Pre-training (Advanced Topics)

CS 288 Spring 2026

UC Berkeley

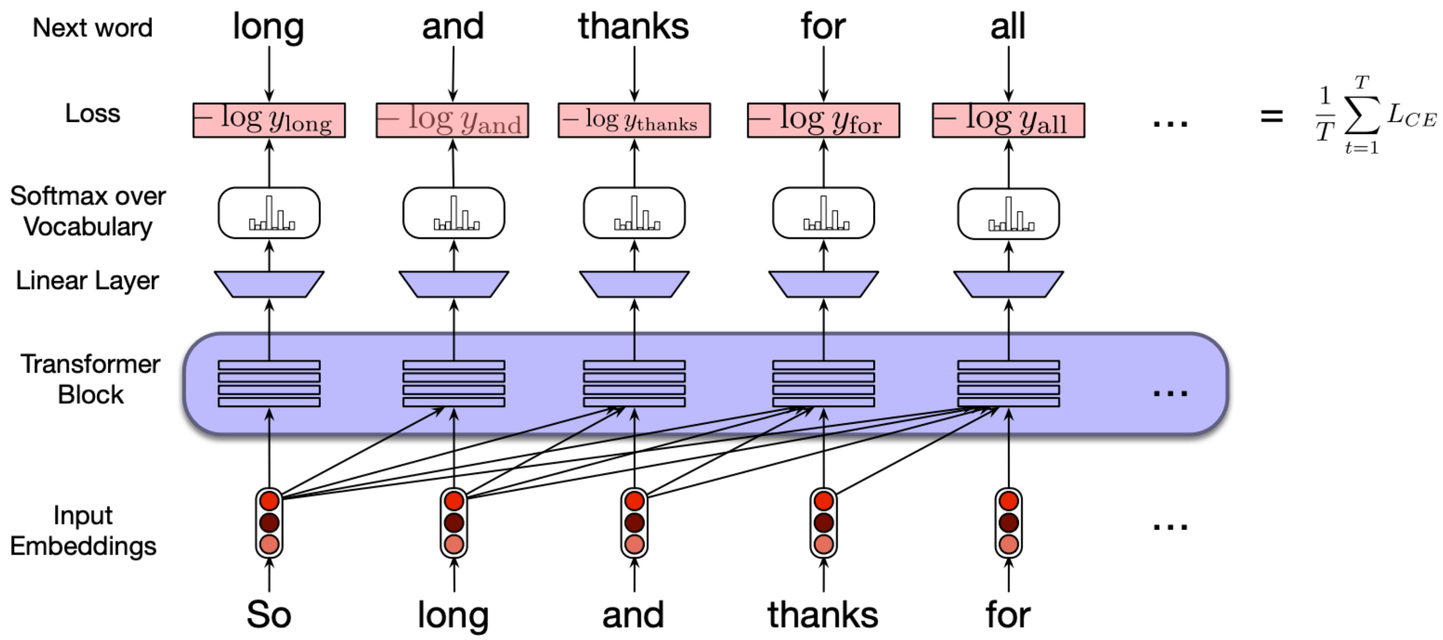
cal-cs288.github.io/sp26

Berkeley EECS

BAIR

Today: Pre-training advanced topics

- Fixed: Training objective (generative language modeling), architecture (decoder-only Transformers)



(Radford et al, 2018): Improving Language Understanding by Generative Pre-Training

Today: Pre-training advanced topics

- Fixed: Training objective (generative language modeling), architecture (decoder-only Transformers)
- Are we done then? No!
    - Science of scaling = *Scaling laws*
        - Training longer is as important as scaling the model size
    - How do we train longer? You need a massive text corpus. Where do we get it? = *Training data curation*

Today: Pre-training advanced topics

- Fixed: Training objective (generative language modeling), architecture (decoder-only Transformers)
- Are we done then? No!
    - Science of scaling = ***Scaling laws***
        - Training longer is as important as scaling the model size
    - How do we train longer? You need a massive text corpus. Where do we get it? = ***Training data curation***

Much of this is contemporary research!

Scaling Laws

Scaling laws = Scaling is all you need?

No! Actually, it is opposite = Scaling is not all you need

- You can do a totally wrong scaling
- You should do smarter (informed) scaling
- Scaling laws are all about
    - How even OpenAI and Google did wrong scaling before figuring out scaling laws
    - how to do informed, predictable scaling

Why scaling laws?

- Training GPT-3 costs millions of dollars. Failed YOLO runs are expensive!
- For example, Qwen3 models were trained on 36T tokens (~10^{24} FLOPs).
- Model builders (even GPU-rich) don't have infinite compute.

What are practical strategies to allocate resources to reduce train costs?

Scaling Laws: Let's predict bigger runs from smaller runs!

Scaling Laws from Baidu (2017)

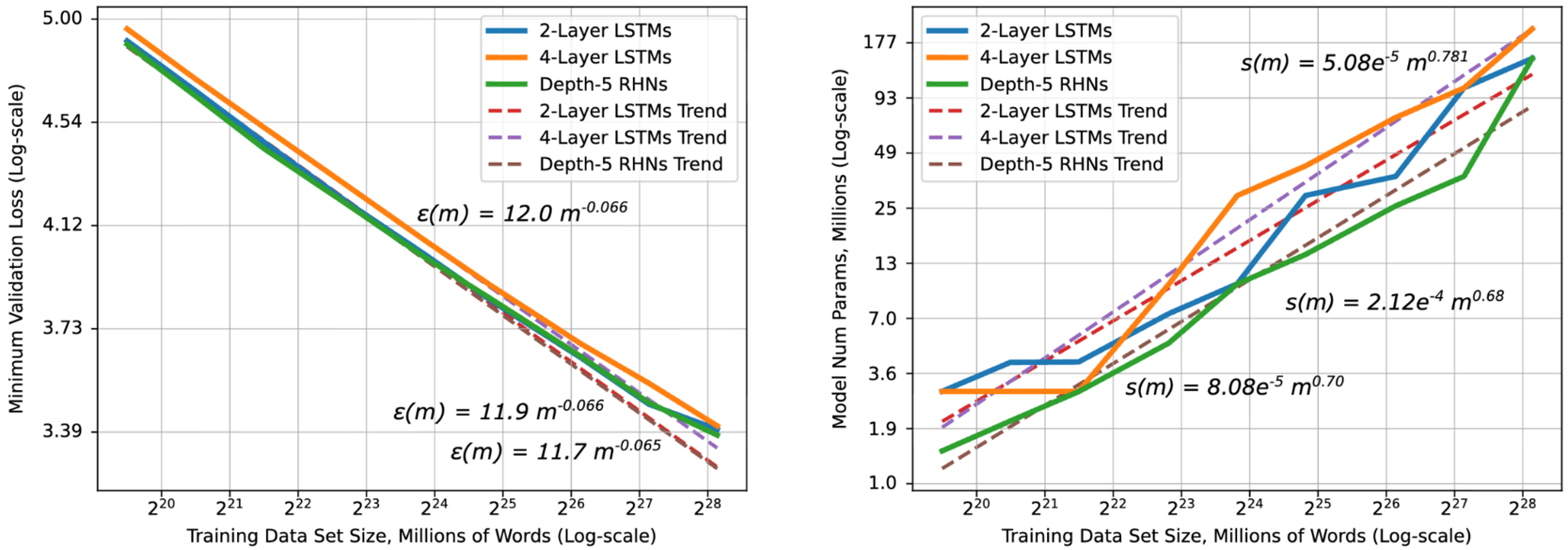


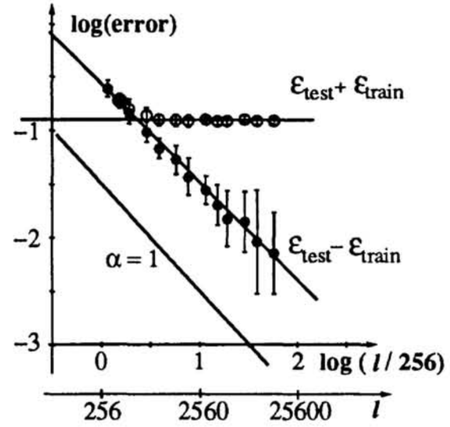
Figure 2: Learning curve and model size results and trends for word language models.

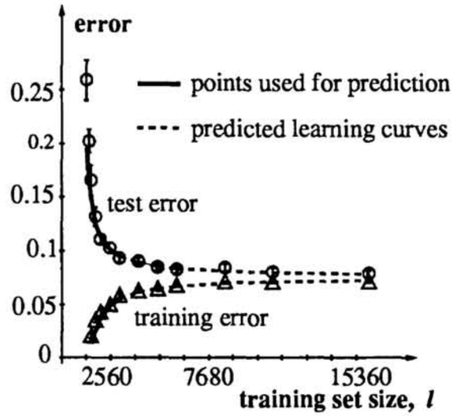
Deep Learning Scaling is Predictable, Empirically (Hestness et al., 2017)

(30 years earlier) Prototypical Scaling Laws

Learning Curves: Asymptotic Values and Rate of Convergence

Corinna Cortes, L. D. Jackel, Sara A. Solla, Vladimir Vapnik, and John S. Denker
AT&T Bell Laboratories
Holmdel, NJ 07733





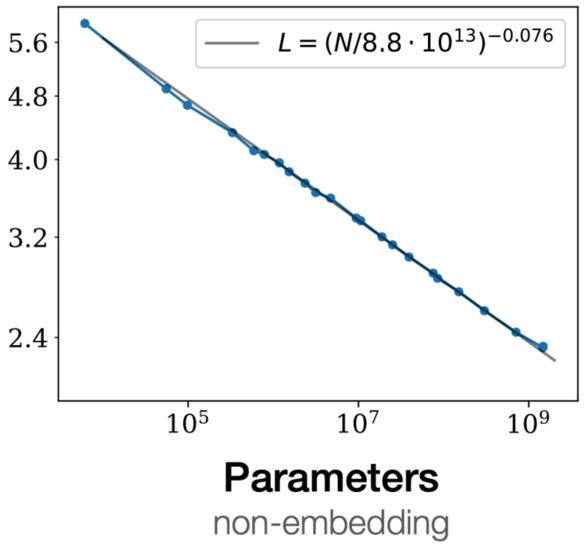
Scaling laws in LLMs

- (Kaplan et al., 2020) Scaling Laws for Neural Language Models → Kaplan scaling laws (OpenAI) 4 months before GPT-3
- (Hoffmann 2022) Training Compute-Optimal Large Language Models → Chinchilla scaling laws (Google DeepMind) 2 years after GPT-3

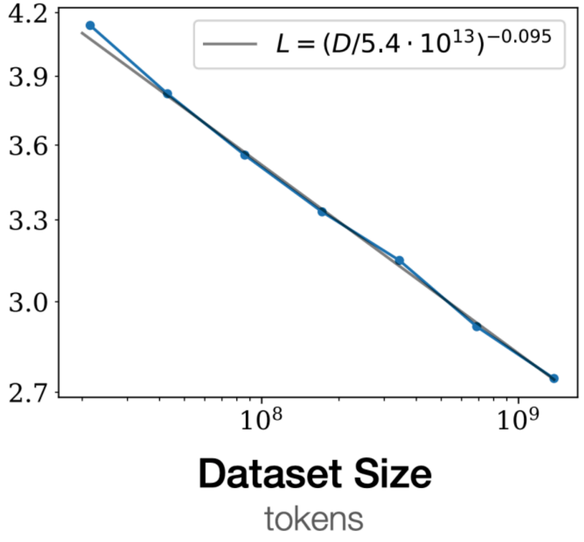
Standard scaling law recipe

1. Grid search w/ different parameters, varying two key scaling dimensions: model size (parameter count) & training length (data size)
2. Identify the Pareto frontier, i.e., given the fixed training budget, what is the best config?
3. Fit simple models to the results to make trends predictable beyond the tested scale
4. Show it extrapolates!

