Challenges and Opportunities Building Open LLMs

Iz Beltagy

We need language models that are (truly) open!

Transparent

Reproducible

Accessible

Which one of these models is “Open”?

GPT4, ChatGPT, BARD
Llama, LLama2,
MPT, Falcon
Pythia, GPT-J .. (EleutherAI)
BLOOM

Do you want to use an existing LLM as a blackbox to build an application

or

Research Language models and advance them?

Which one of these models is “Open”?

	SOTA	API	Model weights	Data	Training Code	Ablations	Wandb logs
GPT4, ChatGPT, BARD	✓	✓	✗	✗	✗	✗	✗
Llama, LLama2,	✓	✓	✓	✗	✗	✗	✗
MPT, Falcon	✓	✓	✓	✓✗	✗	✗	✗
Pythia, GPT-J .. (EleutherAI)	✗		✓	✓	✓	✓	✗
BLOOM	✗		✓	✓	✓	✓	✗
olmo + dolma	This is the goal	✓	✓	✓	✓	✓	✓

Science of LMs:

- Open, documented, and reproducible: enable a larger, diverse AI community to understand, study, evaluate, and advance LMs and their components; narrow the private/public gap

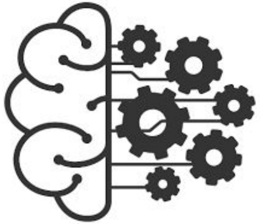
LMs for Science:

- Advance scientific understanding and discovery by training on scientific text (eventually serve Semantic Scholar projects/users)
- Promote AI literacy through transparency and public demos

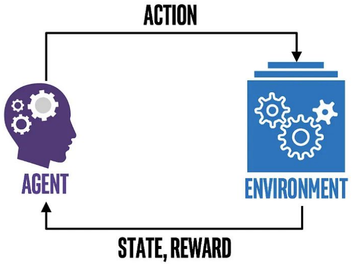
Target OLMo 1.0 releases

Open Data
- Pretraining data
- Demonstration data

Evaluation:
- Evaluation suite
- Tasks & efficiency



Training & Inference:
- Open training code
- Models @7B & 70B
- Inference code
- Instruction tuning



Public Demo & API
- Demo
- Human interaction & feedback

Model Impact
- Impact License

Agenda

Model Construction

1. **Dolma:** Data to feed OLMo's appetite
2. **Evaluation**
3. **Training**

Model Adaptation: the Tülu model

What's next?

OLMo Data

Dolma: An Open Corpus of 3 Trillion Tokens for Language Model Pretraining Research

Luca Soldaini[α*] Rodney Kinney[α*] Akshita Bhagia[α*] Dustin Schwenk[α*]

David Atkinson[α] Russell Authur[α] Khyathi Chandu[α] Jennifer Dumas[α]
Li Lucy[β] Xinxi Lyu[α] Ian Magnusson[α] Aakanksha Naik[α]
Crystal Nam[α] Matthew E. Peters[α] Abhilasha Ravichander[α] Zejiang Shen[τ]
Emma Strubell[χ,α] Nishant Subramani[χ,α] Oyvind Tafjord[α] Evan Pete Walsh[α]

Hannaneh Hajishirzi[α,ω] Noah A. Smith[α,ω] Luke Zettlemoyer[ω]
Iz Beltagy[α] Dirk Groeneveld[α] Jesse Dodge[α]

Kyle Lo[α*]

[α]Allen Institute for AI [β]University of California, Berkeley [χ]Carnegie Mellon University
[τ]Massachusetts Institute of Technology [ω]University of Washington

Pretraining Dataset

Given all the text in the web and other sources, how to build your pretraining data

Design of pretraining dataset is understudied in the literature

- Not perfectly clear what makes a high-quality pretraining dataset
- Not perfectly clear how pretraining data characteristics translate to downstream performance

What is Dolma?

