So you want to build an LLM…

Jonathan Frankle
Chief Scientist, MosaicML

www.github.com/mosaicml/llm-foundry
www.github.com/mosaicml/streaming
www.github.com/mosaicml/composer

Friendly Advice

Start small and work your way up.

Trust nothing you read in the literature.
Test everything for yourself.

Do not trust your intuition, received wisdom, or a rumor you heard about OpenAI. Test everything.

Do the math.

Let’s Talk Cost

How much does it cost to train?

https://github.com/mosaicml/llm-foundry/tree/main/scripts/train/benchmarking

https://lambdalabs.com/service/gpu-cloud

How much does it cost to train?

FLOPs = 6 * N * D

D = 20 * N (for Chinchilla)

7B parameters

A100 = 312TFLOP/s

How much does it cost to train?

FLOPs = 6 * N * D

D = 20 * N (for Chinchilla)

Actual FLOPs = FLOPs * MFU

What is MFU? MFU vs. HFU

How much does it cost to train?

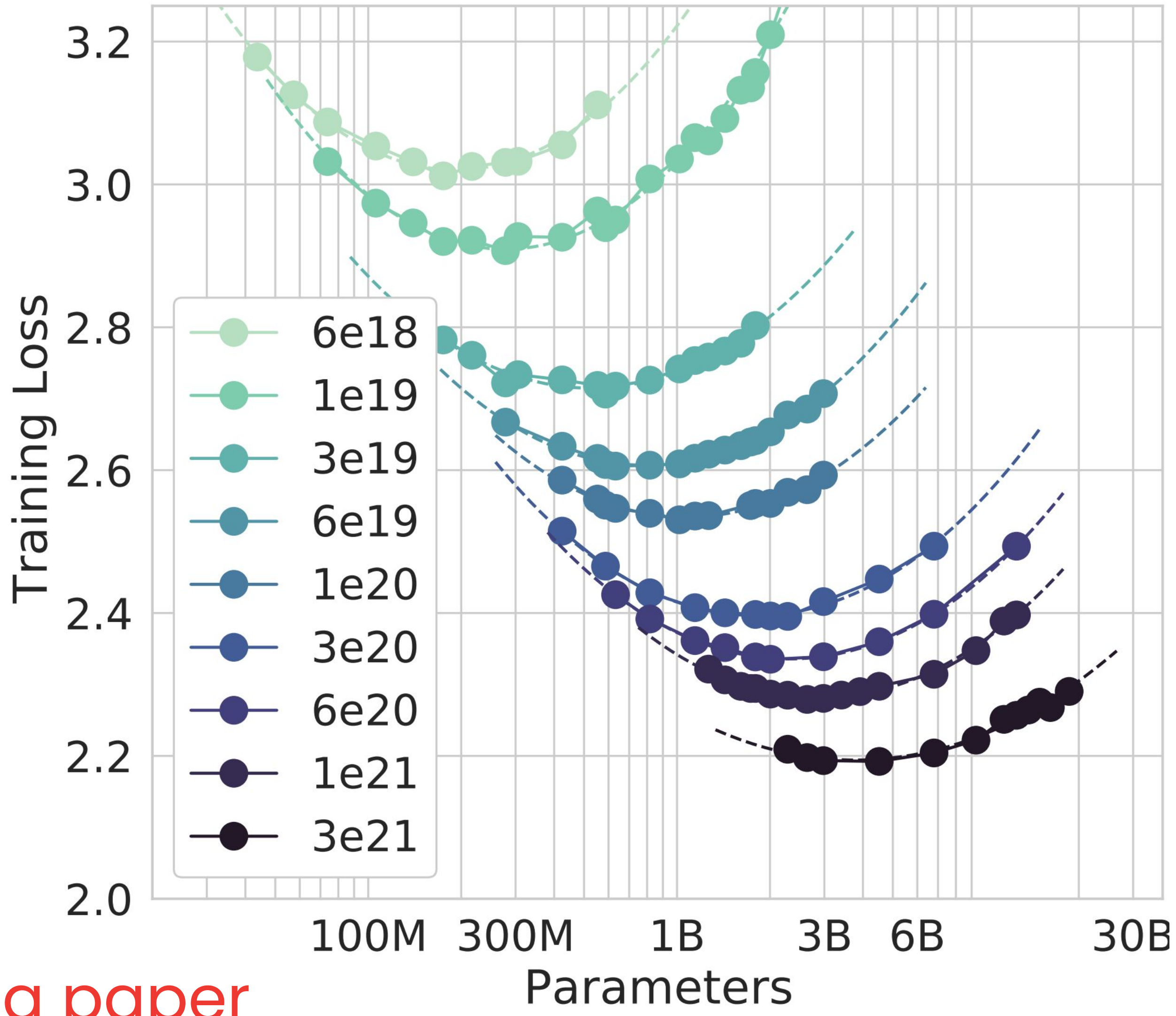
FLOPs = 6 * N * D

D = 20 * N (for Chinchilla)

Actual FLOPs = FLOPs * MFU

^ Ignores self-attention
https://arxiv.org/abs/2205.14135

Chinchilla or Llama?