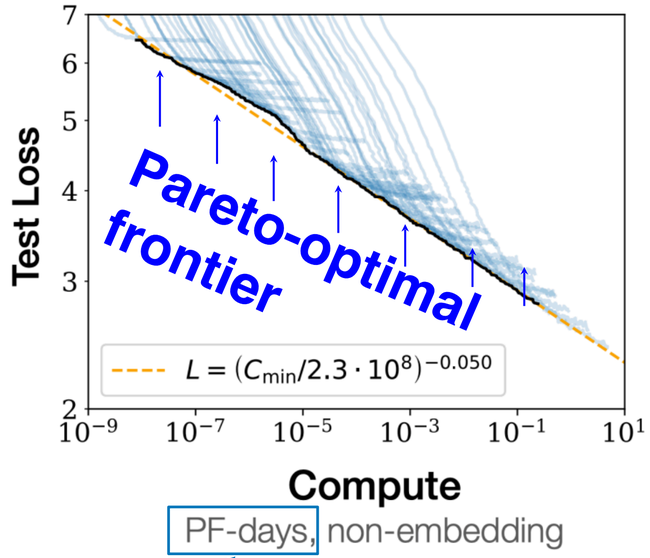
Kaplan et al's scaling laws: Claim 1



$\text{Loss} \propto (\text{Model params})^{-\gamma}$



$\text{Loss} \propto (\text{Data})^{-\beta}$

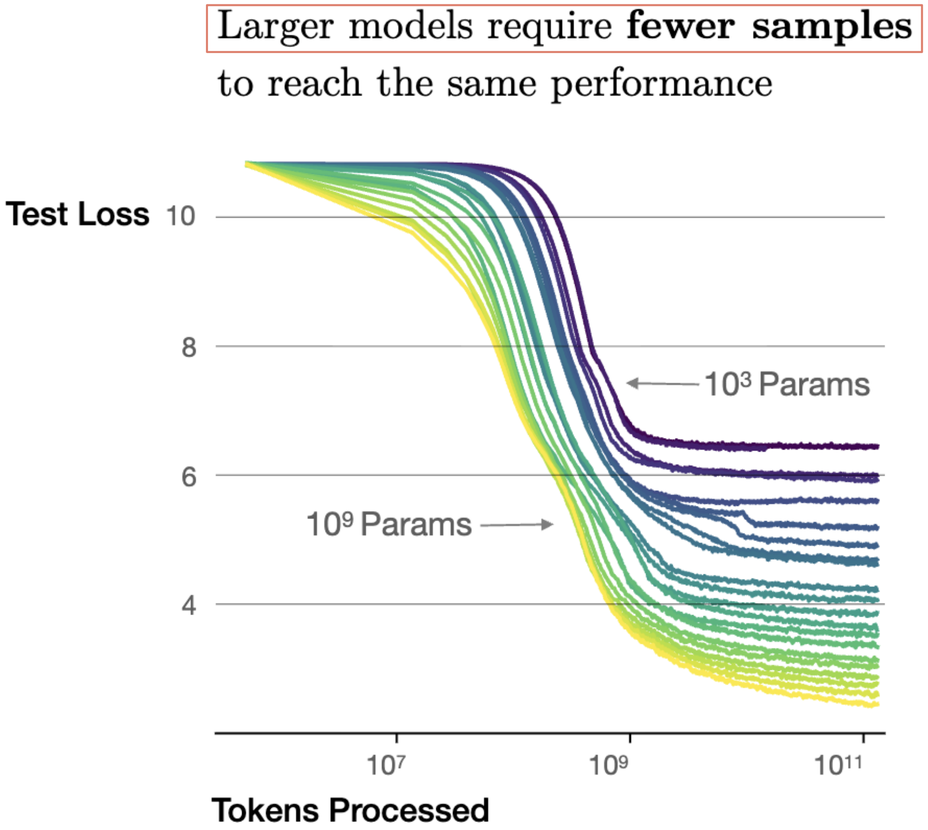


$\text{Loss} \propto (\text{Compute})^{-\alpha}$

PF-day = (1PetaFLOPs/second) * (3600s/hr) * (24hr)

Claim 1: Loss goes down predictably wrt compute, data, model size!

Kaplan et al's scaling laws: Claim 2



Claim 2: Overfitting is harder than we thought

Scaling Laws for Neural Language Models (Kaplan et al., 2020)

Kaplan et al's scaling laws: Putting things together

1. With all datapoints, jointly fit ***N*** (# params) and ***D*** (# tokens) according to:

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

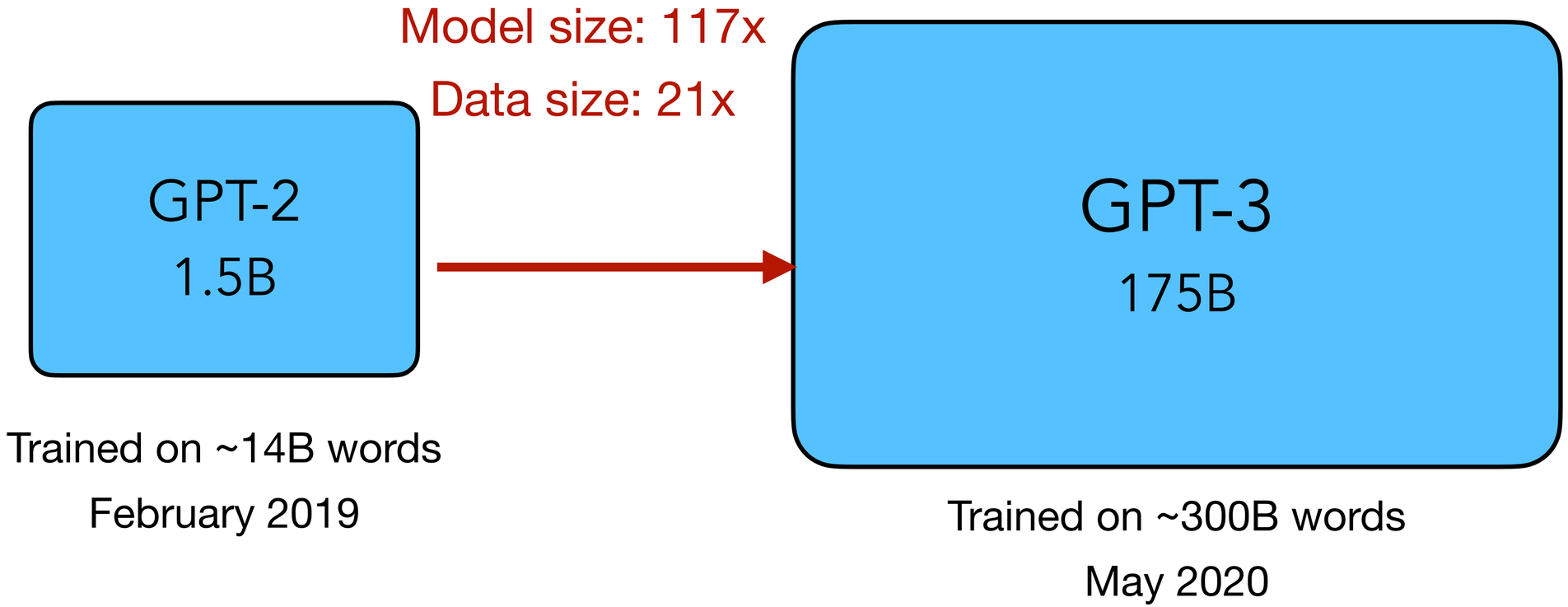
where ***E*** denotes irreducible loss.

2. Obtain values for ***A***, ***B***, ***α***, ***β***, and ***E*** by regression optimization.
3. Now given your desired FLOPs (=6***ND***), find your best ***N*** and ***D*** via math.

Conclusion: $D \propto N^{0.27}$!!

Translation: If you have 10x more compute, increase model size 6.2x and data size 1.6x

Outcome of Kaplan et al. = GPT-3!



Didn't exactly follow $D \propto N^{0.27}$ but followed the sprit of it:
prioritizing model size over data size

Chinchilla laws: Background

Google trying to reproduce GPT-3

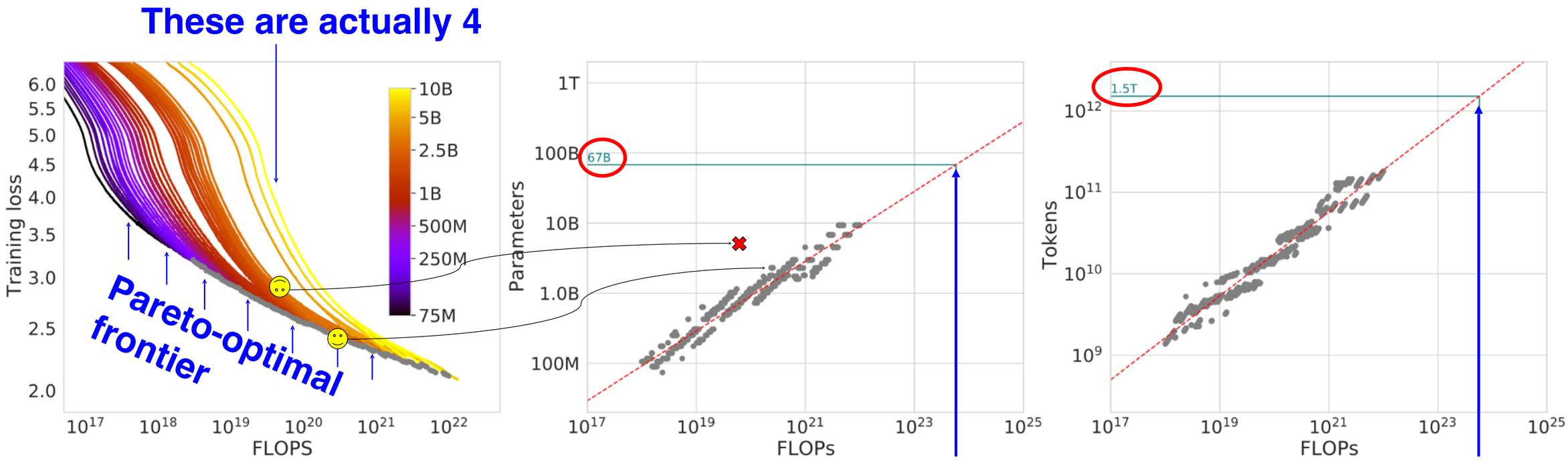
They created Gopher (December 2021), a 280-billion model, not as good as GPT-3

280B 175B

	Method	*Gopher*	GPT-3
Natural Questions (dev)	0-shot	10.1%	14.6%
	5-shot	24.5%	-
	64-shot	28.2%	29.9%
TriviaQA (unfiltered, test)	0-shot	52.8%	64.3 %
	5-shot	63.6%	-
	64-shot	61.3%	71.2%
TriviaQA (filtered, dev)	0-shot	43.5%	-
	5-shot	57.0%	-
	64-shot	57.2%	-

“What’s wrong?!? Let’s fit our own scaling laws”

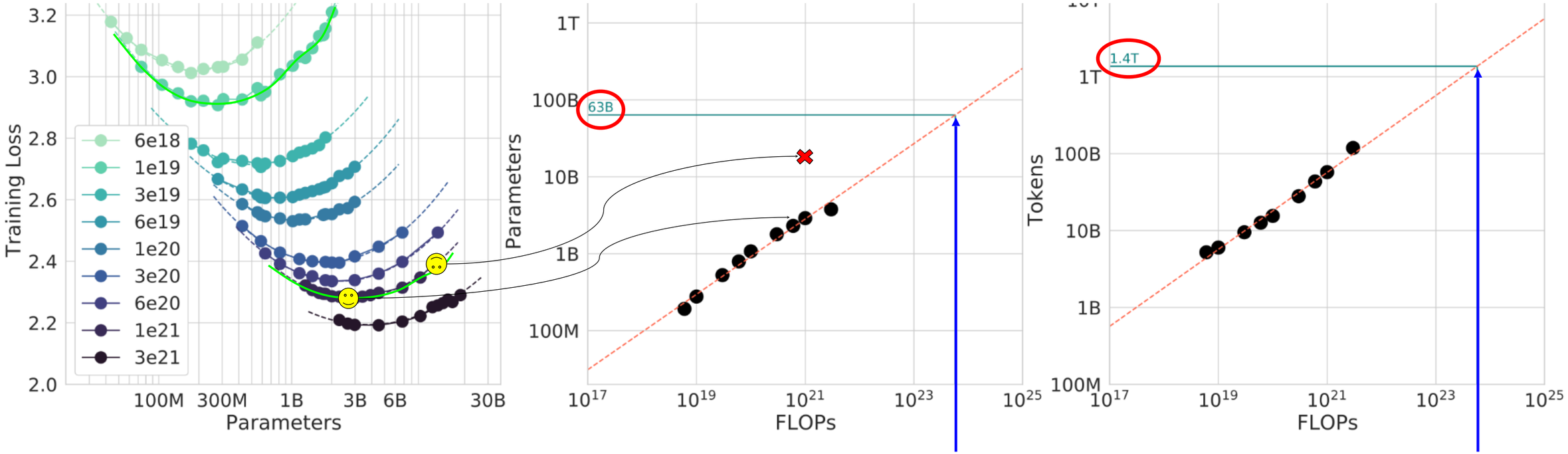
Chinchilla laws: Approach 1



Same compute as Gopher (~GPT-3 compute)

Chinchilla laws: Approach 2

1. Vary model size ***N*** for fixed set of FLOPs (=6***ND***)
2. For best models in each Iso-FLOPs group, fit the line again!



