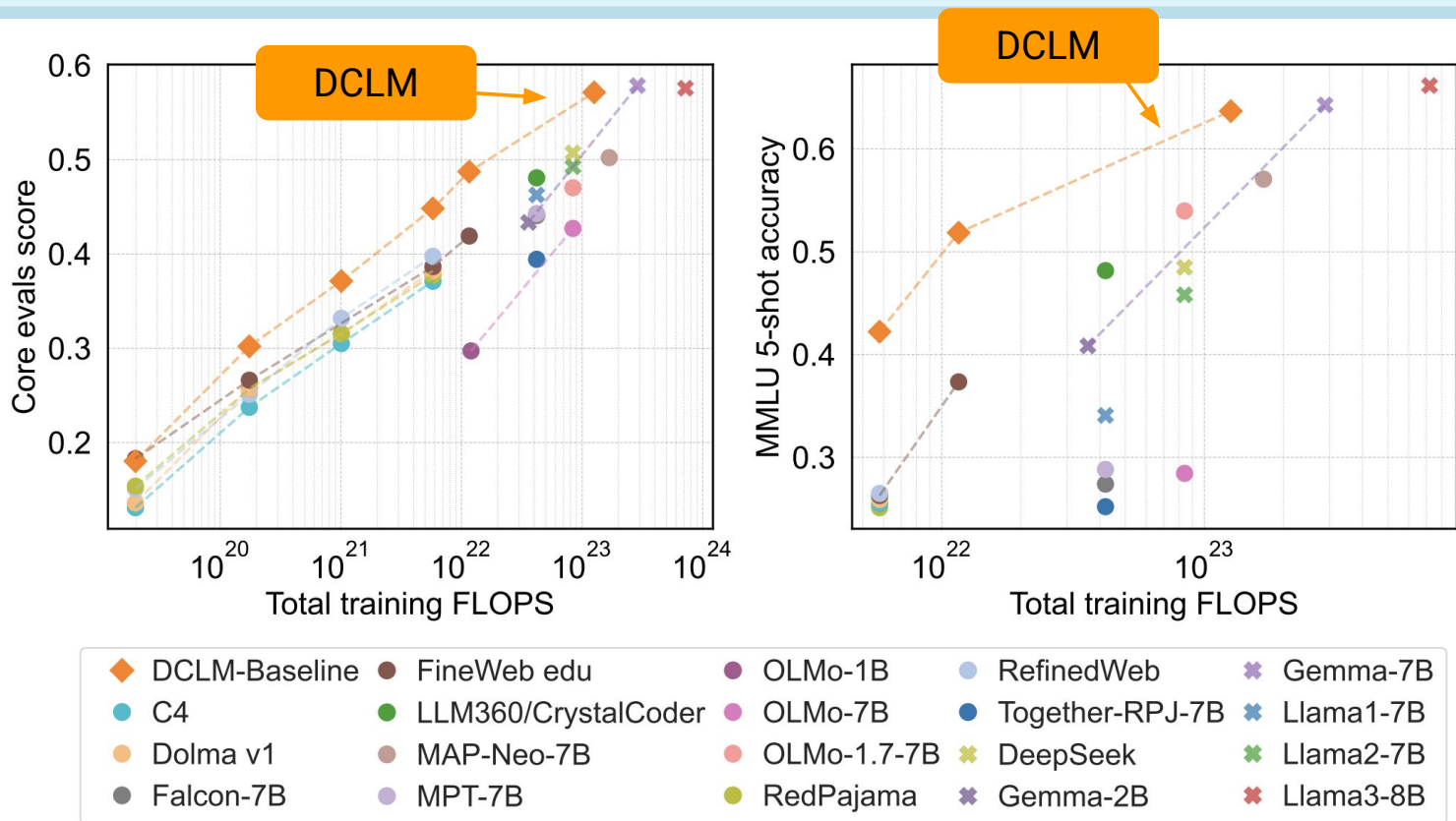
1. Acquisition: Unfiltered web-text corpus from all Common Crawl data prior to 2023
2. Linearization: Extract text from HTML → Called DCLM-Pool (240T tokens)
3. Heuristic-based filtering → 19.9% of DCLM-Pool retained
4. Deduplication → 13.7% of DCLM-Pool retained
5. Model-based filtering → 1.4% of DCLM-Pool retained → called DCLM-Baseline (3.8T tokens)

-
1. Acquisition: Unfiltered web-text corpus from all Common Crawl data (96 dumps)
 2. Linearization: Extract text from HTML
 3. Heuristic filtering I → 36T tokens
 4. Deduplication → 20T tokens
 5. Heuristic filtering II → 15T tokens → Called FineWeb
 6. Model-based filtering → 1.3T tokens → Called FineWeb-Edu

Final FineWeb-Edu



Final DCLM-Baseline: Scale to 3T-token training!



Pre-training data (Summary)

Summary (1/2)

- Common Crawl based pre-training data curation, first led by OpenAI (GPT-2/3) and Google (C4; T5), made $O(100B)$ -token datasets possible (and now, the standard is $O(10-100T)$ -token).



Case study: DCLM & FineWeb

- Every step is important – even seemingly trivial ones, like linearization and deduplication, have deep science behind them
- Science here doesn't mean fancy novel methods, but careful controlled experiments, well-chosen ablations, right choice of evaluation on benchmarks giving early, predictable signals

Summary (2/2)

- Lots of efforts building better open-source training data
 - C4, The Pile, RefinedWeb, Dolma, etc...
 - DCLM and FineWeb: Beating Llama3 and Gemma at a 7B scale (on benchmarks in the paper)
- These efforts are building on top of each other
 - {C4, RefinedWeb, etc} led to {DCLM, Fineweb}
 - {DCLM, Fineweb} led to {OLMo 2, etc}
 - More recently: Olmo 3, Nemotron, Marin, ...

Discussion questions

- If model-based filtering is ultimately applied, why heuristic filtering at all? Wouldn't model-based methods always be stronger?
- Both papers did heuristic filtering → deduplication → model-based filtering. Is it a coincidence?
- DCLM shows that its "AskLLM" baseline underperforms a simple fastText classifier, while FineWeb's "AskLLM"-like approach is highly effective. What's the difference?
- Before discussing DCLM/FineWeb papers, we covered other data sources in "Acquisition" (books, math, and code). Would the same heuristic/model-based filtering pipeline apply? How might filtering strategies differ by source type?
- In 2019, C4 intentionally removed code. Now, we're intentionally curating code data. What changed?
- If infinite compute (token budget) is given, do we still want to do filtering?