From the Chinchilla paper

After-Training Data Cost

1 instruction-response pair: $30

1 pairwise comparison for RLHF: $8

1 multi-turn chat conversation: $130

After-Training Data Cost

1 instruction-response pair: $30

1 pairwise comparison for RLHF: $8

1 multi-turn chat conversation: $130

Quality Is All You Need. Third-party SFT data is available from many different sources, but we found that many of these have insufficient diversity and quality — in particular for aligning LLMs towards dialogue-style instructions. As a result, we focused first on collecting several thousand examples of high-quality SFT data, as illustrated in Table 5. By setting aside millions of examples from third-party datasets and using fewer but higher-quality examples from our own vendor-based annotation efforts, our results notably improved. These findings are similar in spirit to Zhou et al. (2023), which also finds that a limited set of clean instruction-tuning data can be sufficient to reach a high level of quality. We found that SFT annotations in the order of tens of thousands was enough to achieve a high-quality result. We stopped annotating SFT after collecting a total of 27,540 annotations. Note that we do not include any Meta user data.

After-Training Data Cost

1 instruction-response pair: $30

1 pairwise comparison for RLHF: $8

1 multi-turn chat conversation: $130

We also observed that different annotation platforms and vendors can result in markedly different downstream model performance, highlighting the importance of data checks even when using vendors to source annotations. To validate our data quality, we carefully examined a set of 180 examples, comparing the annotations provided by humans with the samples generated by the model through manual scrutiny. Surprisingly, we found that the outputs sampled from the resulting SFT model were often competitive with SFT data handwritten by human annotators, suggesting that we could reprioritize and devote more annotation effort to preference-based annotation for RLHF.

After-Training Data Cost

1 instruction-response pair: $30

1 pairwise comparison for RLHF: $8

1 multi-turn chat conversation: $130

Table 26 shows detailed statistics on Meta human preference data. In total, we collected 14 batches of human preference data (i.e., Meta Safety + Helpfulness) on a weekly basis, consisting of over 1 million binary model generation comparisons. In general, later batches contain more samples as we onboard more annotators over time and the annotators also become more familiar with the tasks and thus have better work efficiency. We also intentionally collect more multi-turn samples to increase the complexity of RLHF data and thus the average number of tokens per sample also increase accordingly over batches.

Let's Pick Data

Pick your proportions for 2T tokens

mosaicML MPT-30B Training Data

Data Source	Number of Tokens in Source (Billion)	Proportion	Effective Number of Tokens (Billion)	Epochs
mC4 3.1.0 - English (200+ words)	2417.99	??	335	??
c4 - English - SemDedup 80%	100.42	??	299	??
RedPajama - CommonCrawl	878.45	??	85	??
The Stack - Selected Languages	463.78	??	100	??
RedPajama - Wikipedia	4.87	??	40	??
The Stack - Markdown	107.07	??	45	??
Semantic Scholar ORC	48.95	??	33	??
RedPajama - Books	26.02	??	30	??
RedPajama - arXiv	28.1	??	19	??
RedPajama - StackExchange	20.54	??	14	??


STOP

What is your goal with this model?

General purpose chat, for now.

Pick your proportions for 2T tokens

mosaicML MPT-30B Training Data

Data Source	Number of Tokens in Source (Billion)	Proportion	Effective Number of Tokens (Billion)	Epochs
mC4 3.1.0 – English (200+ words)	2417.99	??	335	??
c4 – English – SemDedup 80%	100.42	??	299	??
RedPajama – CommonCrawl	878.45	??	85	??
The Stack – Selected Languages	463.78	??	100	??
RedPajama – Wikipedia	4.87	??	40	??
The Stack – Markdown	107.07	??	45	??
Semantic Scholar ORC	48.95	??	33	??
RedPajama – Books	26.02	??	30	??
RedPajama – arXiv	28.1	??	19	??
RedPajama – StackExchange	20.54	??	14	??