Same compute as Gopher (~GPT-3 compute)

Chinchilla laws: Putting things together

1. With all datapoints, jointly fit ***N*** (# params) and ***D*** (# tokens) according to:

$$\hat{L}(N, D) \triangleq E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

where ***E*** denotes irreducible loss.

2. Obtain values for ***A***, ***B***, ***α***, ***β***, and ***E*** by regression optimization.
3. Now given your desired FLOPs (=6***ND***), find your best ***N*** and ***D*** via math.

Conclusion: ~~$D \propto N^{0.27}$~~ $D \propto N$!!

Translation: If you have 10x more compute, increase model size ~~6.2x~~ **3.2x** and data size ~~1.6x~~ **3.2x**

More concretely: You need (# of tokens) = **20** x (# of params)

Chinchilla conclusions

Conclusion: ~~$D \propto N^{0.27}$~~ $D \propto N$!!

Translation: If you have 10x more compute, increase model size ~~6.2x~~ **3.2x** and data size ~~1.6x~~ **3.2x**

More concretely: You need (# of tokens) = **20** x (# of params)

GPT-3

175B

Trained on ~300B words

May 2020

→ It should have been 175 billion x 20

= **3.5 trillion tokens**?!

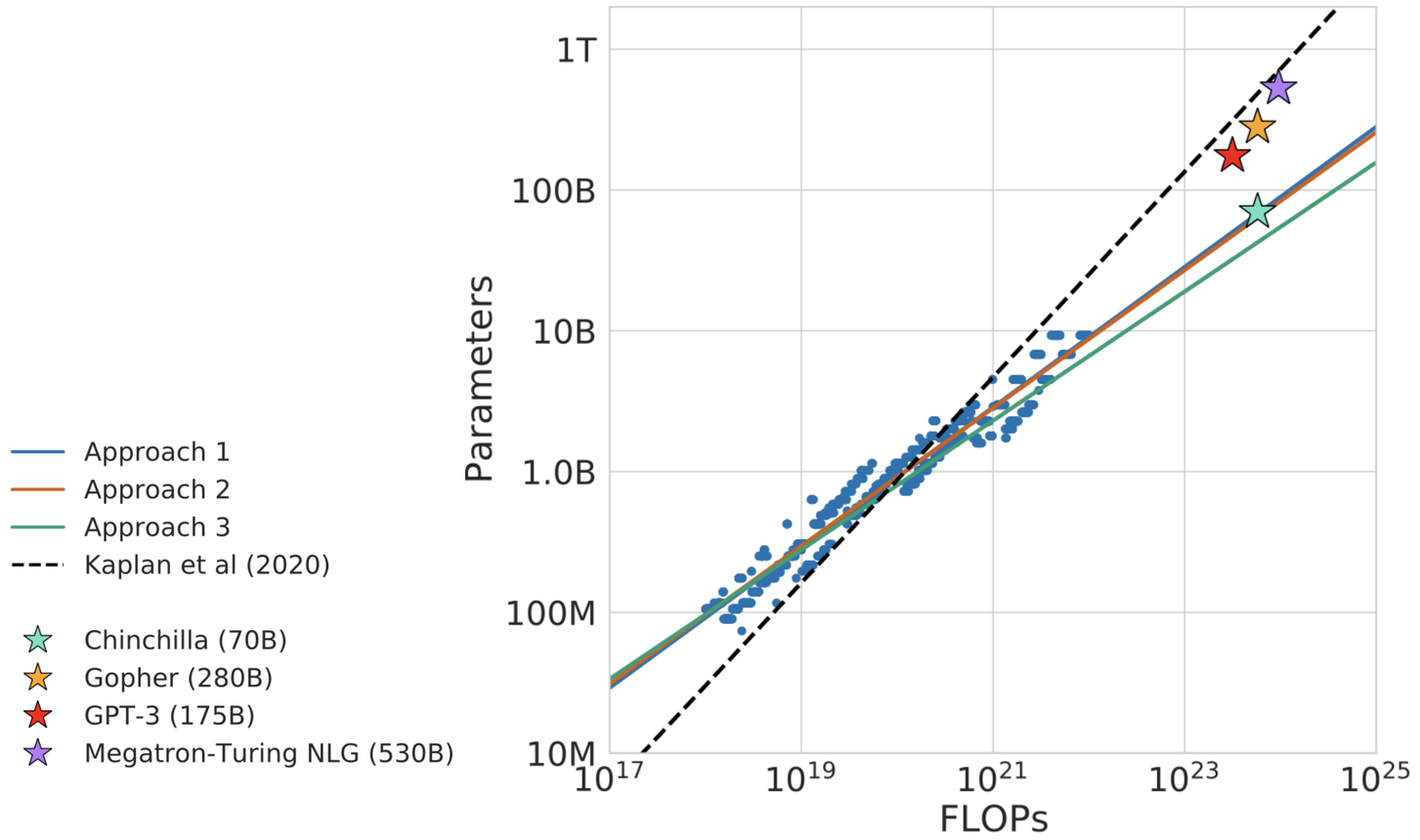
Chinchilla conclusions

Conclusion: ~~$D \propto N^{0.27}$~~ $D \propto N$!!

Translation: If you have 10x more compute, increase model size ~~6.2x~~ **3.2x** and data size ~~1.6x~~ **3.2x**

More concretely: You need (# of tokens) = **20** x (# of params)

- **Gopher and GPT-3 did not really need to be that big —they are undertrained**
- **Reduce parameter count, and instead train longer!**



Outcome: Chinchilla (70B)!

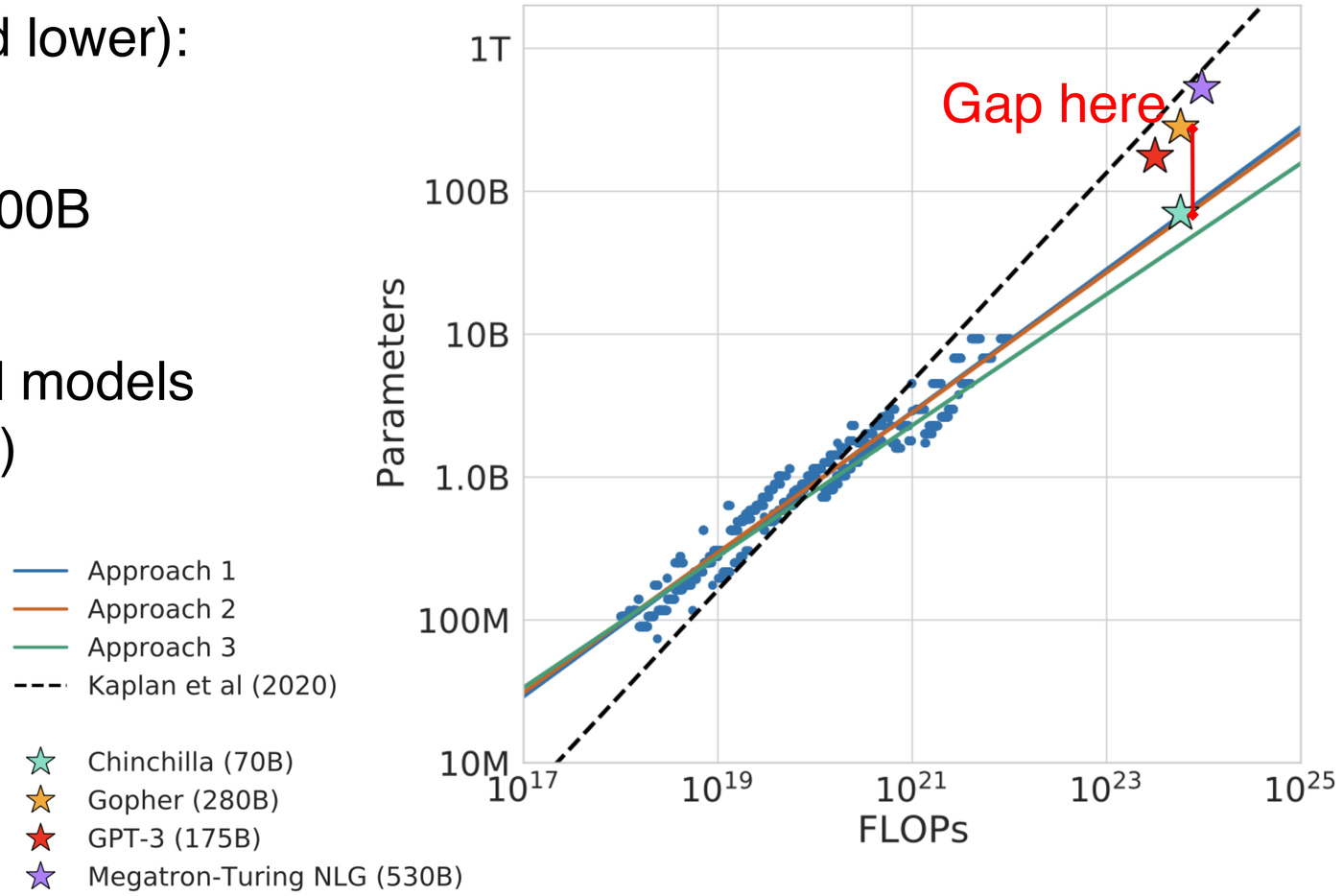
- Chinchilla is same compute as Gopher, but 4x more data and ¼ params
- 4-7% improvements most places overall (outperforming Gopher and GPT-3)!

		280B	175B	70B
	Method	*Gopher*	GPT-3	*Chinchilla*
Natural Questions (dev)	0-shot	10.1%	14.6%	16.6%
	5-shot	24.5%	-	31.5%
	64-shot	28.2%	29.9%	35.5%
TriviaQA (unfiltered, test)	0-shot	52.8%	64.3 %	67.0%
	5-shot	63.6%	-	73.2%
	64-shot	61.3%	71.2%	72.3%
TriviaQA (filtered, dev)	0-shot	43.5%	-	55.4%
	5-shot	57.0%	-	64.1%
	64-shot	57.2%	-	64.6%

	Gopher	GPT-3	MT-NLG 530B	*Chinchilla*
HellaSWAG	79.2%	78.9%	80.2%	**80.8%**
PIQA	81.8%	81.0%	**82.0%**	81.8%
Winogrande	70.1%	70.2%	73.0%	**74.9%**
SIQA	50.6%	-	-	**51.3%**
BoolQ	79.3%	60.5%	78.2%	**83.7%**

Why different conclusions from Kaplan et al?

- OpenAI mostly looked at (100M and lower): turned out to be “too small”
  - Chinchilla: 16B parameters and 500B tokens
- A fixed learning rate schedule for all models (i.e., used intermediate checkpoints)
  - Looks minor, isn’t it?
  - Telling us about how non-trivial it is to set up the right scaling laws



What happened after Chinchilla?

- Many smaller models that perform better
- Number of tokens is going up a lot!

