



FineWeb: decanting the web for the finest text data at scale

larships,

Commodities/Services Provision

Accommodation

Nikhil and Yaohan

Politics

Philosophical/Spiritual Introspection

Theology

What is FineWeb & FineWeb-Edu ?

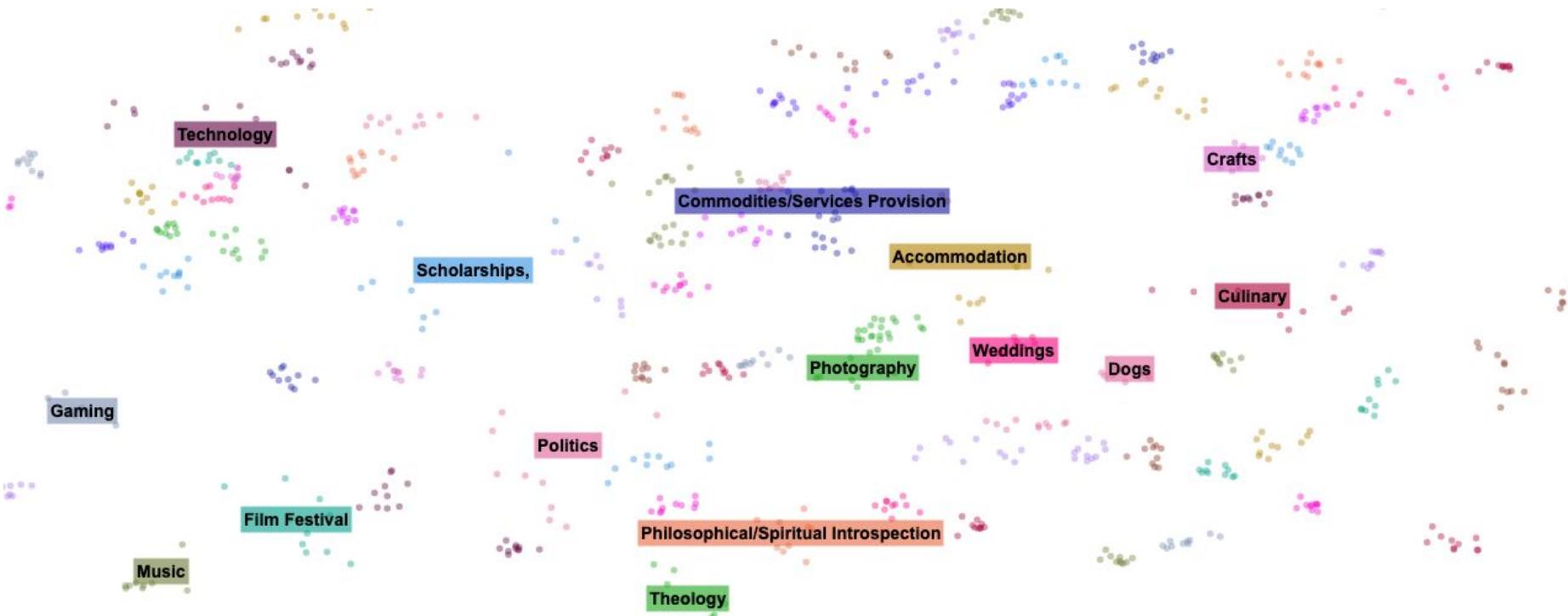
FineWeb

- A new, large-scale (**15-trillion tokens, 44TB disk space**) dataset for **LLM pretraining**.
- Derived from 96 [CommonCrawl](#) snapshots and produces **better-performing LLMs than other open pretraining datasets**.

FineWeb-Edu

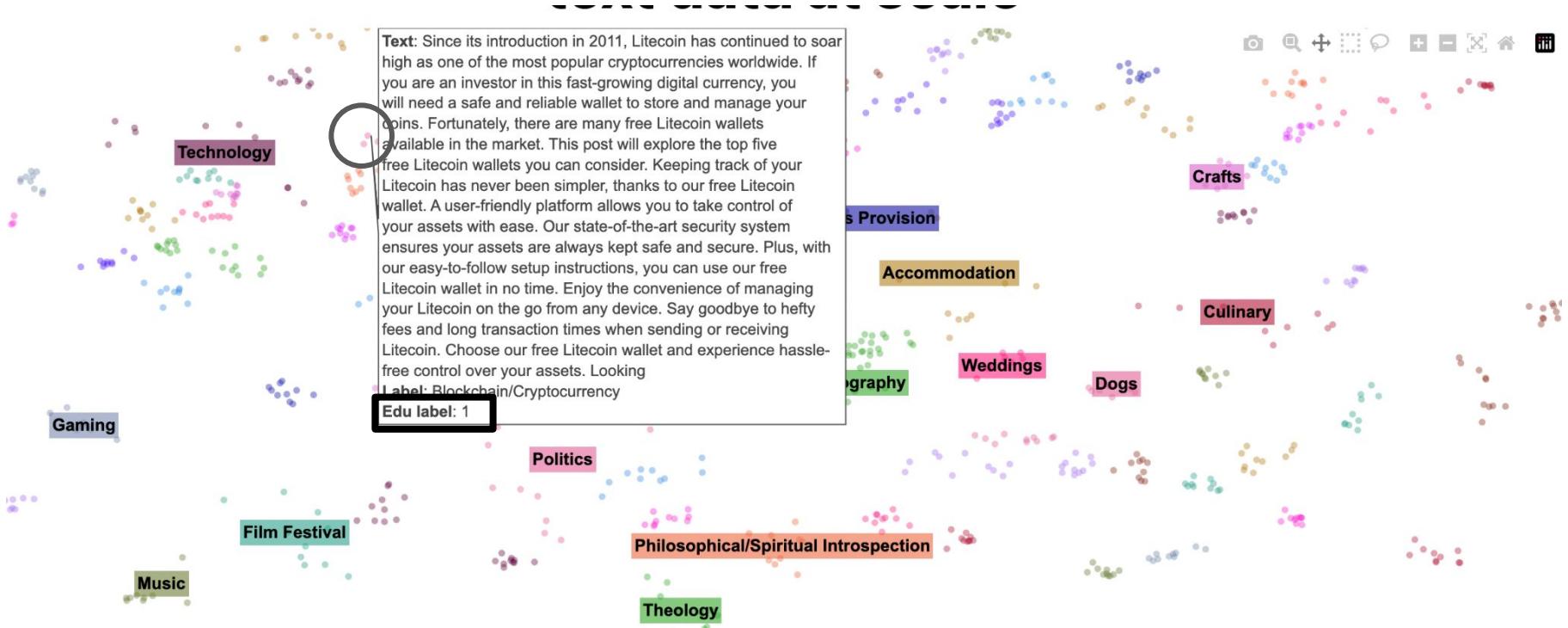
- A **subset** of FineWeb constructed using scalable automated high-quality annotations for **educational value**.
- **Outperforms all** openly accessible web-datasets on a number of **educational benchmarks** such as MMLU, ARC, and OpenBookQA