DOLMA: Data to Feed OLMo’s Appetite

Dataset for Pretraining OLMo
→ Lots of text (3.1T tokens)
→ Large scale, high quality

Toolkit!
→ Transforms raw text to a pretraining corpus
→ Good design & performance
→ Common filters; fast global deduplication

Data Distribution & Features

Data mix:

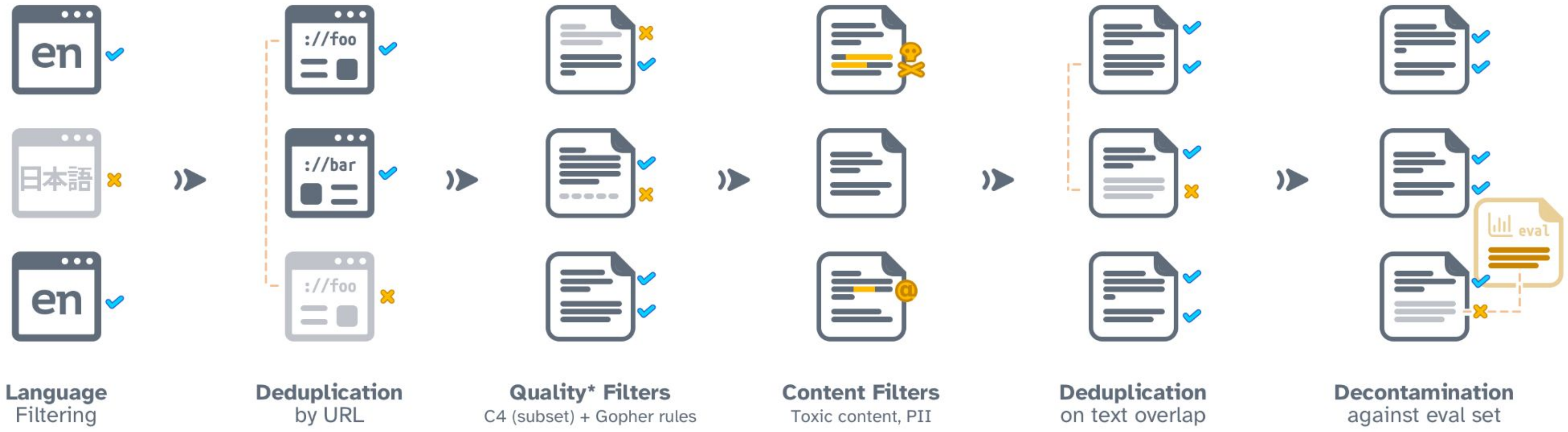
- ❑ English only
- ❑ Web data
- ❑ Just enough code
- ❑ Diverse domains

2.6T	Web	Commoncrawl & C4
430B	Code	Stack
57B	Science	Semantic Scholar
8B	Knowledge	Wikipedia & books

Processing:

- ❑ Quality filtering:
  - ❑ toxicity detection; personal information identification, …
- ❑ Deduplication
- ❑ Decontamination