Year	2020	2021	2022	2023	2024	2025	2025
Model	GPT3	Gopher	Chinchilla	LLama1	Llama3	Qwen3	Olmo2
Params	175B	280B	70B	65B	70B	235B	32B
Tokens	240B	300B	1.4T	1.4T	15T	36T	6T

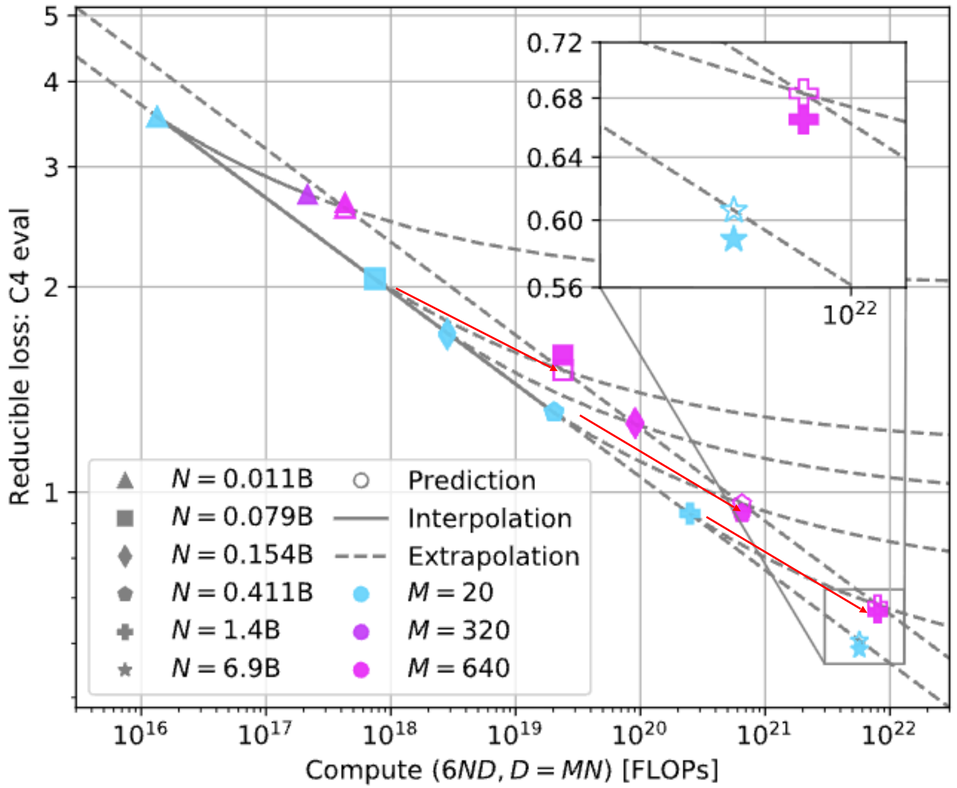
- Even more extreme setting of over-training (train on even more data than “train-time optimal”)

WHY??

1) Inference efficiency: Small models are easier and cheaper to run at inference
2) “Practical” training cost is actually lower with “smaller model, more data” than “big model, less data”

Different kinds of scaling laws (1/5)

Scaling laws in the “over-training” regime

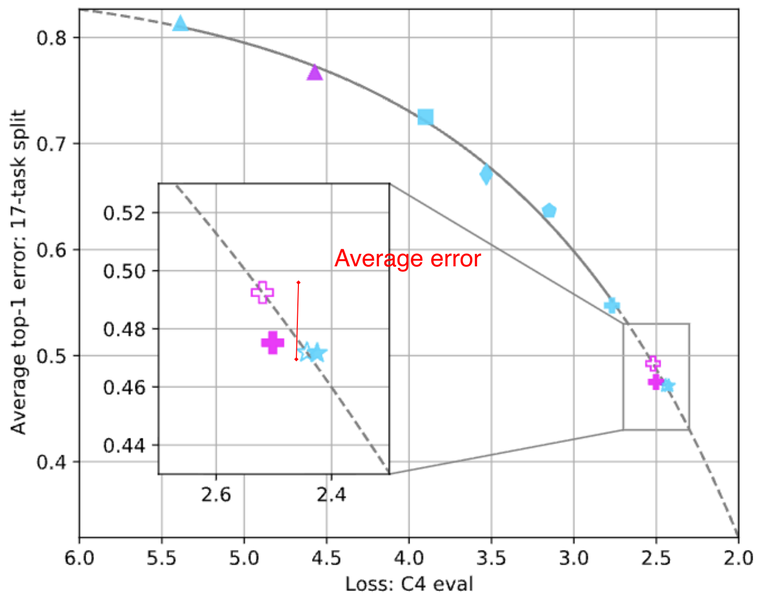


Overtrained models follow different scaling law, shifts over a bit

Gadre et al. 2024. "Language models scale reliably with over-training and on downstream tasks"

Different kinds of scaling laws (2/5)

Scaling laws for downstream task performance



Average benchmark performance is sort of predictable (with the right data and metrics), although individual isn't

Table 2: **Downstream relative prediction error at 6.9B parameters and 138B tokens.** While predicting accuracy on individual zero-shot downstream evaluations can be challenging ("Individual"), predicting *averages* across downstream datasets is accurate ("Avg.").

	Individual top-1 error				Avg. top-1 error
Train set	ARC-E [23]	LAMBADA [77]	OpenBook QA [68]	HellaSwag [126]	17-task split
C4 [27, 88]	28.96%	15.01%	16.80%	79.58%	0.14%
RedPajama [112]	5.21%	14.39%	8.44%	25.73%	0.05%
RefinedWeb [82]	26.06%	16.55%	1.92%	81.96%	2.94%

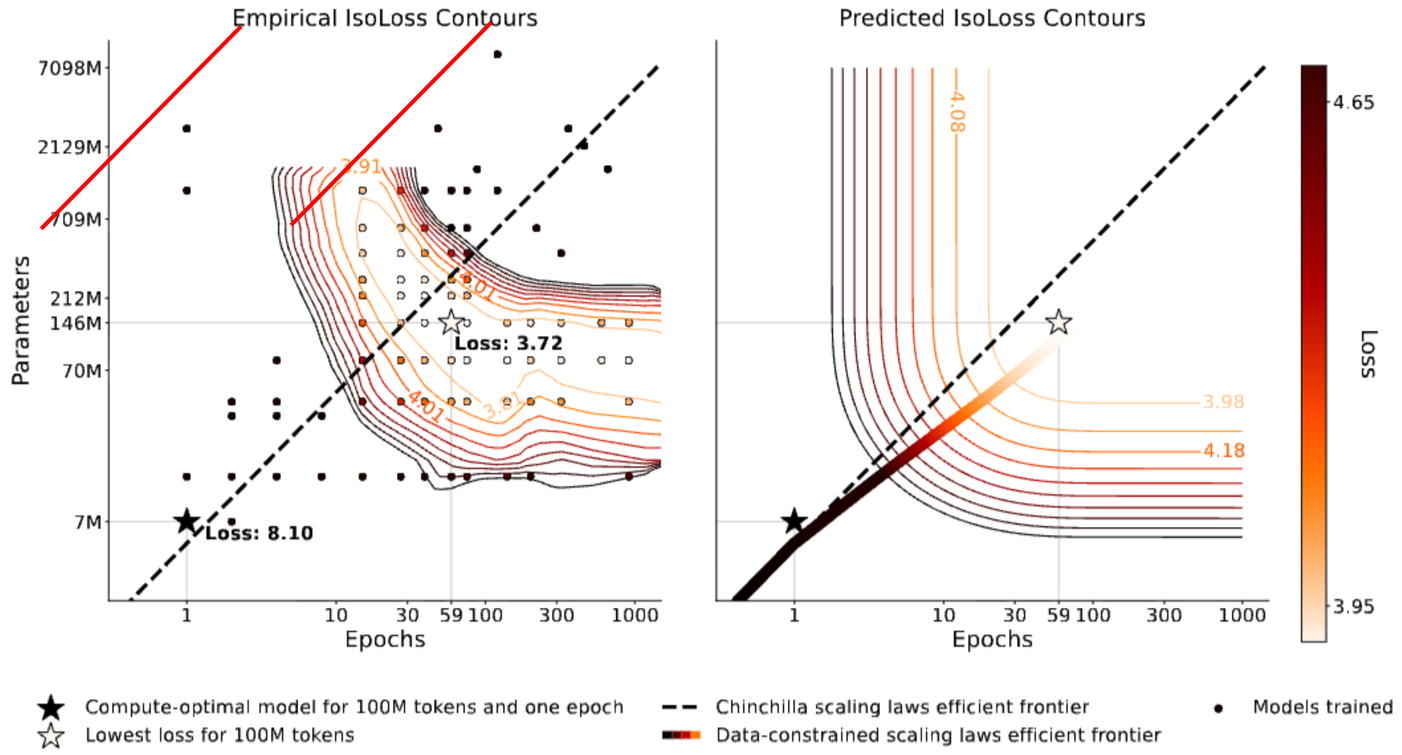
Error for specific tasks can get pretty high, but average is low

26

Gadre et al. 2024. "Language models scale reliably with over-training and on downstream tasks"

Different kinds of scaling laws (3/5)

Scaling laws for the data-constrained settings



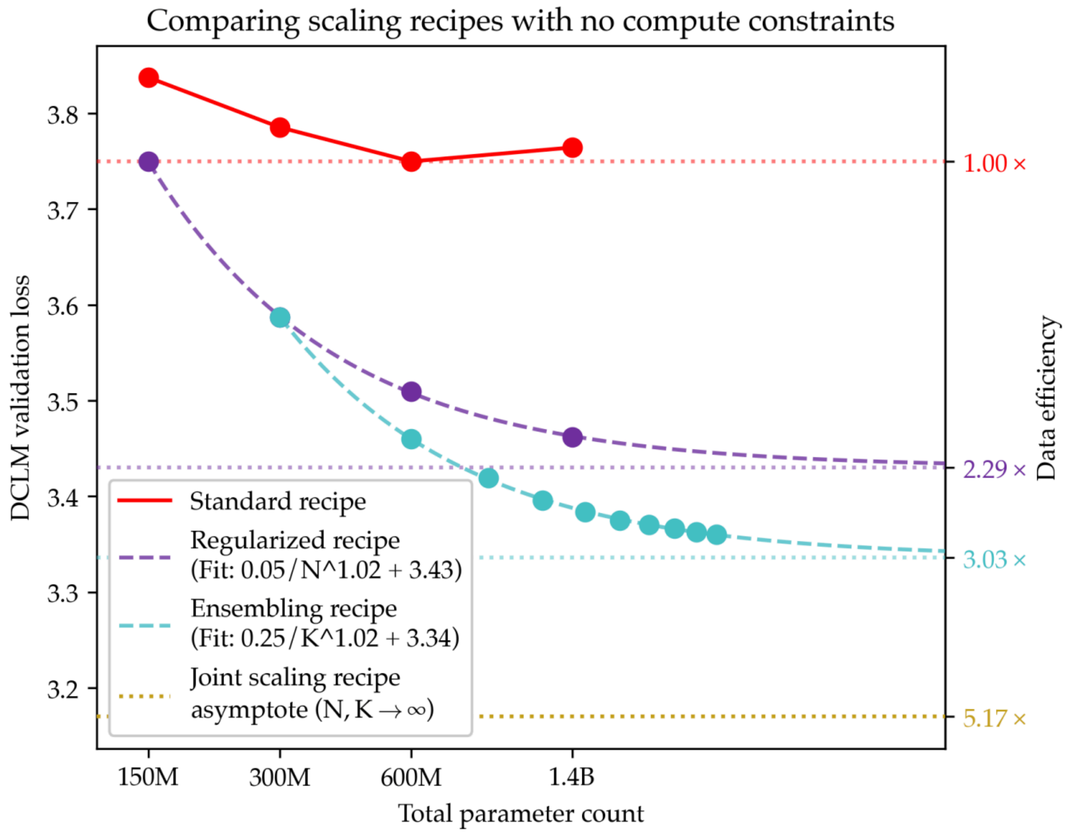
- Chinchilla: assumes unlimited/ unique data
- Motivation: Data might run out
- Setting: Fix 100M unique tokens. If you have more compute, how do you allocate it?