What is FineWeb & FineWeb-Edu ?



The  FineWeb dataset, clustered and annotated with educational score labels

What is FineWeb & FineWeb-Edu ?



The  FineWeb dataset, clustered and annotated with educational score labels

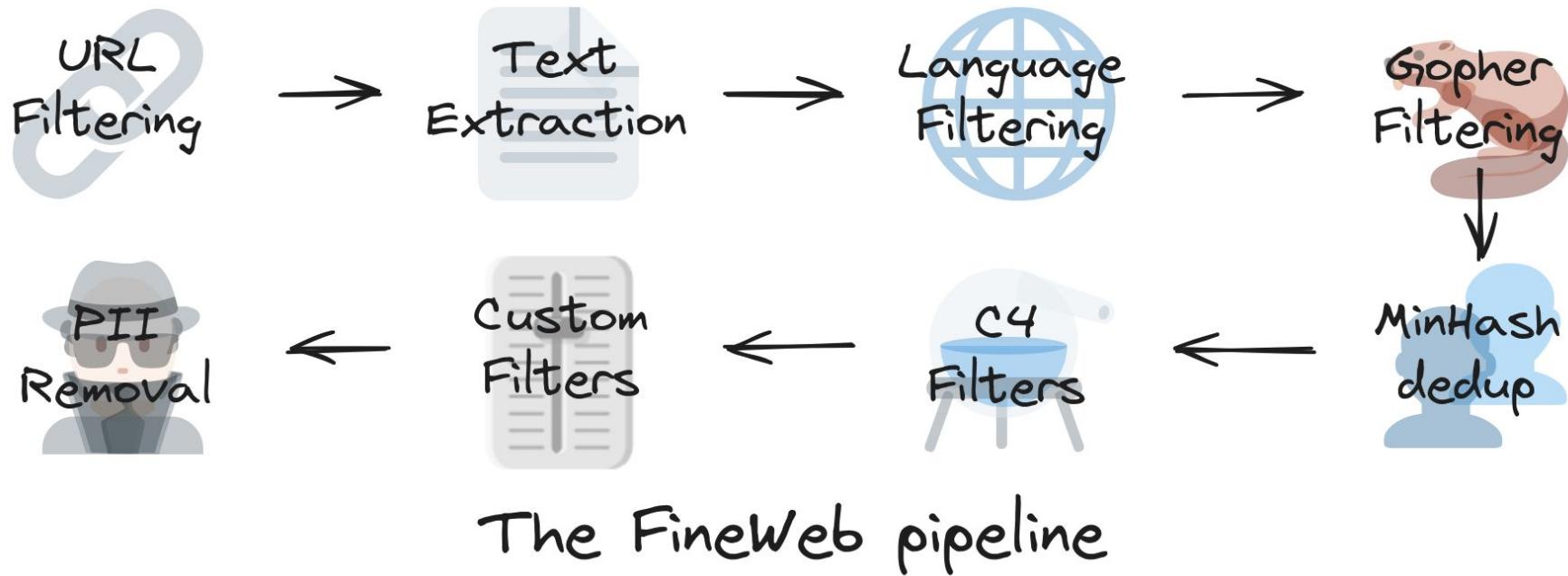
Why do we need such datasets?



Why do we need such datasets?



The 🍷 FineWeb recipe



Text Extraction from ...?

CommonCrawl!