Example of Processing: Web



How Closed LMs Prepare Their Data

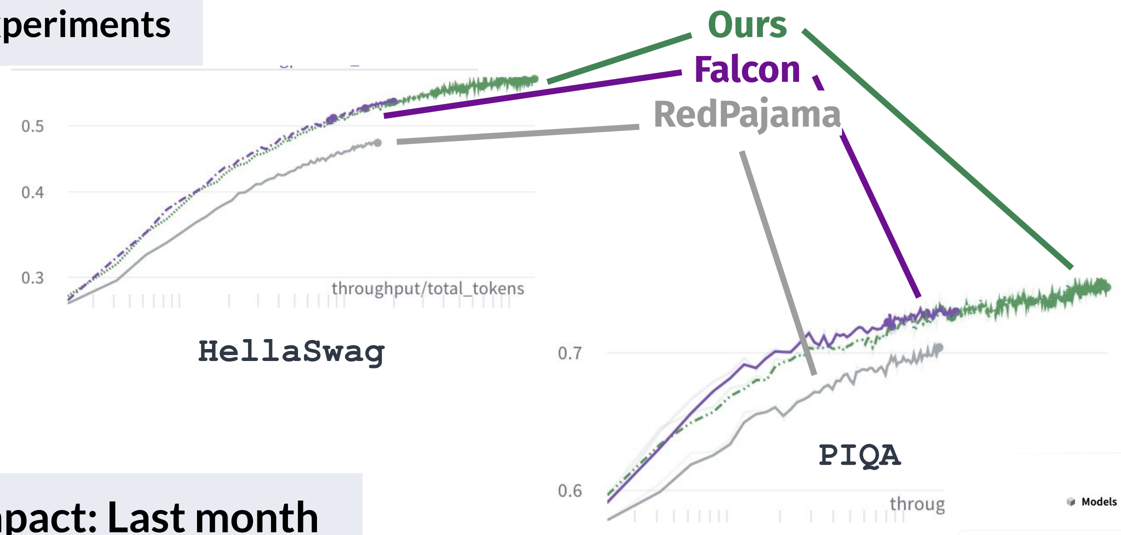
Model	Num Tokens**	Data Provenance?	PII ID + filtering method	Toxicity ID + filtering method	Lang ID + filtering method	Quality filtering method	Dedup method	Decontam method
LLaMA 2 (Jul 2023)	2T	~	✔	?	✔	?	?	?
PaLM 2 (May 2023)	?	~	✔	✔	✔	✔	?	✔
GPT-4 (Mar 2023)	?	?	?	✔	?	?	?	✔
Claude* (Mar 2023)	?	?	?	?	?	?	?	?
LLaMA (Feb 2023)	1.4T	✔	?	?	✔	✔	✔	?
GLM (Oct 2022)	400B	~	?	?	?	?	?	✔
OPT (May 2022)	180B	✔	?	?	✔	?	✔	?
PaLM (Apr 2022)	780B	~	?	?	✔	✔	✔	?
Gopher (Dec 2021)	300B	~	?	✔	✔	✔	✔	✔
Jurassic-1 (Aug 2021)	300B	~	?	?	?	?	?	?
GPT-3 (May 2020)	400B	✔	?	?	?	✔	✔	✔

Other Open Datasets

Dataset	Example language models	Tokens**	Sources	License	PII Filter	Toxicity Filter	Language	Quality Filtering	Dedup	Decontam
OSCAR (Jul 2019)	BLOOM (via ROOTS)	1.08B	Common Crawl	Varies by data subset*	○	○	Multilingual (152 langs)	○	●	○
C4 (Oct 2019)	T5, FLAN-T5	156B	Common Crawl	ODC-BY	○	●	English	●	○	○
The Pile (Dec 2020)	GPT-J, GPT-NeoX, Pythia	300B	22 datasets e.g. Common Crawl, scientific text, books, code, Wikipedia, news	Varies by data subset	○	○	English	●	●	●***
ROOTS (Mar 2023)	BLOOM	341B	517 datasets e.g. Github, news, books, scientific text, Wikipedia	Varies by data subset	●	●	Multilingual (59 langs)	●	●	○
RedPajama (Apr 2023)	LLaMa reproduction	1.2T	Common Crawl, C4, Github, Arxiv, Books, Wikipedia, StackExchange	Varies by data subset	○	○	English	●	●	○
RefinedWeb (Jun 2023)	Falcon	600B****	Common Crawl	ODC-By 1.0	○	●	English	●	●	○
Ours (Dolma)	OLMo (Ongoing)	3.08T	Common Crawl, C4, peS2o, Gutenberg, Github, Wikipedia + Wikibooks	ImpACT MR	●	●	English	●	●	●