Pick your proportions for 2T tokens

mosaic[ML] MPT-30B Training Data

Data Source	Number of Tokens in Source (Billion)	Proportion	Effective Number of Tokens (Billion)	Epochs
mC4 3.1.0 - English (200+ words)	2417.99		335	0.14
c4 - English - SemDedup 80%	100.42	29.9%	299	2.98
RedPajama - CommonCrawl	878.45	8.5%	85	0.10
The Stack - Selected Languages	463.78	10.0%	100	0.22
RedPajama - Wikipedia	4.87	4.0%	40	8.21
The Stack - Markdown	107.07	4.5%	45	0.42
Semantic Scholar ORC	48.95	3.3%	33	0.67
RedPajama - Books	26.02	3.0%	30	1.15
RedPajama - arXiv	28.1	1.9%	19	0.68
RedPajama - StackExchange	20.54	1.4%	14	0.68

Pick your proportions for 2T tokens

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

Key Questions

Should you mix at all? Freshness vs. repetition.

Key Questions

Should you mix at all? Freshness vs. repetition.

Quality vs. quantity?

Key Questions

Should you mix at all? Freshness vs. repetition.

Quality vs. quantity?

Should you deduplicate?

Key Questions

Deduplicating Training Data Makes Language Models Better

Katherine Lee*† Daphne Ippolito*†‡ Andrew Nystrom† Chiyuan Zhang†

Douglas Eck† Chris Callison-Burch‡ Nicholas Carlini†

Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .
C4	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!

Abstract

We find that existing language modeling datasets contain many near-duplicate examples and long repetitive substrings. As a result, over 1% of the unprompted output of language models trained on these datasets is copied verbatim from the training data. We develop two tools that allow us to deduplicate training datasets—for example removing from C4 a single 61 word English sentence that is repeated over 60,000 times. Deduplication allows us to train models that emit memorized text ten times less frequently and require fewer training steps to achieve the same or better accuracy. We can also reduce train-test overlap, which affects over 4% of the validation set of standard datasets, thus allowing for more accurate evaluation. Code for deduplication is released at https://github.com/google-research/deduplicate-text-datasets.

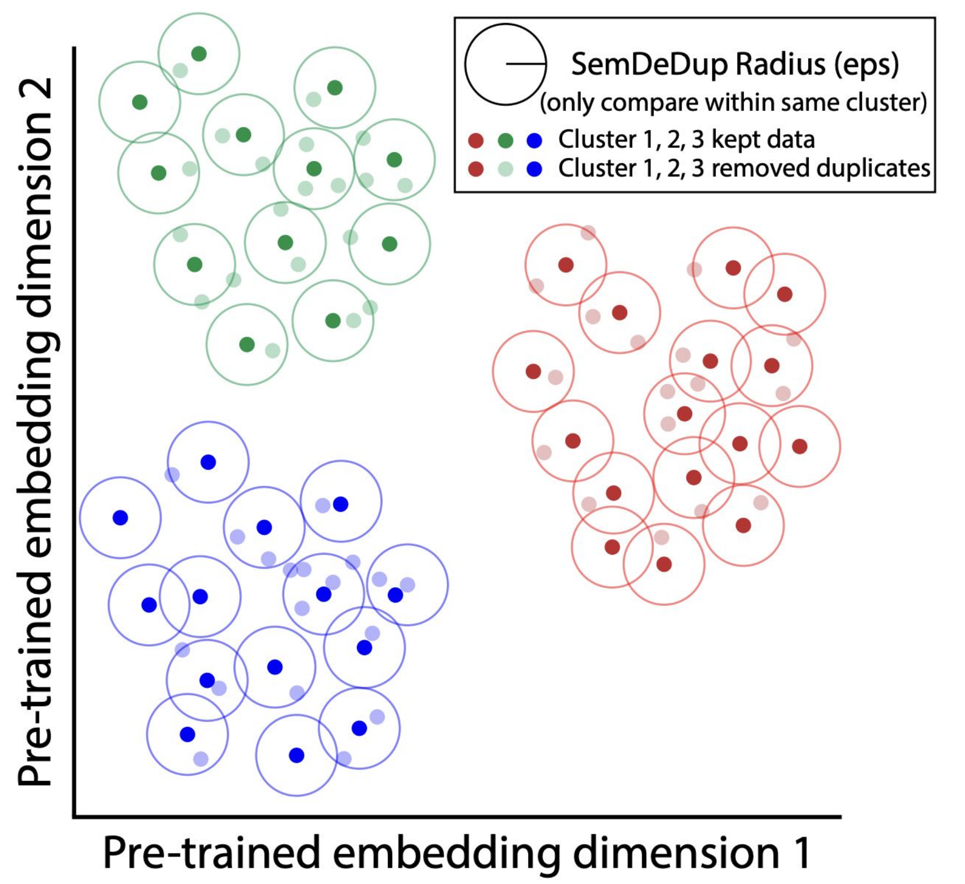
Key Questions

SemDeDup: Data-efficient learning at web-scale through semantic deduplication

Amro Abbas[1] **Kushal Tirumala**[1*] **Dániel Simig**[1*] **Surya Ganguli**[2] **Ari S. Morcos**[1*]