WARC (Web ARChive format)

Encapsulate



trafilatura

Text

WET (WARC Encapsulated Text)



too much boilerplate
and menu text

Base Filtering

Remove adult content



URL filtering

English only



Language filtering

Quality and repetition filters



Gopher filtering

Deduplication

MinHash: a fuzzy hash-based deduplication technique

- Collect each document's **5-grams**
- Compute MinHashes using **112 hash functions** in total, split into **14 buckets of 8 hashes** each
- Documents are matched if they have **the same 8 minhashes** in at least one of the 14 buckets

Deduplication

MinHash

	Hash Func 1	Hash Func 2	Hash Func 3	Hash Func 4	Hash Func 5	Hash Func 6	Hash Func 7	Hash Func 8	Hash Func 9	Hash Func 10	Hash Func 112
5-gram 1													
5-gram 2			Min								Min		
5-gram 3	Min				Min				Min				
5-gram 4				Min				Min		Min			
...						Min							
...							Min						
...													

Score

Deduplication

Apply MinHash **globally**

- Applied to all 96 snapshots chronologically
- Removed up to 90% of data in oldest snapshots ; 4 trillion tokens left
- Unexpectedly **lower** performance than Refined-Web

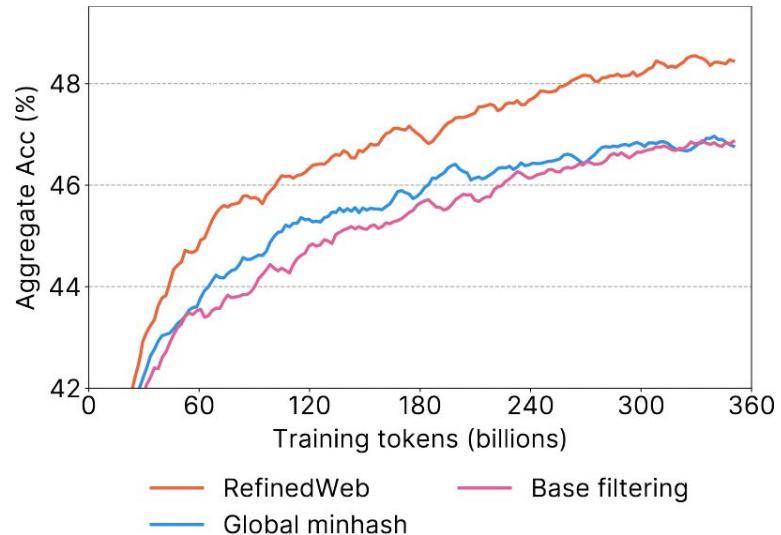


Figure 3: **Global minhash deduplication study.** Applying minhash deduplication globally to the dataset provides only a modest performance uplift, with the resulting model far behind one trained on Refined-Web.

Deduplication

Apply MinHash **globally**

- Applied to all 96 snapshots chronologically
- Removed up to 90% of data in snapshots ; 4 trillion tokens left
- Unexpectedly **lower** performance than Refined-Web

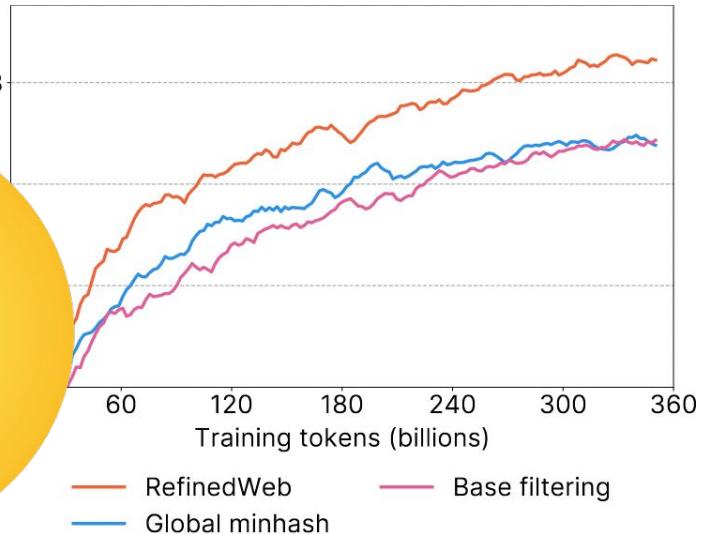
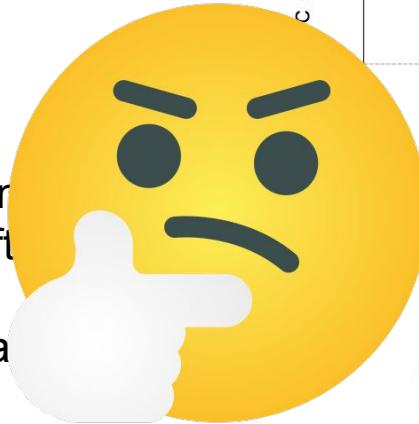


Figure 3: **Global minhash deduplication study.** Applying minhash deduplication globally to the dataset provides only a modest performance uplift, with the resulting model far behind one trained on Refined-Web.