Low repetition (<4 epochs) approximates chinchilla

Takeaway: you can squeeze more juice from data with more compute

Different kinds of scaling laws (4/5)

Scaling laws for the “infinite compute” scenario



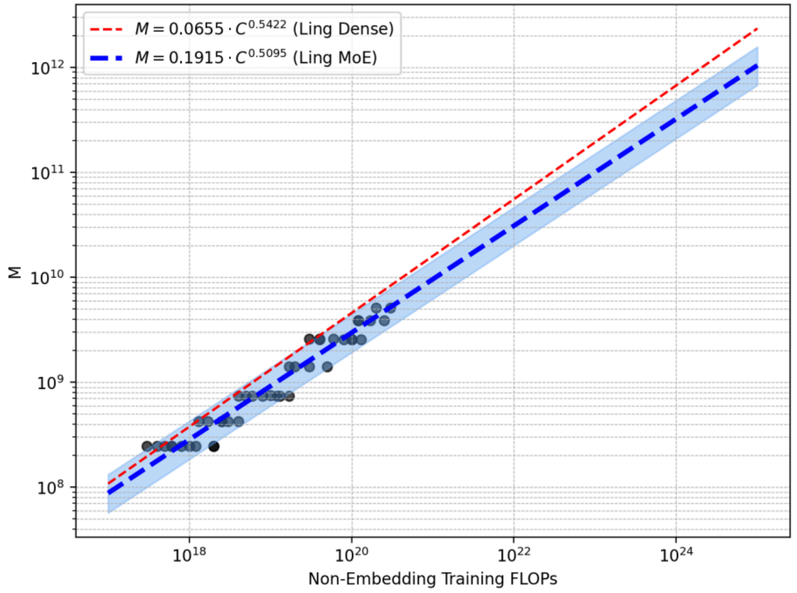
- pre-training under fixed data and no compute constraints
- Existing data-constrained approaches of increasing epoch count and parameter count eventually overfit
- It’s possible to significantly improve by properly tuning regularization, e.g., increasing weight decay to be 30× larger than standard practice.
- Ensembling independently trained models achieves a significantly lower loss asymptote than the regularized recipe.

Kim et al. 2025. "Pre-training under infinite compute"

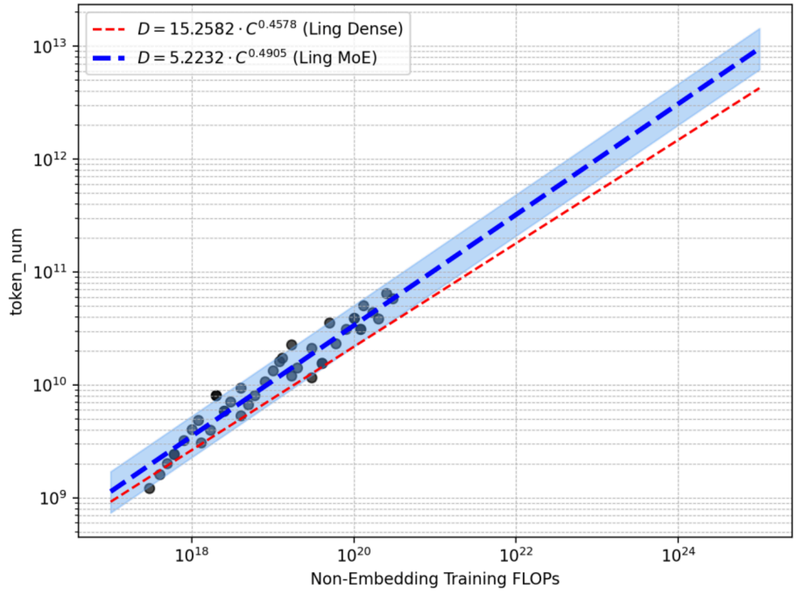
Different kinds of scaling laws (5/5)

Scaling laws for the Mixture-of-Experts architecture [a later lecture]

	Optimal Model Scale (M^{opt})	Optimal Data Size (D^{opt})
Dense	$M^{\text{opt}} = 0.0655 \cdot C^{0.5422}$	$D^{\text{opt}} = 15.2582 \cdot C^{0.4578}$
MoE	$M^{\text{opt}} = 0.1915 \cdot C^{0.5095}$	$D^{\text{opt}} = 5.2232 \cdot C^{0.4905}$



(a) Optimal Model Scale (M^{opt}) Scaling



(b) Optimal Data Size (D^{opt}) Scaling

Tian et al. 2025. Towards Greater Leverage: Scaling Laws for Efficient Mixture-of-Experts Language Models

Does scaling law make sense? (1/2)

Common critics

- May specific to specific setups, e.g., hyper parameter choices, architecture choices
- Does extrapolation actually work? Very large models may not fit!
- Overlooks the relationship between pre-training loss and the validation accuracy for specific downstream tasks
  - Lourie et al. (2025): Most tasks scale unpredictably -- Only 18 of 46 tasks are predictable

Does scaling law make sense? (2/2)

Recommended perspectives

- Scaling laws are not laws! They are empirical trends.
- They exist for necessity — resource allocation is a “real” issue
    - Without scaling laws, your only other option is “YOLO” run
        - But “failed” runs a
    - History tells us — GPT-3
- “Downstream task performan
  task performance predicable”
    - It’s possible to have ina
      this out properly), but fr

Pre-training Data

From pre-training to data

- So far: Data scaling is very important, and even GPT-3 didn't properly scale the data
- Since then: frontier lab's continued efforts in scalable training data curation
- Also the biggest secret: even when models are open-weight, most models do not open source nor disclose much information about the training data

> "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

- Today: we cover history, and public's best efforts to replicate frontier pre-training data

History of pre-training data

Pre-training data before GPT-2

- **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**
- 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

- 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”.
  - 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.

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 classifier to distinguish high-quality data from low-quality Common Crawl**
    - A logistic regression classifier
    - Positive examples: **WebText, Wikipedia, and several book corpora**
    - Negative examples: **Unfiltered Common Crawl**
  - Kept a doc iff np.random.pareto(α) > 1 – document_score w/ α = 9
    - Take mostly docs with high scores, but still include some low-score docs
- (Fuzzy) Deduplication: Removed docs with high overlap with other docs
- **Add known high-quality reference corpora**
  - An expanded version of the WebText dataset
  - Two books corpora (Books1 and Books2)
  - English Wikipedia

GPT-3 [Brown et al. 2020]

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

- After all the filtering, roughly equivalent to **400B tokens**
- "[D]atasets we view as higher-quality are sampled more frequently, such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times."

What Other Public Training Datasets Exist Today?

Dataset	Example LMs	Tokens	Sources
C4 (Oct 2019)	T5, FLAN-T5	175B	Common Crawl
Pile (Dec 2020)	GPT-J, GPT-NeoX, Pythia	387B	Common Crawl, arXiv, PubMed, Books3, Gutenberg, Wikipedia, etc…
The Stack v1 (Nov 2022)	StarCoder	200B	Software Heritage
RedPajama v1 (Apr 2023)	INCITE	1.2T	Common Crawl, C4, Github, arXiv, Gutenberg, Books3, Wikipedia, Internet Archive (Stack Exchange)
RefinedWeb (Jun 2023)	Falcon	580B*	Common Crawl
Dolma (Aug 2023)	OLMo	3.1T	Common Crawl, C4, Semantic Scholar, Pushshift Reddit, Gutenberg, the Stack, Wikipedia, Wikibooks

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

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

Dataset	Example LMs	Tokens	Sources
OpenWebMath (Oct 2023)	Llema	15B	Common Crawl
RedPajama v2 (Oct 2023)	-	30T	Common Crawl
Amber (Dec 2023)	Amber	1.3T	C4, RefinedWeb, the Stack, RedPajama v1
Dolma 1.7 (Apr 2024)	OLMo 0424	2.3T	Dolma, RefinedWeb, RP's StackExchange, Flan, OpenWebMath, …
FineWeb (May 2024)	-	15T	Common Crawl
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**"

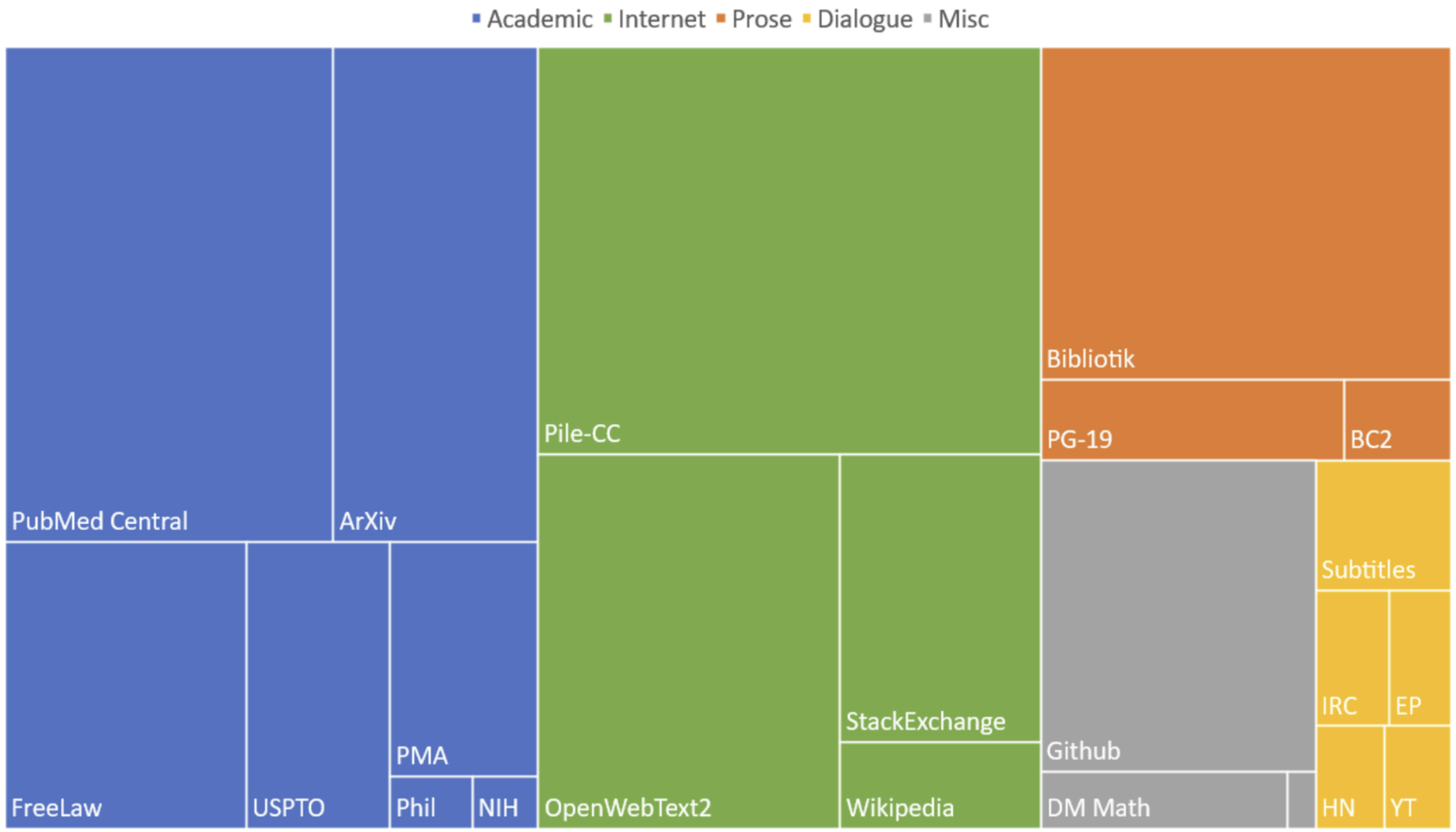
What Other Public Training Datasets Exist Today?