[1]Meta AI (FAIR) [2]Department of Applied Physics, Stanford University

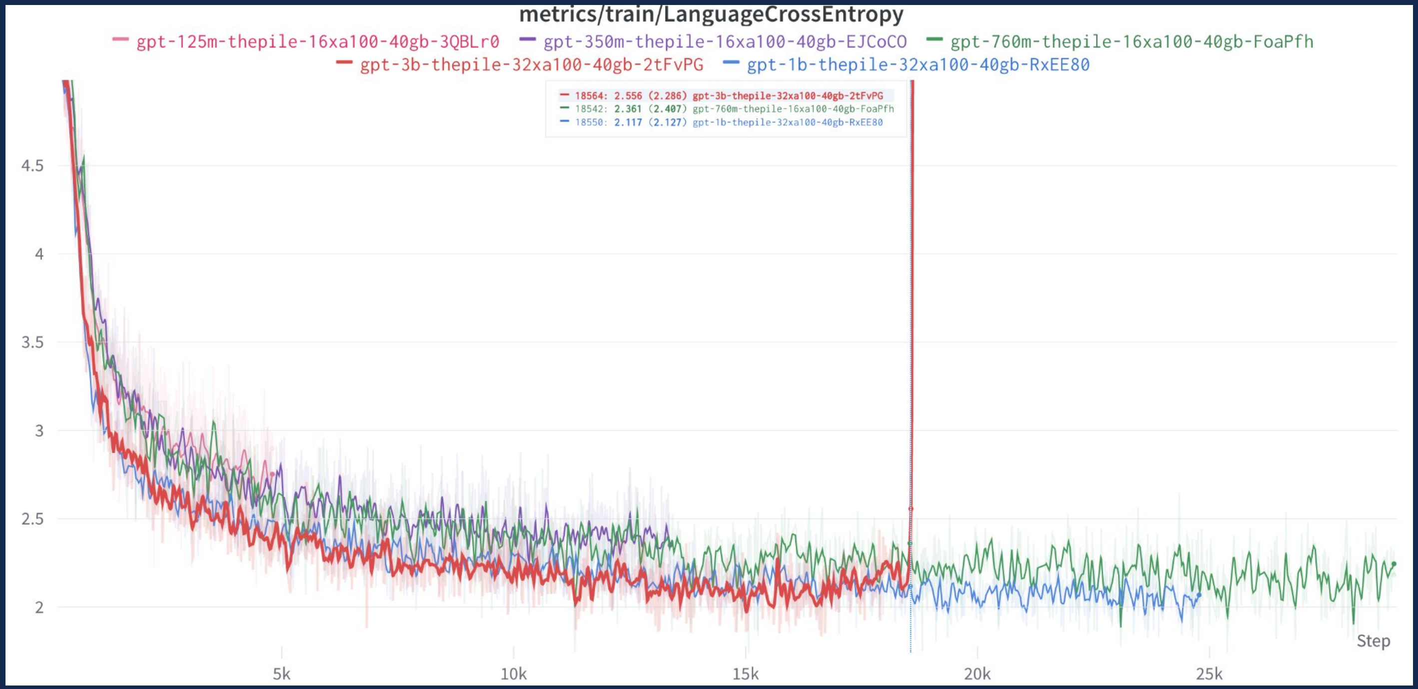
Abstract: Progress in machine learning has been driven in large part by massive increases in data. However, large web-scale datasets such as LAION are largely uncurated beyond searches for exact duplicates, potentially leaving much redundancy. Here, we introduce SemDeDup, a method which leverages embeddings from pre-trained models to identify and remove "semantic duplicates": data pairs which are semantically similar, but not exactly identical. Removing semantic duplicates preserves performance and speeds up learning. Analyzing a subset of LAION, we show that SemDeDup can remove 50% of the data with minimal performance loss, effectively halving training time. Moreover, performance increases out of distribution. Also, analyzing language models trained on C4, a partially curated dataset, we show that SemDeDup improves over prior approaches while providing efficiency gains. SemDeDup provides an example of how simple ways of leveraging quality embeddings can be used to make models learn faster with less data.