Deduplication

Why lower? Another Experiment

- Originally **kept** data:
Kept after global minhash(31 billion tokens)
- Originally **removed** data:
Removed data by global minhash(460 billion)

↓

Individual minhash
(Deduplicate each snapshot independently from the other crawls)

171 billion tokens

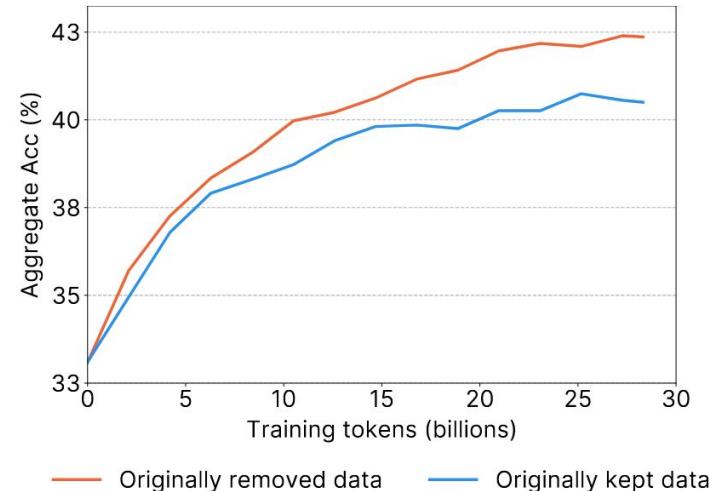


Figure 4: **2013-48 global minhash impact study.** Global deduplication upsamples lower-quality data in the last deduplicated crawl, resulting in worse performance of the retained data compared to the removed data.

Deduplication

Globally VS Individually? **Individually !**

Apply MinHash **individually**

- Applied to each snapshot independently
- Resulted in 20 trillion tokens
- **Matched** RefinedWeb's performance

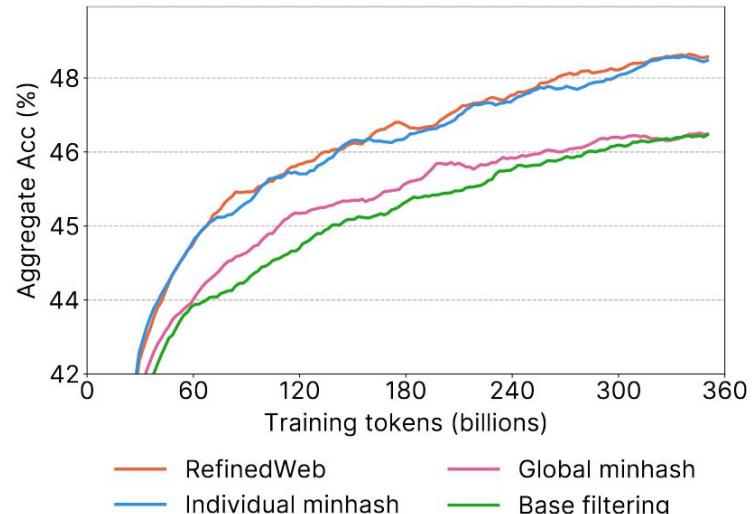
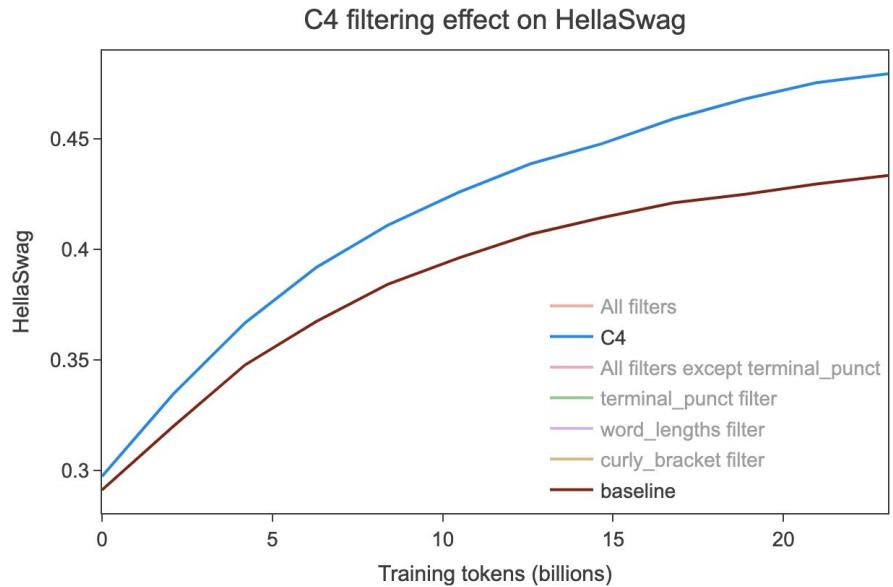