Dataset	Example LMs	Tokens	Sources
C4 (Oct 2019)	T5, FLAN-T5	175B	Common Crawl
Pile (Dec 2020)	GPT-J, GPT-NeoX, Pythia	387B	Common Crawl, arXiv, PubMed, Books3, Gutenberg, Wikipedia, etc…
The Stack v1 (Nov 2022)	StarCoder	200B	Software Heritage
RedPajama v1 (Apr 2023)	INCITE	1.2T	Common Crawl, C4, Github, arXiv, Gutenberg, Books3, Wikipedia, Internet Archive (Stack Exchange)
RefinedWeb (Jun 2023)	Falcon	580B*	Common Crawl
Dolma (Aug 2023)	OLMo	3.1T	Common Crawl, C4, Semantic Scholar, Pushshift Reddit, Gutenberg, the Stack, Wikipedia, Wikibooks

Dataset	Example LMs	Tokens	Sources
OpenWebMath (Oct 2023)	Llema	15B	Common Crawl
RedPajama v2 (Oct 2023)	-	30T	Common Crawl
Amber (Dec 2023)	Amber	1.3T	C4, RefinedWeb, the Stack, RedPajama v1
Dolma 1.7 (Apr 2024)	OLMo 0424	2.3T	Dolma, RefinedWeb, RP's StackExchange, Flan, OpenWebMath, …
FineWeb (May 2024)	-	15T	Common Crawl
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]

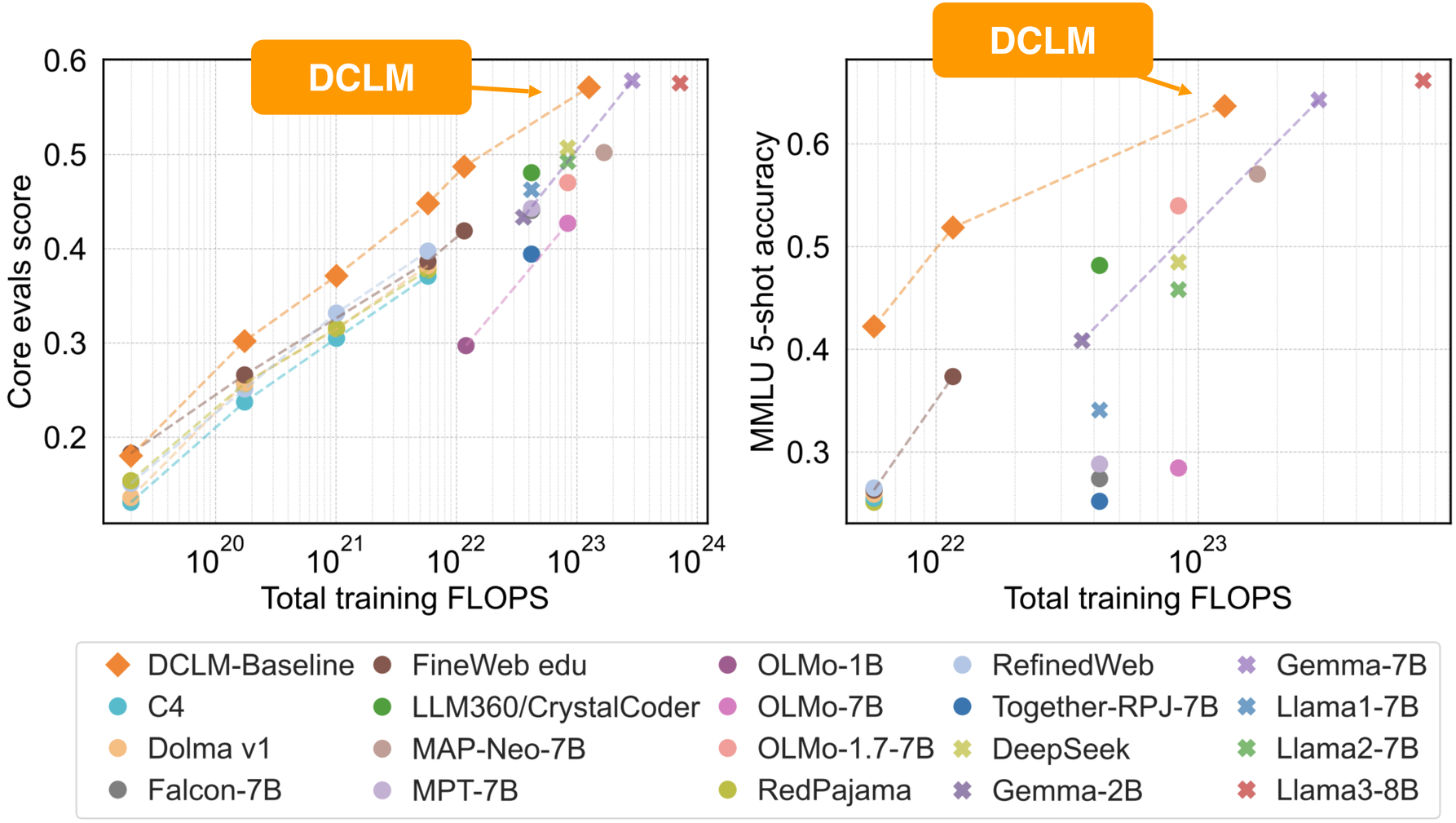
Source	Doc Type	UTF-8 bytes *(GB)*	Documents *(millions)*	Unicode words *(billions)*	Llama tokens *(billions)*
Common Crawl	web pages	9,812	3,734	1,928	2,479
GitHub	code	1,043	210	260	411
Reddit	social media	339	377	72	89
Semantic Scholar	papers	268	38.8	50	70
Project Gutenberg	books	20.4	0.056	4.0	6.0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

Example 4: DCLM and FineWeb (2024)

- Focusing on **Common Crawl only — extensive ablations** to maximize performance from it
- Conducted at a **(fairly) academic scale**!
  - DCLM: From UW + Apple
  - FineWeb: Huggingface
- 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: Produced a 7B model that is SoTA among those with open-source data

Li et al. 2024. "DataComp-LM: In search of the next generation of training sets for language models”
Penedo et al. 2024. "FineWeb: decanting the web for the finest text data at scale"

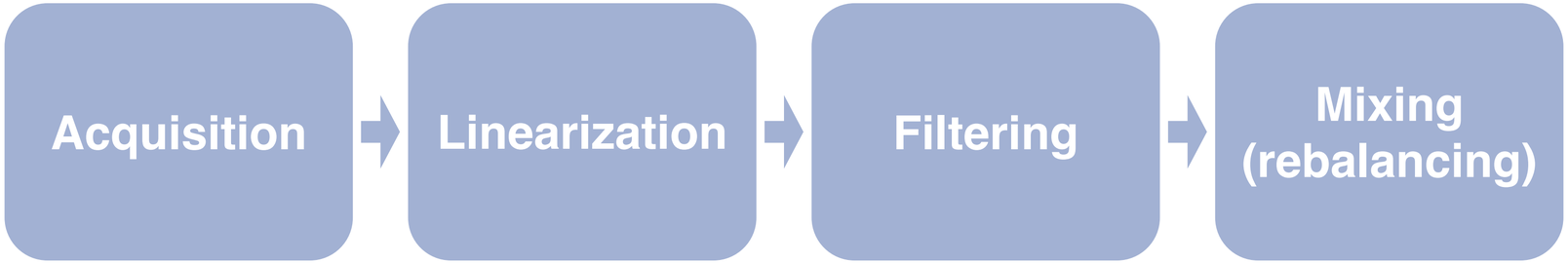
Example 4: DCLM and FineWeb (2024)



Li et al. 2024. "DataComp-LM: In search of the next generation of training sets for language models”
Penedo et al. 2024. "FineWeb: decanting the web for the finest text data at scale"

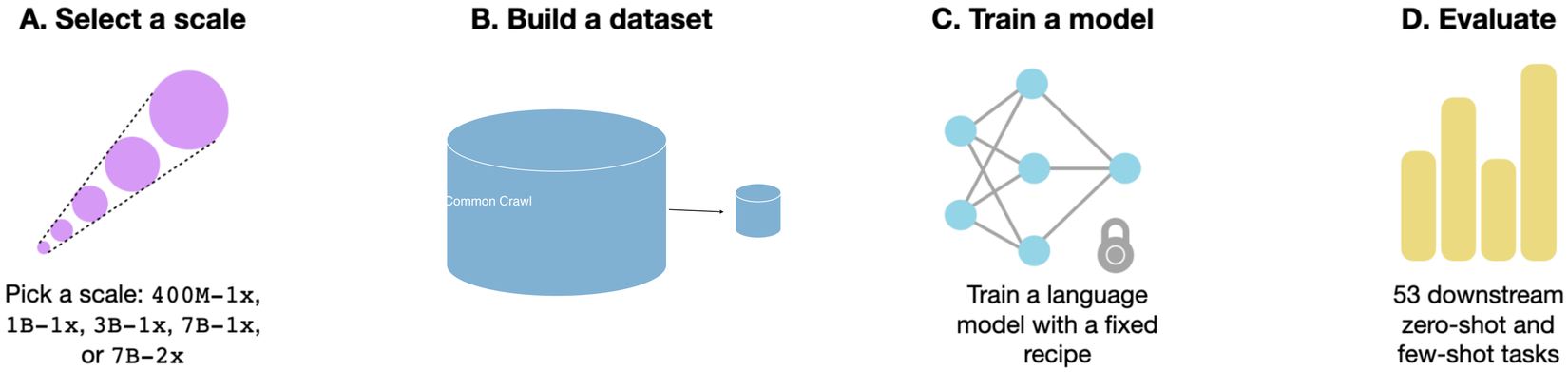
Pre-training data curation

Data curation: Overview



Preliminary: How to evaluate data?

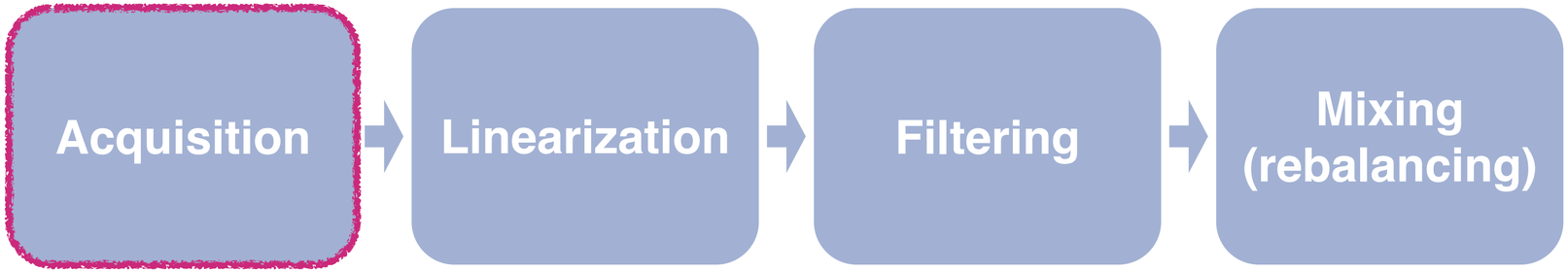
- Fix model architecture, training recipe, optimizer, scale, evaluation, etc — Only vary the training data!



- Challenges: Pre-training is expensive!
  - Invest in smallest experimental design that generalizes
    - “Scaling laws”
    - “Annealing”-style ablations (see Section 3.1.3 of the Llama 3 technical report)
  - Choose evaluation sets based on “how to get early, stable signals” rather than “difficulty”

Figure from the DCLM paper

Data curation: 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 quantitive 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 LaTeX 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

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

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

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

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

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: Debate

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-1T

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.

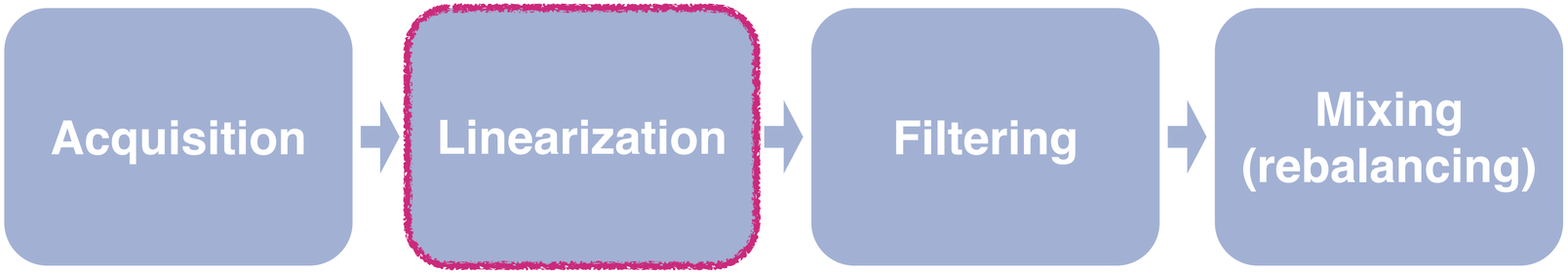
wired.com/story/new-documents-unredacted-meta-copyright-ai-lawsuit/