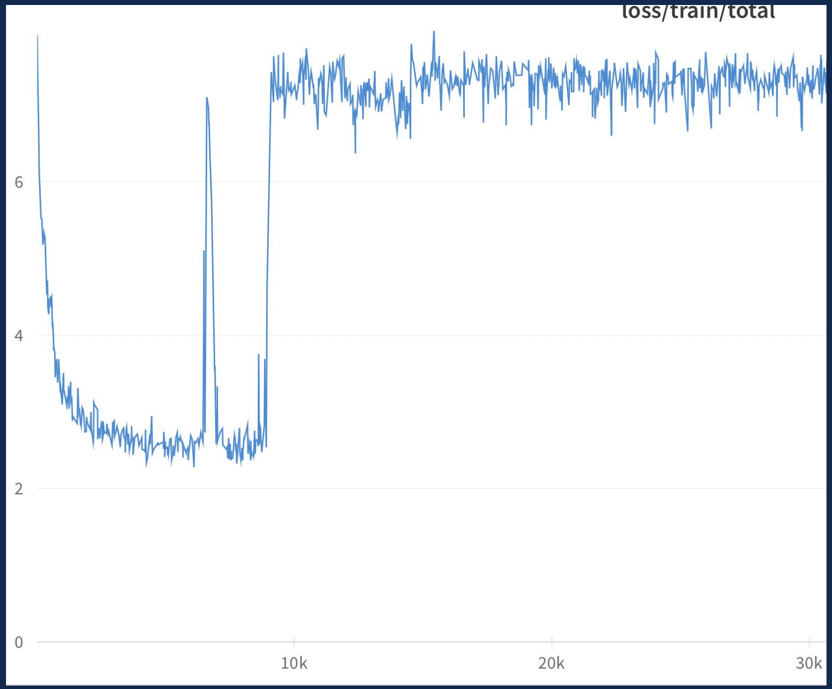


And then you train…

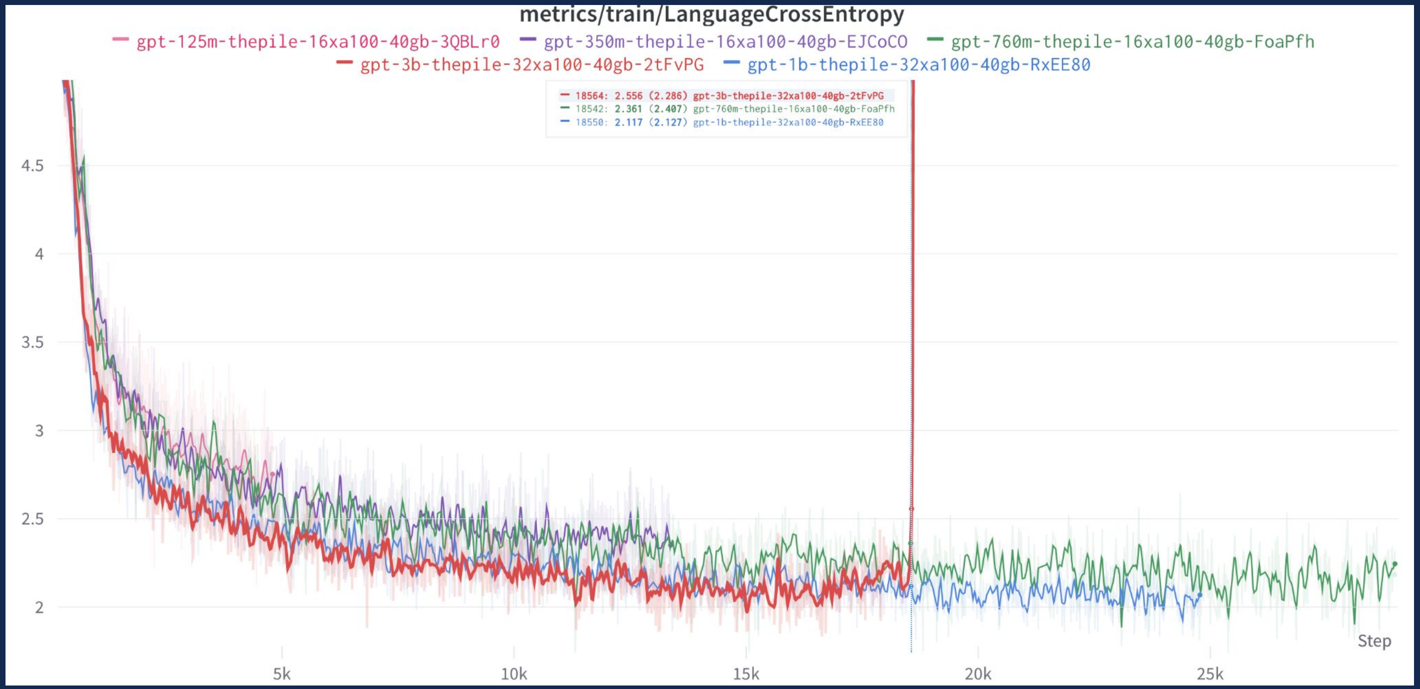
And then you train…
…and all hell breaks loose