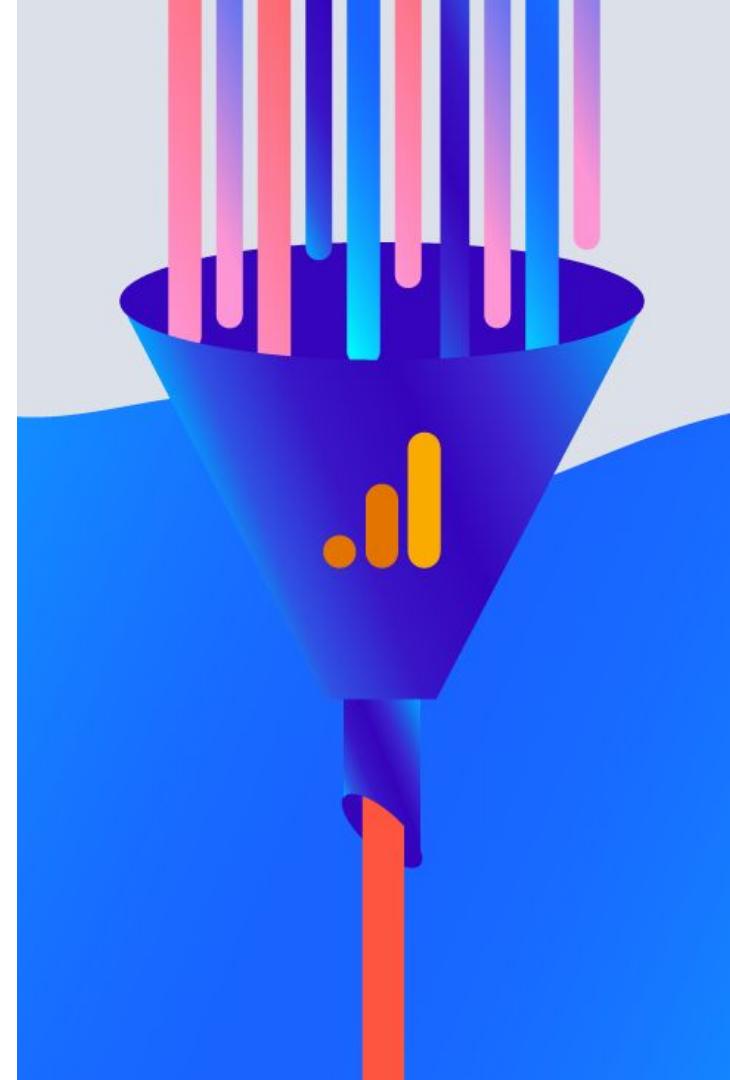
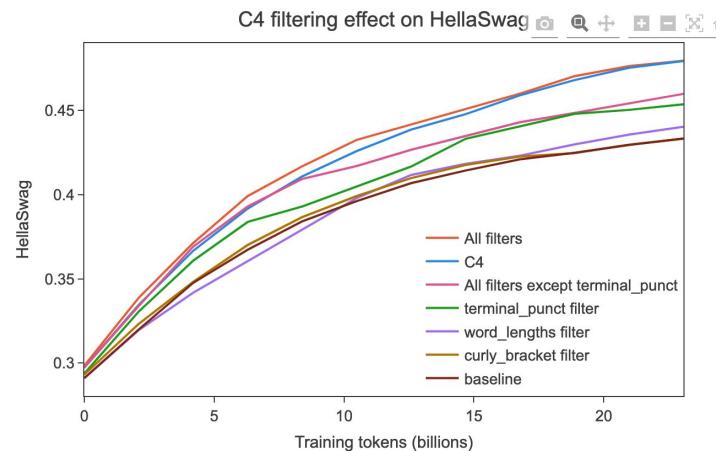
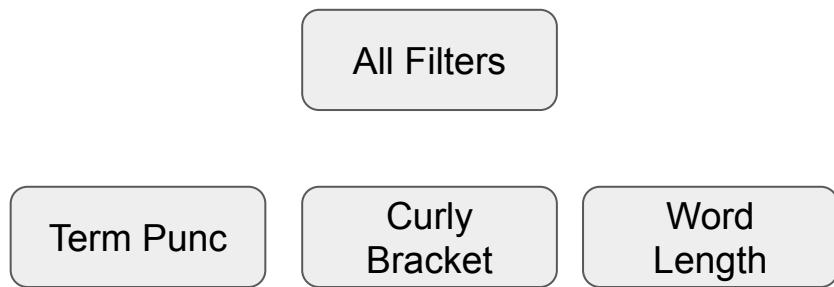
Figure 5: **Individual minhash deduplication study.** Unlike Global minhash, deduplicating individually improves the average score.

But wait, something perplexing was seen

C4^[1] a heavily filtered dataset was still performing better than base filtering and independent MinHash