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

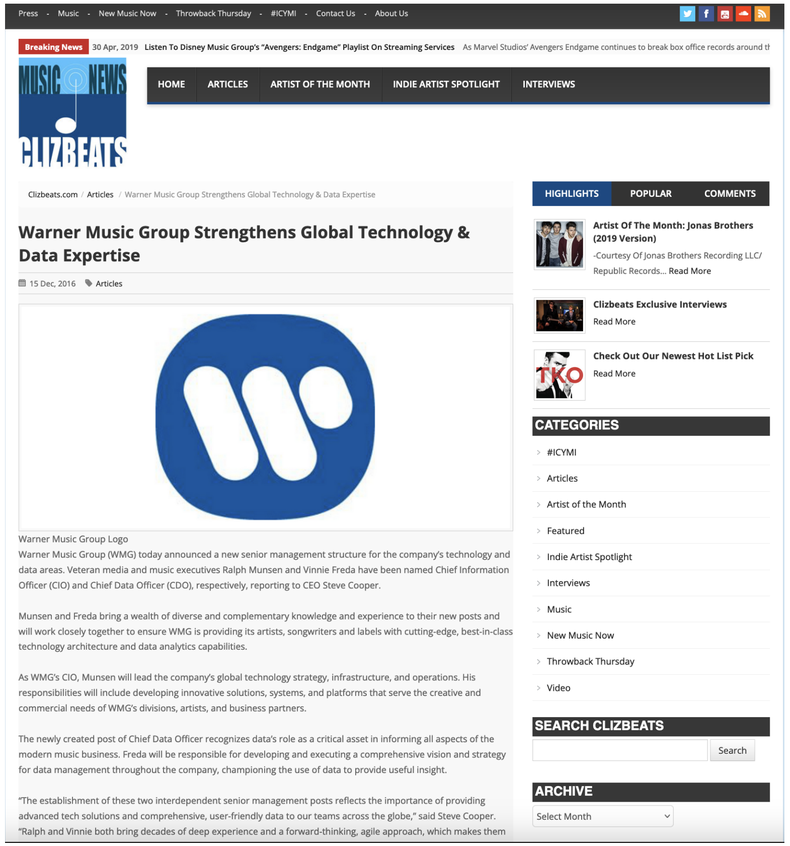
wired.com/story/stack-overflow-will-charge-ai-giants-for-training-data/

Data curation: Overview



2. Linearization

- HTML → Plain Text
- PDF → Plain Text
- …




```
  "text": "New Music Now\nThrowback Thursday\n#ICYMI\n30 Apr, 20
19LIFETIME TO DEBUT FOLLOW UP SPECIAL SURVIVING R. KELLY: THE IM
PACT WITH HOST SOLEDAD O'BRIEN ON MAY 4 Following the debut of t
he record-breaking premiere of Lifetime's Surviving R. Kelly thi
s past January, the network will debut… Read More\n30 Apr, 2019L
isten To Disney Music Group's "Avengers: Endgame" Playlist On St
reaming Services As Marvel Studios' Avengers Endgame continues t
o break box office records around the world, we are happy… Read
More\n24 Apr, 2019SONY MUSIC ENTERTAINMENT NAMES SYLVIA RHONE CH
AIRMAN 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, 2019ANDERSON .PAAK ANNOUNCE
S 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 To
p 200 for his latest album Ventura… Read More\nArtist of the Mon
th\nIndie Artist Spotlight\nClizbeats.com/\nArticles /\nWarner M
usic Group Strengthens Global Technology & Data Expertise\n15 De
c, 2016 matt Articles\nWarner Music Group Logo\nWarner Music Gro
up (WMG) today announced a new senior management structure for t
he company's technology and data areas. Veteran media and music
executives Ralph Munsen and Vinnie Freda have been named Chief I
nformation Officer (CIO) and Chief Data Officer (CDO), respectiv
ely, reporting to CEO Steve Cooper.\nMunsen and Freda bring a we
alth of diverse and complementary knowledge and experience to th
eir new posts and will work closely together to ensure WMG is pr
oviding its artists, songwriters and labels with cutting-edge, b
est-in-class technology architecture and data analytics capabili
ties.\nAs WMG's CIO, Munsen will lead the company's global techn
ology strategy, infrastructure, and operations. His responsibili
ties will include developing innovative solutions, systems, and
platforms that serve the creative and commercial needs of WMG's
divisions, artists, and business partners.\nThe newly created po
st of Chief Data Officer recognizes data's role as a critical as
set in informing all aspects of the modern music business. Freda
 will be responsible for developing and executing a comprehensiv
e vision and strategy for data management throughout the company
, championing the use of data to provide useful insight.\n"The e
stablishment of these two interdependent senior management posts
 reflects the importance of providing advanced tech solutions an
```

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

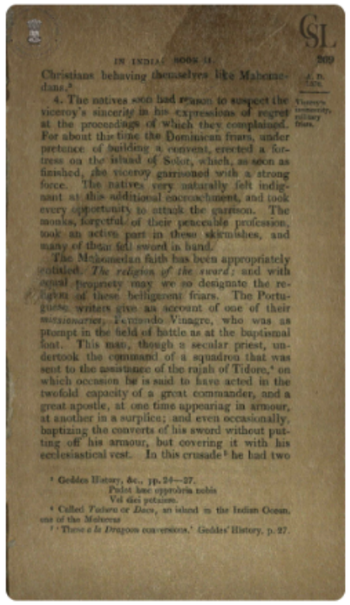
BAKERSFIELD SWAP-O-RAMA, 4501 Wible Road, Bakersfield, Calif. 93313; 805-831-9342. Saturday and Sunday.

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, Saturday and Sunday.

Continue reading the main story

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 viceroy'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 viceroy garrisoned with a strong force. The natives very naturally felt indignant 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 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

High quality

11 11 IN INDIA: BOOK

Christians behaving themselves like Mahomedans.3 4. The natives soon had reason to suspect the Viceroy viceroy's sincerity in his expressions of regret insmeerity 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 viceroy garrisoned with a strong

force. The natives very naturally felt indignant 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 ntitled, The religion of the sword; and with qual propriety may we so designate the reigion of these belligerent friars. The Portuguess 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,4 on which occasion he is said to have acted in the twofold capacity of a great commander, and a great

High quality

Pudet hæc opprobria nobis Vel dici potuisse.

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

Christians behaving themselves like Ma borne- a. t>.

dans.3 ."5/0-

4. The natives soon had reason to suspect the viceroy,

viceroy'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 for tress on the island of Sol or, which, as soon as

finished, the viceroy garrisoned 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 i'lfinomedan faith has been appropriately

Low quality

missionaries, Fernando Vinagre, who was as

prompt in the field of battle as at the baptismal

font. This man, though a secular priest, un dertook 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 Geddea History, &c., pp. 24—27.

Pudet hsec opprobria nobis

Vel dici potuisse.

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

one of the Moluccas

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

IN INDIA * BOOK TI. S69

IN INDIA * BOOK TI.

S

I

;

S69

Christians behaving themselves like Ma borne- a. t>.

dans.3

."5/04. The natives soon had reason to suspect the viceroy,

viceroy'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 for

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

finished, the viceroy garrisoned with a strong

Low quality

many of tbg.tr fell sword in hand.

The i'lfinomedan faith has been appropriately

entitled., The religion o f the sword,; and with

equal propriety may we so designate the re.■. i'gv.m of these belligerent friars. The Portugu

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, un

dertook the command of a squadron that was

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 Geddea History, &c., pp. 24—27.

Pudet hsec opprobria nobis

Vel dici potuisse.

Christians behaving themselves like Ma borne-

a. t>.

dans.3

."5/0-

4. The natives soon had reason to suspect the viceroy,

viceroy'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 for

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

finished, the viceroy garrisoned with a strong

force.

The natives' very naturally felt indig-

S

nant at this additional encroachment, and took

every

The

monk

took

Low quality

many of tbg.tr fell sword in hand.

The i'lfinomedan faith has been appropriately

entitled., The religion of the sword,; and with

equal propriety may we so designate the re-

.■. i'gv.m of these belligerent friars.

The Portu-

gu

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, un

dertook 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,

Side story: Linearization of Common Crawl

CommonCrawl data is available in two main formats:

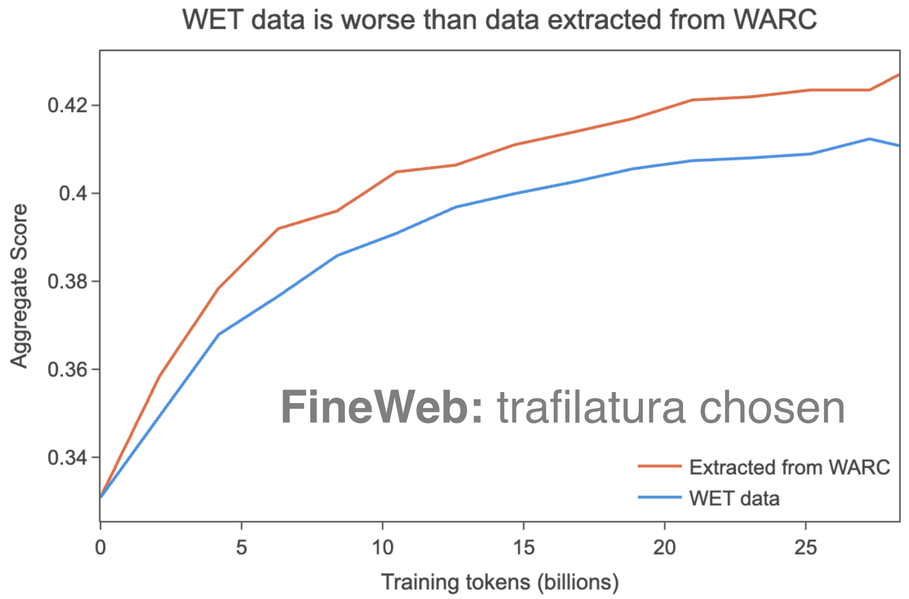
- **WET** (WARC Encapsulated Text) files provide **a plan text only version**.
  - Most prior work takes the WET files as their starting point.
- **WARC** (Web ARChive format) files contain the **raw data from the crawl** (HTML and metadata)

DCLM and FineWeb reported ablations on these choices for the first time, finding that **their own linearization from WARC better than using Common Crawl provided WET files**

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

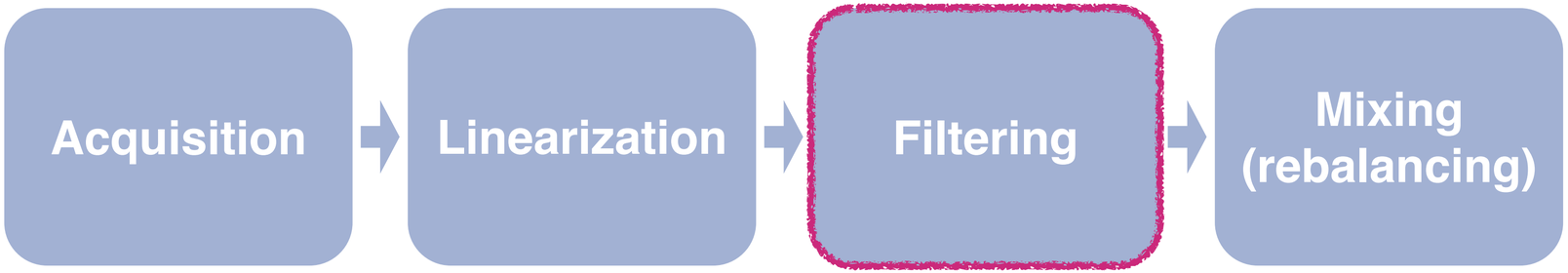
Text Extraction	CORE	EXTENDED
`resiliparse`	24.1	**13.4**
`trafilatura`	**24.5**	12.5
WET files	20.7	12.2

DCLM: Resiliparse chosen because of efficiency



Li et al. 2024. "DataComp-LM: In search of the next generation of training sets for language models"
Penedo et al. 2024. "FineWeb: decanting the web for the finest text data at scale"