Loss Spikes


metrics/train/LanguageCrossEntropy
gpt-125m-thepile-16xa100-40gb-3QBLr0
gpt-350m-thepile-16xa100-40gb-EJCoCO
gpt-760m-thepile-16xa100-40gb-FoaPfh
gpt-3b-thepile-32xa100-40gb-2tFvPG
gpt-1b-thepile-32xa100-40gb-RxEE80
18564: 2.556 (2.286) gpt-3b-thepile-32xa100-40gb-2tFvPG
18542: 2.361 (2.407) gpt-760m-thepile-16xa100-40gb-FoaPfh
18550: 2.117 (2.127) gpt-1b-thepile-32xa100-40gb-RxEE80
4.5
4
3.5
3
2.5
2
Step
5k
10k
15k
20k
25k


loss/train/total
6
4
2
0
10k
20k
30k

Loss Spikes



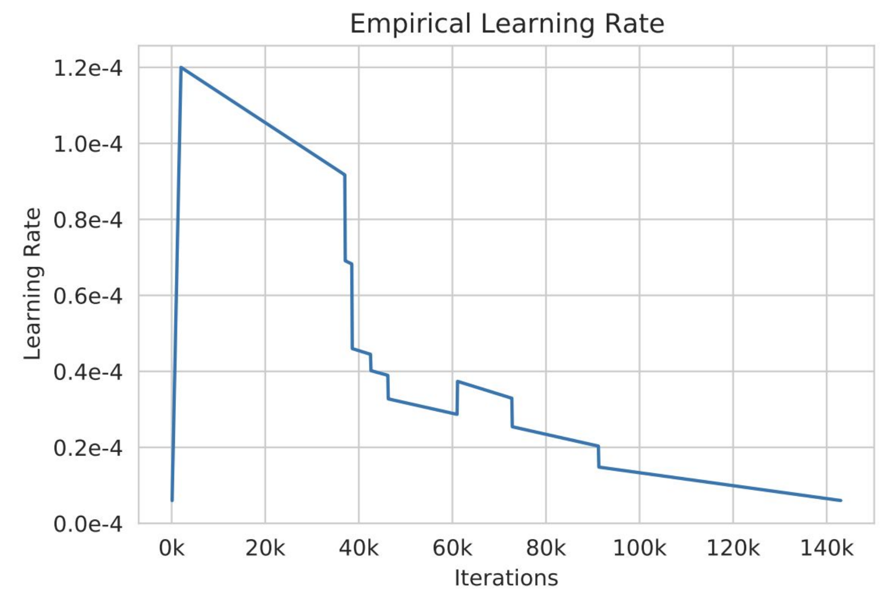


Figure 1: **Empirical LR schedule.** We found that lowering learning rate was helpful for avoiding instabilities.

Mitigations: Rollback, change seed, retry, pray OR

Fix the architecture so loss spikes don't happen

Hardware Failures

Jonathan Frankle 12 days ago
Today has been a bad day for GPUs. Please press F to pay your respects to our fallen comrades.
F 18

Node-Health-Bot APP 6:54 PM
This little piggy (🐷 node `inst-pwxlx-r7z2-workers`) is 💀**DEAD**💀 on cluster `r7z2`

Priority	**Type**
Critical	*Node Died*
Reason	**Message**
GPU is lost	*GPU at index 2 was detected to be not ready: GPU is lost*

Hardware Failures

Why is this a problem? GPU failure rates are really high.
1 node out of 16 every week, approximately.
Varies by cluster, region, and weather.

Training is not fault tolerant. Every time you have a failure, run dies and you need to recover.

Training only works on certain multiples of GPUs. Batch sizes are only divisible by certain numbers.

Checkpoints and datasets are huge.

Hardware Failures

Mitigations:

- Automatic detection of failures.
- Keeping spare GPUs available (and using them for lower-priority stuff until they're needed)
- Sharded checkpointing.
- Data loaders with random access.

The Details

How big of a model should you use?

Smaller models are better for inference and anecdotally are easier to train.