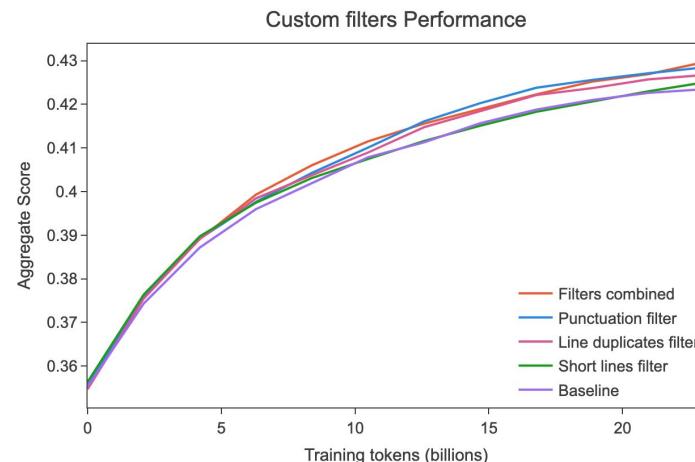
Quality Filtering - Bridging gap to C4



Three custom filters through a statistical approach

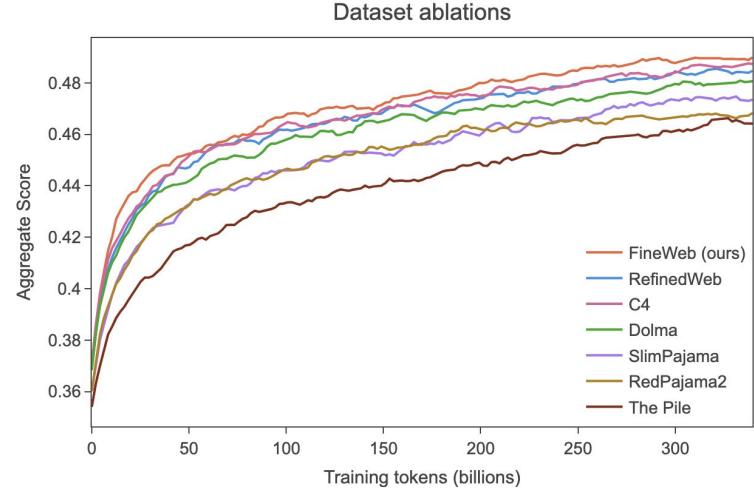
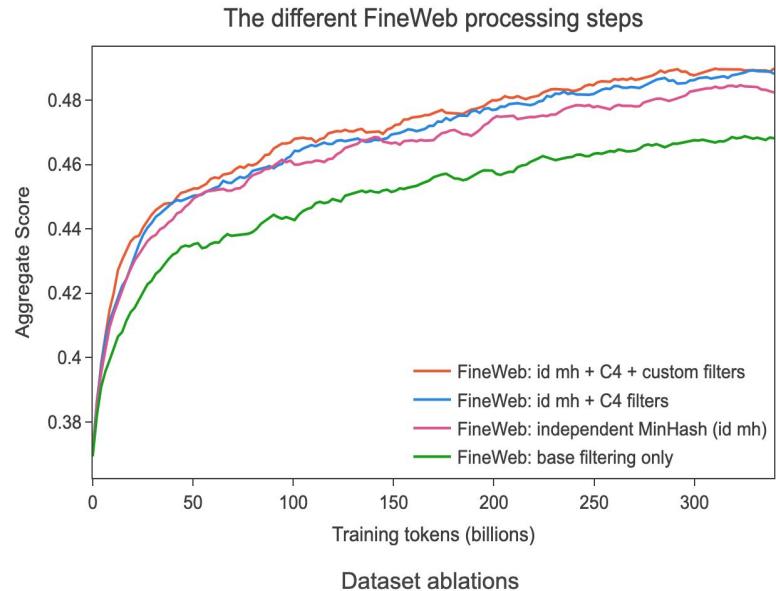


- Remove documents where the fraction of lines ending with punctuation ≤ 0.12 (10.14% of tokens removed) – vs the 30% from the original C4 terminal punct filter
- Remove documents where the fraction of characters in duplicated lines ≥ 0.1 (12.47% of tokens removed) – the original MassiveText threshold for this ratio is ≥ 0.2
- Remove documents where the fraction of lines shorter than 30 characters ≥ 0.67 (3.73% of tokens removed)



The final recipe

- ✓ Surpass C4
- ✓ Have larger corpus
- ✓ Checkpoints every 1000 steps
- ✓ Highest performing models on any open dataset



But there is more - Fineweb Edu

Why?

The popular Phi3 models were trained on 3.3 and 4.8 trillion tokens, with the paper [36] stating:

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.

Similarly, Llama 3 blog post [37] notes:

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.

The synthetic data bit..



PARITY

Synthetic
bootstrapping



Annotation

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 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>"

Classifier

Arctic-Embed: Scalable, Efficient, and Accurate Text Embedding Models

Luke Merrick^{1,*}, Danmei Xu¹, Gaurav Nuti¹, and Daniel Campos¹

¹Snowflake Inc.

* Corresponding author, luke.merrick@snowflake.com

Abstract

This report describes the training dataset creation and recipe behind the family of arctic-embed text embedding models (a set of five models ranging from 22 to 334 million parameters with weights open-sourced under an Apache-2 license). At the time of their release, each model achieved state-of-the-art retrieval accuracy for models of their size on the MTEB Retrieval leaderboard,¹ with the largest model, arctic-embed-l outperforming closed source embedding models such as Cohere's embed-v3 and Open AI's text-embed-3-large. In addition to the details of our training recipe, we have provided several informative ablation studies, which we believe are the cause of our model performance.

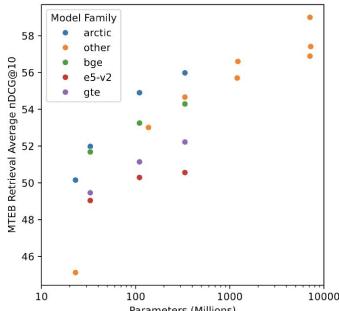


Figure 1: Snowflake's Arctic-embed models are a suite of 5 embedding models, each of which pushes the Pareto frontier in the trade-off between model size and retrieval performance on the MTEB Retrieval Leaderboard.

1 Introduction

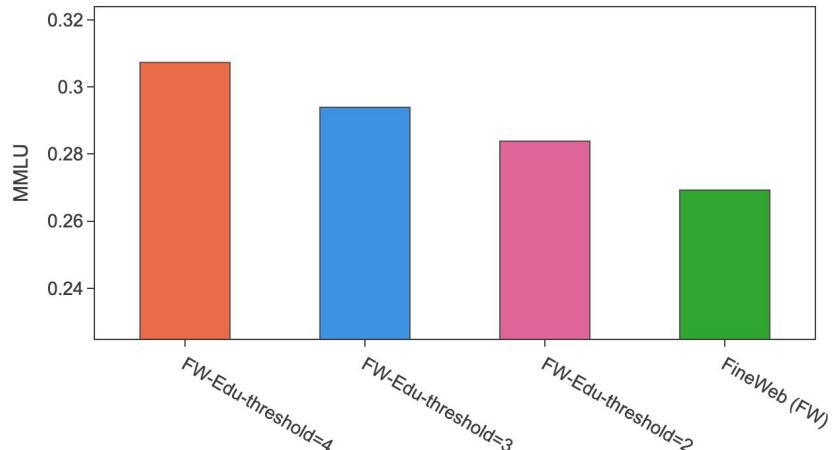
Embedding models' ability to provide accurate retrieval performance without additional tuning