Data curation: Overview



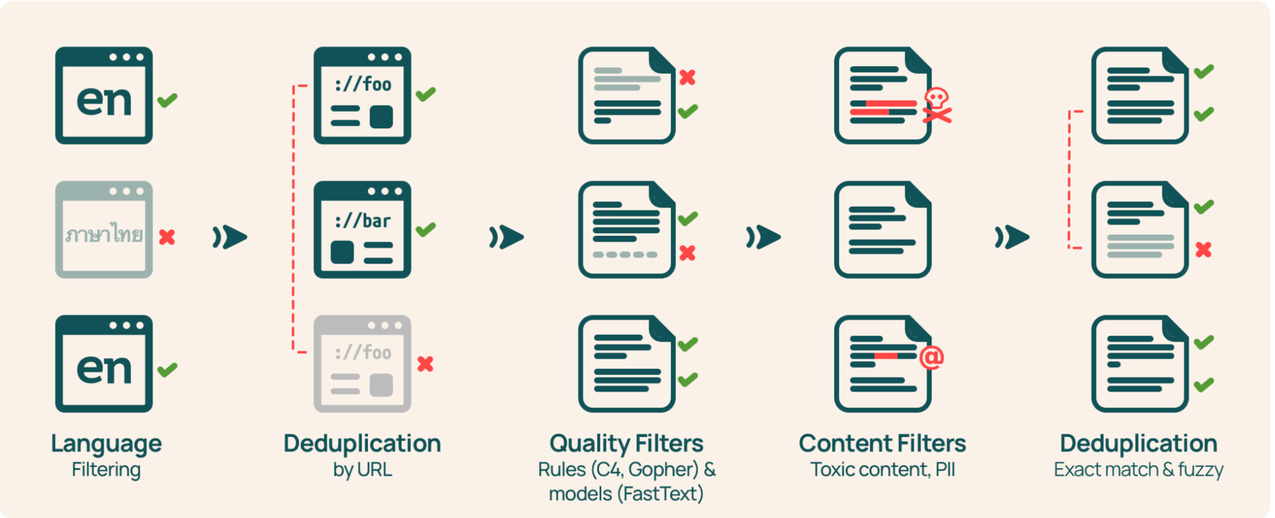
3. Filtering

Considerations

- Language
- Junk
- Content
- Quality
- Duplicates
- etc…

Methods

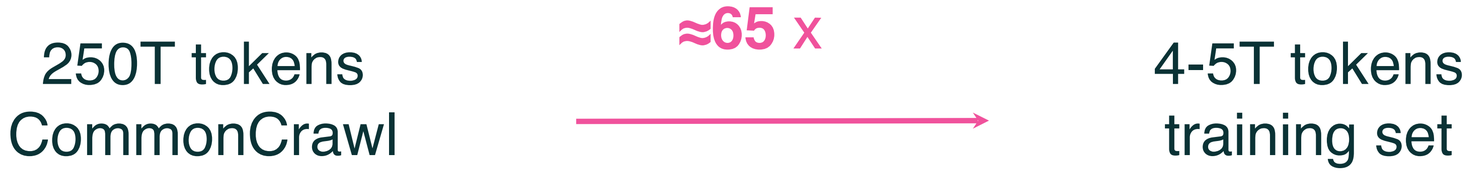
- Heuristic based
- Model based



➤ *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

3. Filtering: Why challenging?

Your data curation methods need to scale!



A bi-gram model

- 500 docs per second per CPU
- $8.5/hr for c7i instance, 192 cores
- Process **24B docs** in **3 days** and **$600 on CPUs**

BERT-Base (110M params)

- 1,600 docs per second per H100
- $2.50/hr
- Process **24B docs** in **3 days** and **$10,000 on 64 H100s**

3. Filtering: Closer look

- Deduplication
- Heuristic filtering
- Model-based filtering

3. Filtering: Closer look

- **Deduplication**
- Heuristic filtering
- Model-based filtering

Deduplication

- Common Crawls often contain many duplicate or near-duplicate data strings. Removing these duplicates can improve performance by reducing memorization and increasing data diversity.
- **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

3. Filtering: Closer look

- Deduplication
- **Heuristic filtering**
    - Didn’t deviated too much from T5’s filtering, at least conceptually (although tons of different variations!)
- Model-based filtering

3. Filtering: Closer look

- Deduplication
- Heuristic filtering
- **Model-based filtering**
    - Varies from cheap, bi-gram based filtering to LLM-based filtering

DCLM's model-based filtering

Train a **bi-gram model** to serve as quality filters

- 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

Heuristic filtering only

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**

FineWeb's model-based filtering

- Key idea: Use an LLM to label which docs are "educational" then train a classifier on that data!
  - Hints from prior papers from industry (Llama 3 from Meta and Phi3 from Microsoft) but without details or ablation results.
- Step 1: Prompt **Llama-3-70B-Instruct** to **annotate** the quality of the doc, based on **its educational contents** (on a scale of 0 to 5)
- Step 2: Fine-tune a **110M-parameter encoder-only** Transformer model on 450,000 Llama 3 annotations for 20 epochs
- Step 3: After training, rounded the scores to 0 to 5, and convert to a **binary classification task** by using **a fixed threshold** to determine if a doc is educational.
- Applying the classifier to the 15T tokens took 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:

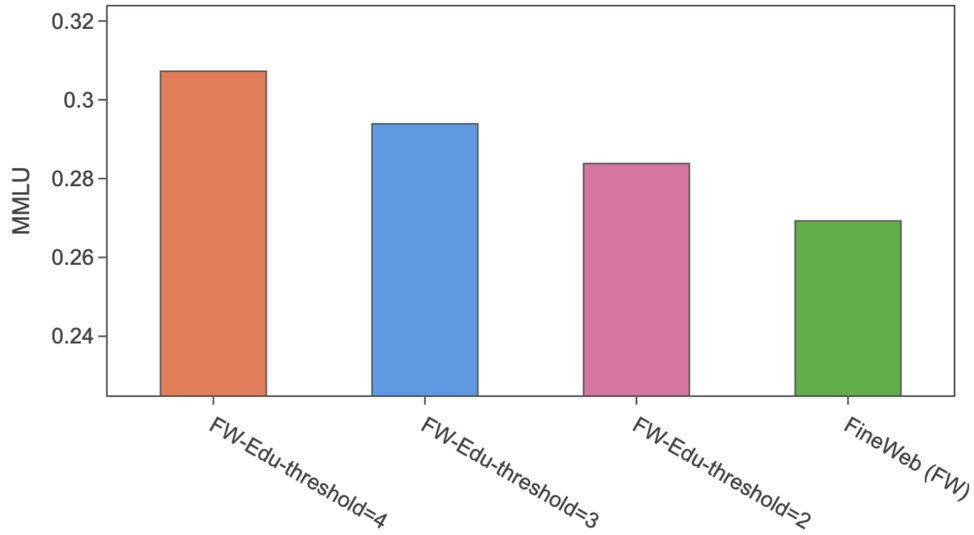
- Add 1 point if the extract provides some basic information relevant to educational topics, even if it includes some irrelevant or non-academic content like advertisements and promotional material.
- Add another point if the extract addresses certain elements pertinent to education but does not align 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 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.

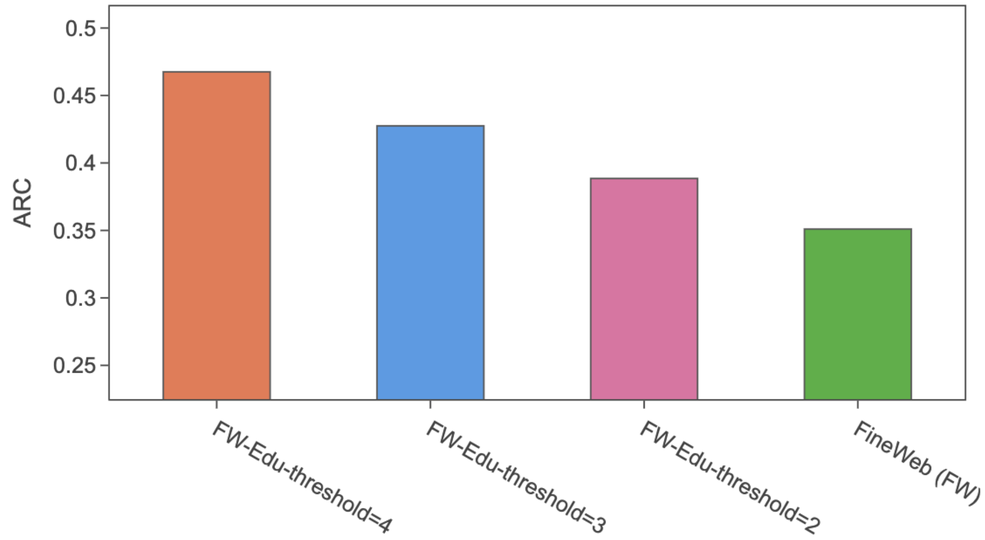
The extract: <extract>.

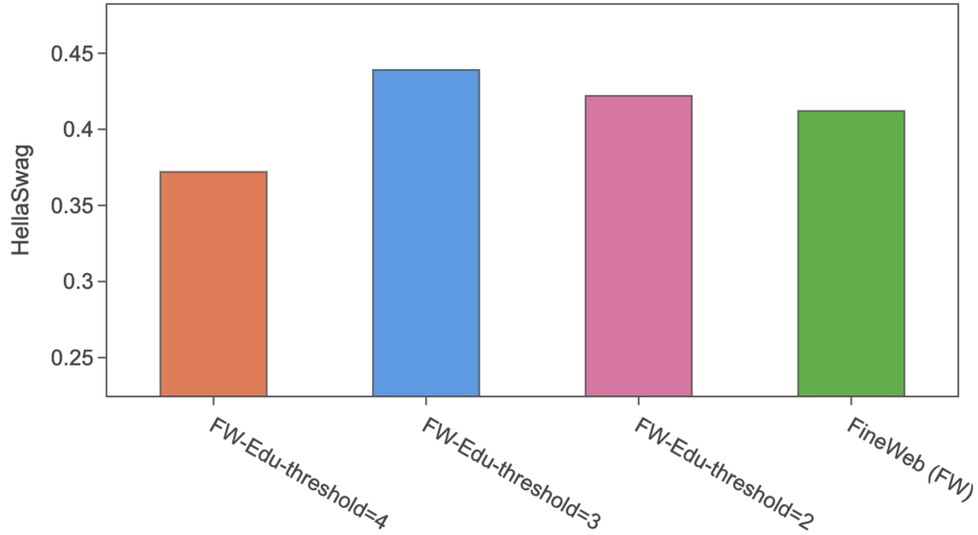
After examining the extract:

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

FineWeb's model-based filtering: Results





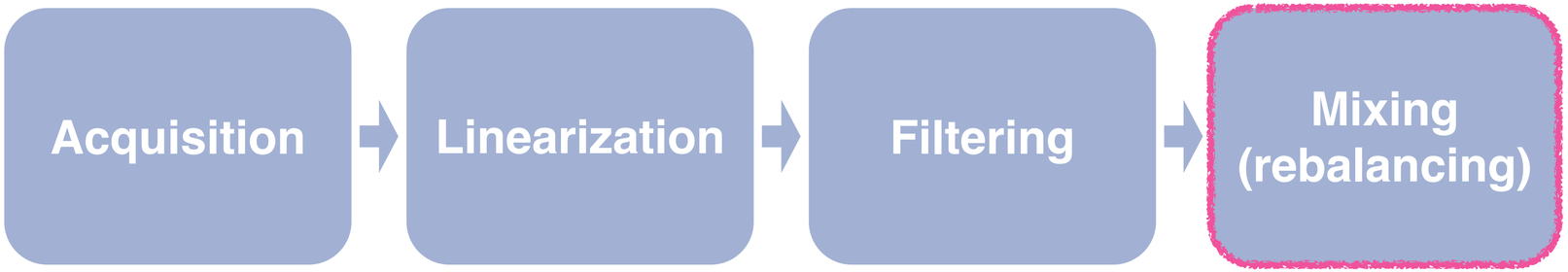


3. Filtering: Closer look

- Deduplication
- Heuristic filtering
- Model-based filtering

Quick quiz: In practice, most work does heuristic filtering, then deduplication, then model-based filtering. Why this order?

Data curation: 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]

Pre-training data summary

Summary: Pre-training data curation

- Very important: both **scale** and **quality**
  - Little known from industry labs due to competitive reasons

- Extensive, ongoing research from the open source community
  - Every step matters, even seemingly trivial ones, e.g., linearization and deduplication
  - These efforts are building on top of each other
    - {C4, RefinedWeb, etc} led to {DCLM, Fineweb}
    - {DCLM, Fineweb} led to {OLMo 2/3, etc}

- Beyond human-curated data: synthetic pre-training data, where LLMs generate new training data
  - Out of scope for this course, but check out Pratyush Maini (CMU / DatologyAI)'s guest lecture for CS 294-288 (Fall 2025) [link]

Open-ended questions

- If model-based filtering is ultimately applied, why heuristic filtering at all? Wouldn't model-based methods always be stronger?
- Model-based filtering: DCLM's "AskLLM" baseline is essentially FineWeb's approach, but DCLM found it worse than a bi-gram model. What's the difference between DCLM's AskLLM and FineWeb's approach?
- We learned filtering mainly based on Common Crawl, but there are other data sources covered 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?

Up next: Post-training!

Questions?

Acknowledgement

UC Berkeley Fall 2025 CS 294-288 Slides made by Jongho Park & Prasann Singhal
Presentation by Kyle Lo (Ai2), Pratyush Maini (CMU & Datology)