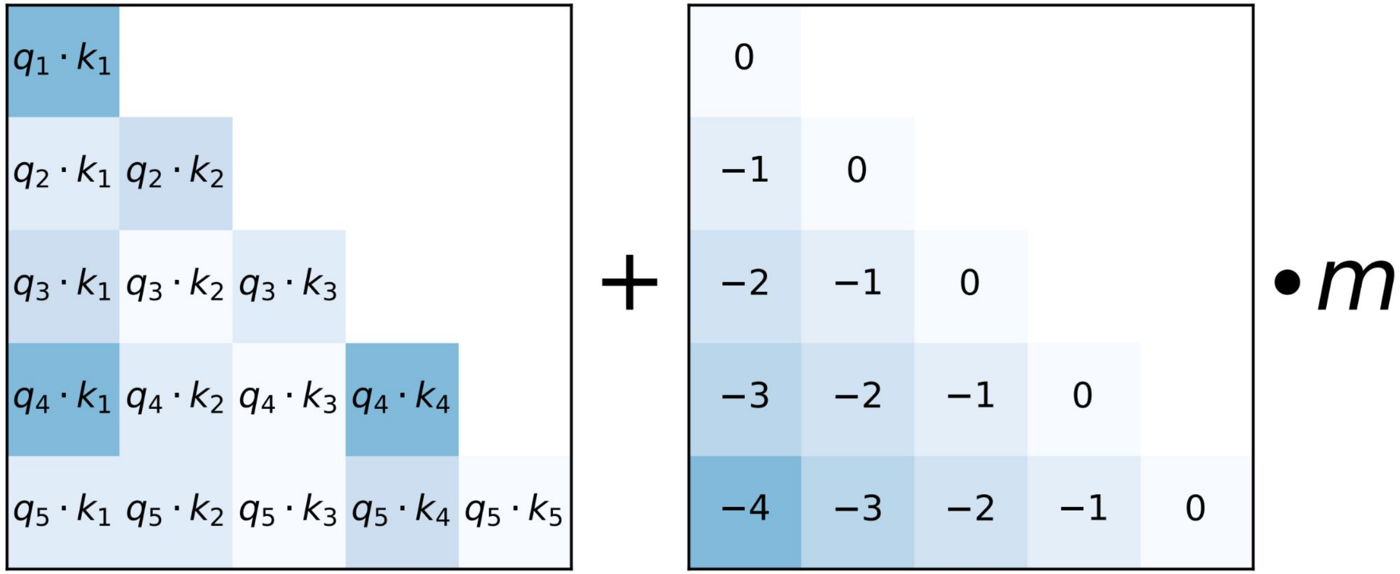
Bigger models are closer to Chinchilla-optimal, i.e., they’re cheaper to train.

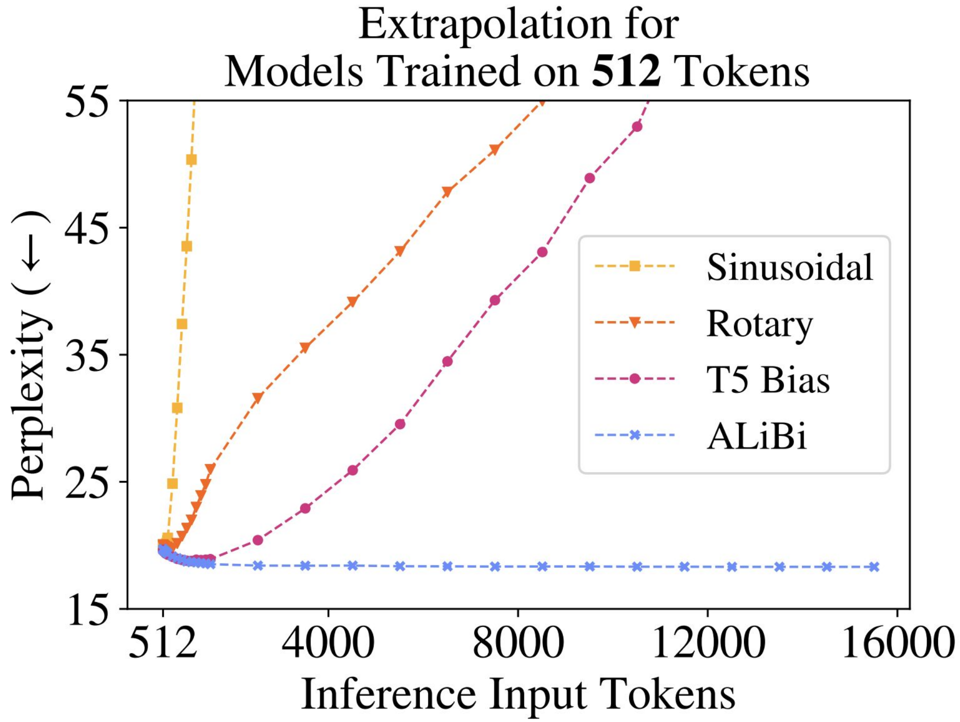
Bigger models may be better at reasoning???