	precision	recall	f1-score	support
0	0.75	0.49	0.59	5694
1	0.78	0.84	0.81	26512
2	0.57	0.61	0.59	10322
3	0.56	0.50	0.53	3407
4	0.58	0.35	0.44	807
5	0.33	0.01	0.02	125
accuracy			0.71	46867
macro avg	0.60	0.47	0.50	46867
weighted avg	0.71	0.71	0.71	46867

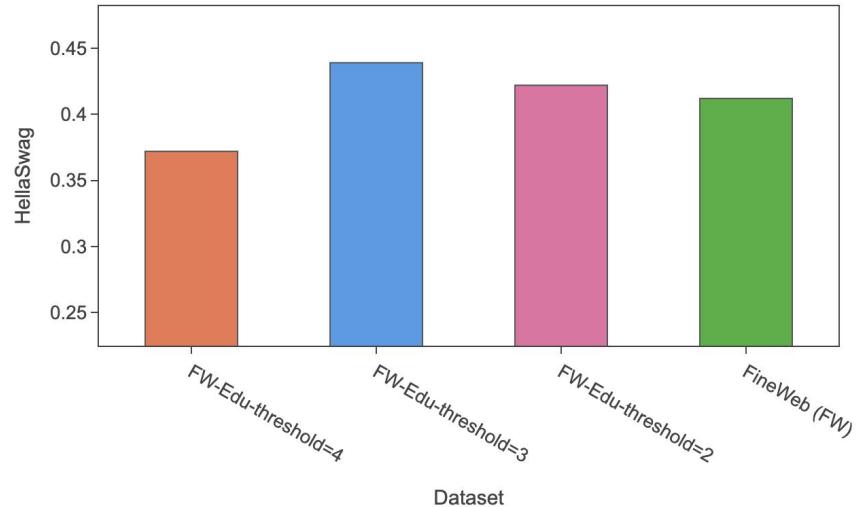
Filtering

1.82B model trained on 8B tokens

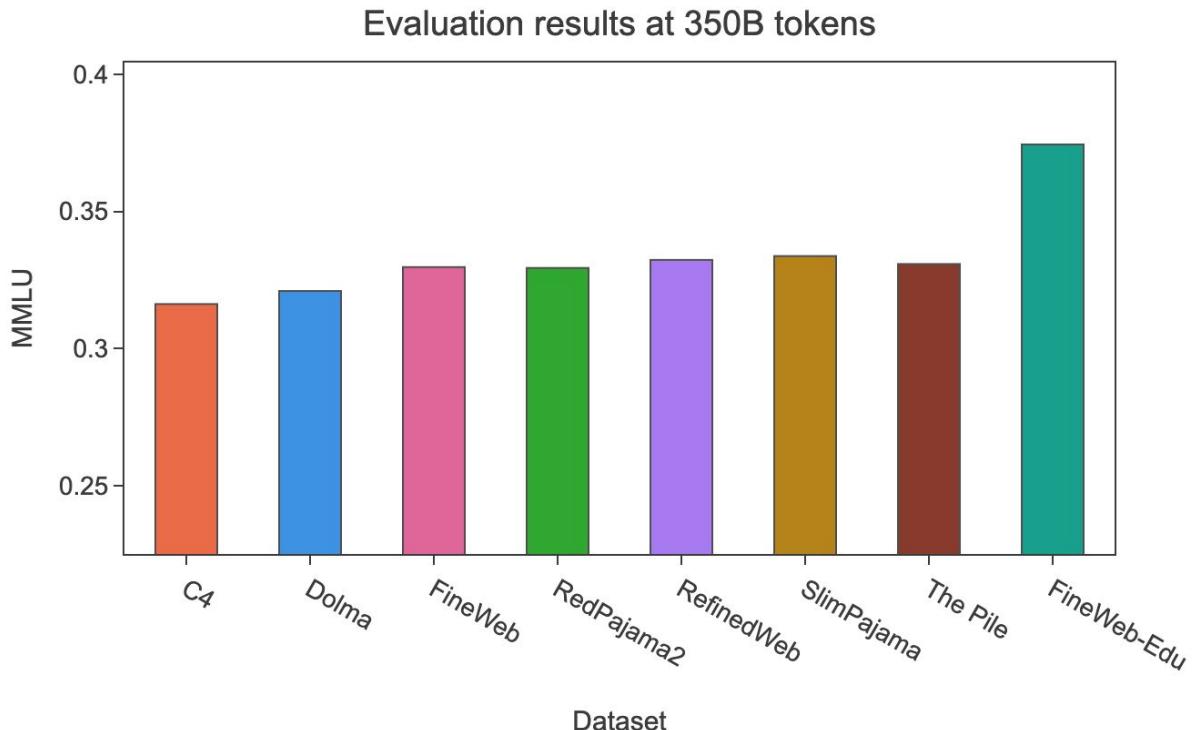
FineWeb-Edu thresholding



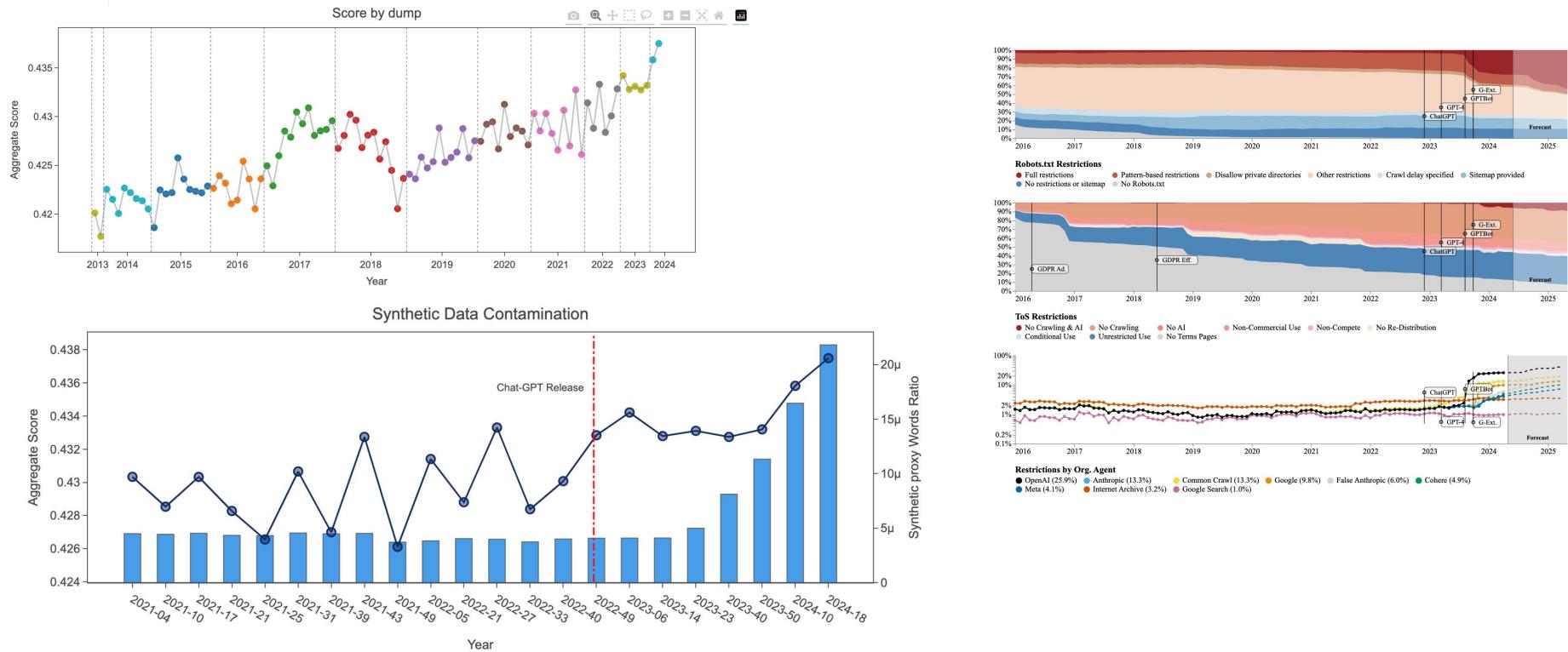
FineWeb-Edu thresholding



Result



Current state - Bonus



Thoughts and takeaways

- So the crawling content is decreasing a lot but so what?
 - Why do we still need crawling? Don't we have all the data we already need?
 - Maybe because we have a very skewed crawl: asia and africa young population who just recently gained access not included
- Larger companies with already good pretrained models will use bootstrapping is the future as seen by Llama3 circumventing the need of crawls.
- Important to have open science to save millions and democratize technology

Other interesting projects going on

OLMo: Accelerating the Science of Language Models

Published on Feb 1 · ★ Submitted by akhalig on Feb 2 #1 Paper of the day