Positional Encodings

Train Short, Test Long: Attention with Linear Biases Enables Input Length Extrapolation

Ofir Press[1,2] **Noah A. Smith**[1,3] **Mike Lewis**[2]





What sequence length to choose?

Do you have the data to support longer contexts?

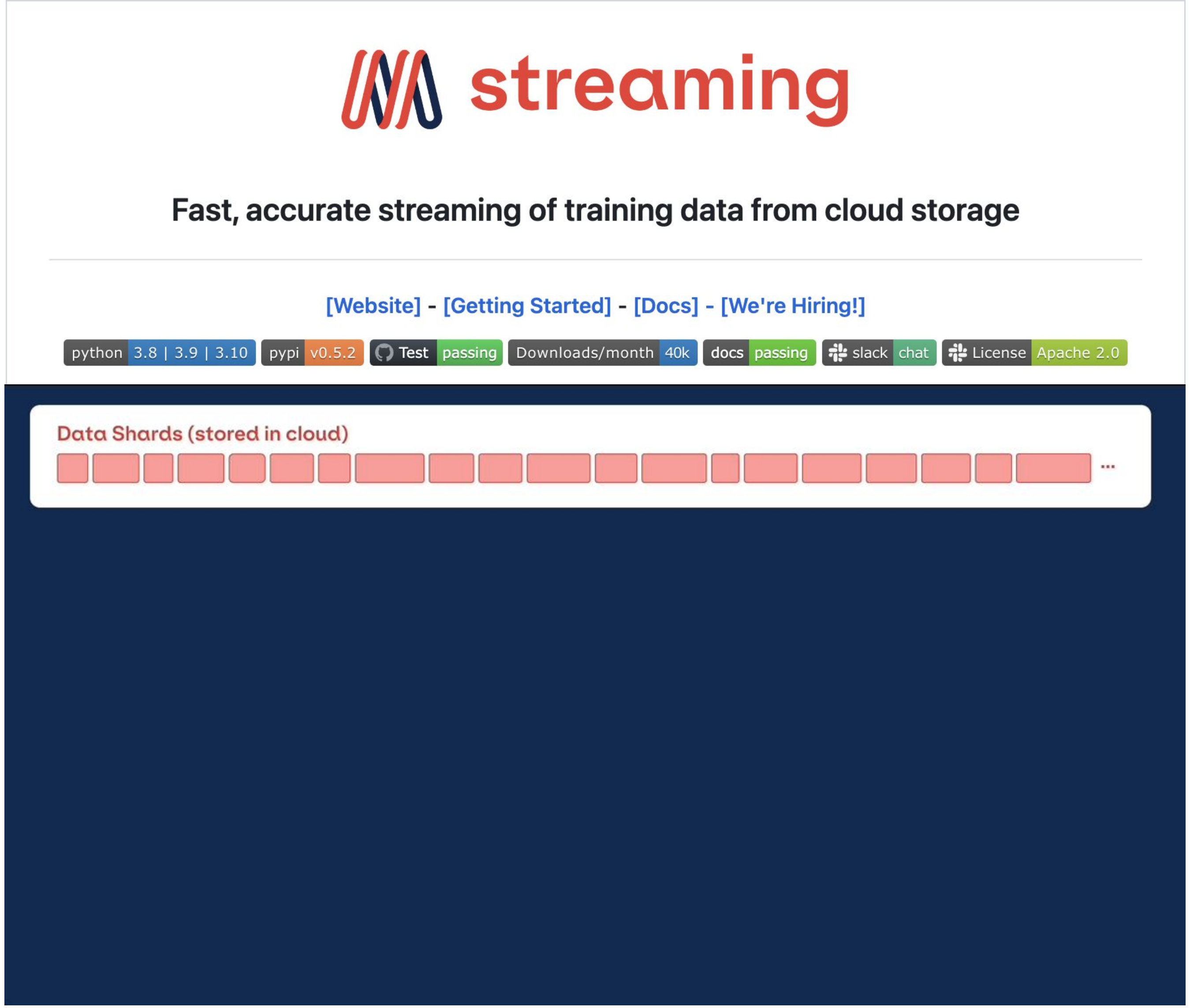
Longer contexts eventually slow down training.

What tokenizer should you use?

What tokenizer should you use?

¯_(ツ)_/¯

How should you store your data?



Friendly Advice

Start small and work your way up.

Trust nothing you read in the literature.
Test everything for yourself.

Do not trust your intuition, received wisdom, or a rumor you heard about OpenAI. Test everything.

Do the math.

So you want to build an LLM…

Jonathan Frankle
Chief Scientist, MosaicML

mosaicML

www.github.com/mosaicml/llm-foundry
www.github.com/mosaicml/streaming
www.github.com/mosaicml/composer