Authors:  [Dirk Groeneweld](#),  [Iz Beltagy](#), [Pete Walsh](#),  [Akhilta Bhagia](#), [Rodney Kinney](#),  [Oyvind Tafjord](#),  [Ananya Harsh Jha](#),  [Hamish Ivison](#),  [Jan Magnusson](#),  [Yizhong Wang](#), [Shane Arora](#), [David Atkinson](#),  [Russell Author](#),  [Khaythi Raghuvi Chandu](#),  [Arman Cohen](#), [Jennifer Dumas](#),  [Yanai Elazar](#),  [Yuling Gu](#),  [Jack Hessel](#),  [Tushar Khot](#), [William Merrill](#),  [Jacob Morrison](#) +21 authors

Abstract

Language models (LMs) have become ubiquitous in both NLP research and in commercial product offerings. As their commercial importance has surged, the most powerful models have become closed off, gated behind proprietary interfaces, with important details of their training data, architectures, and development undisclosed. Given the importance of these details in scientifically studying these models, including their biases and potential risks, we believe it is essential for the research community to have access to powerful, truly open LMs. To this end, this technical report details the first release of OLMo, a state-of-the-art, truly Open Language Model and its framework to build and study the science of language modeling. Unlike most prior efforts that have only released model weights and inference code, we release OLMo and the whole framework, including training data and training and evaluation code. We hope this release will empower and strengthen the open research community and inspire a new wave of innovation.