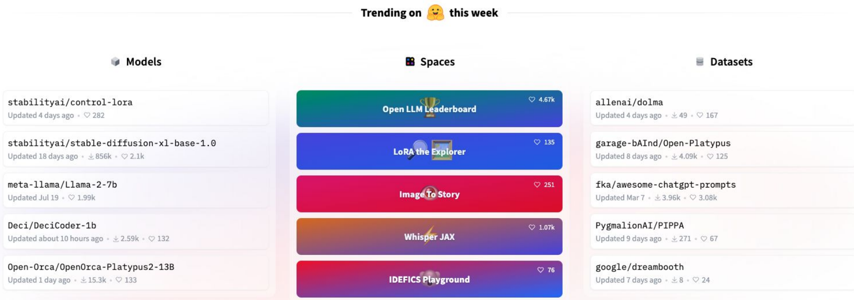
Results

Experiments



Impact: Last month

- Downloads: 320k
- Authorized users: 1,250
- Code repository: >200 stars
- Top trending dataset during the two weeks following release



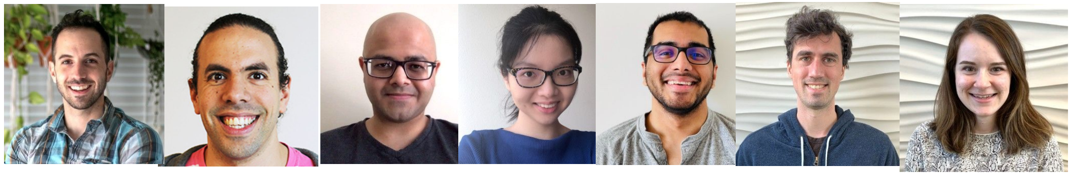
What's next? Dolma 2.0

- **More tokens**
  - ¾ Common Crawl to go; other general domain data providers
- **Better processing**
  - Revisit quality filters through content classifiers; improve other filters
- **Scientific text**
  - More books, more papers. Maybe multimodal?
- **Retrieval & other tools!**
  - Improve Dolma codebase to enable more research

Open Research Questions

- What makes an “oracle” dataset?
- What is the perfect filtering and deduplication method?
- How to best mix domains?
- Where to find pretraining data that doesn’t have copyright issues?

OLMo Evaluation



Pretraining Evaluation is different

Goals: tooling and evaluation to **make sure pretraining is on the right track**

- Evaluation for a trained model
  - Slow and detailed
  - Computationally expensive
  - Can involve human labeling and redteaming
  - e.g: HELM

- Pretraining Evaluation
  - Rapid lightweight evaluation
  - Runs in-loop for early detection of training issues
  - Goes beyond just training and validation loss
  - e.g: Catwalk (ours) and eleutherAI eval harness

Pretraining Evaluation is different

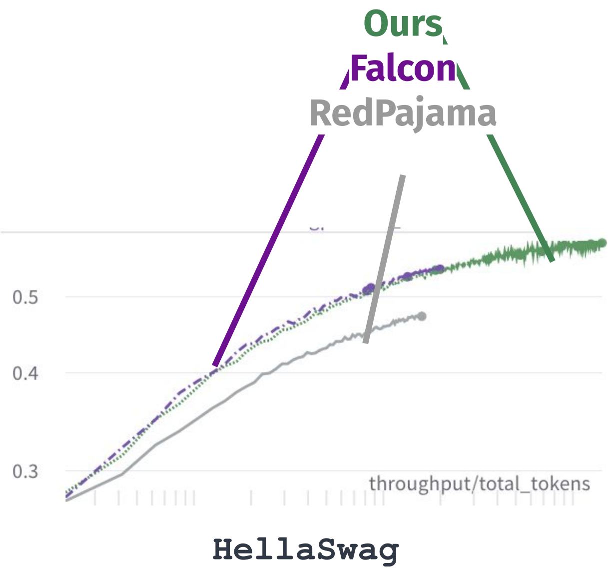
Pretraining Evaluation:

- Goes beyond just training and validation loss
  - Intrinsic evaluation (validation loss)
    - Good for model ablations. Why?

Because validation loss is super strongly correlated with downstream performance

  - Extrinsic evaluation (downstream performance)
    - Good for data ablations. Why?

Because validation loss is not comparable once the training data changes



Monitoring

Goals: tooling and evaluation to **make sure pretraining is on the right track**

What else to monitor other than validation loss and in-loop downstream eval?

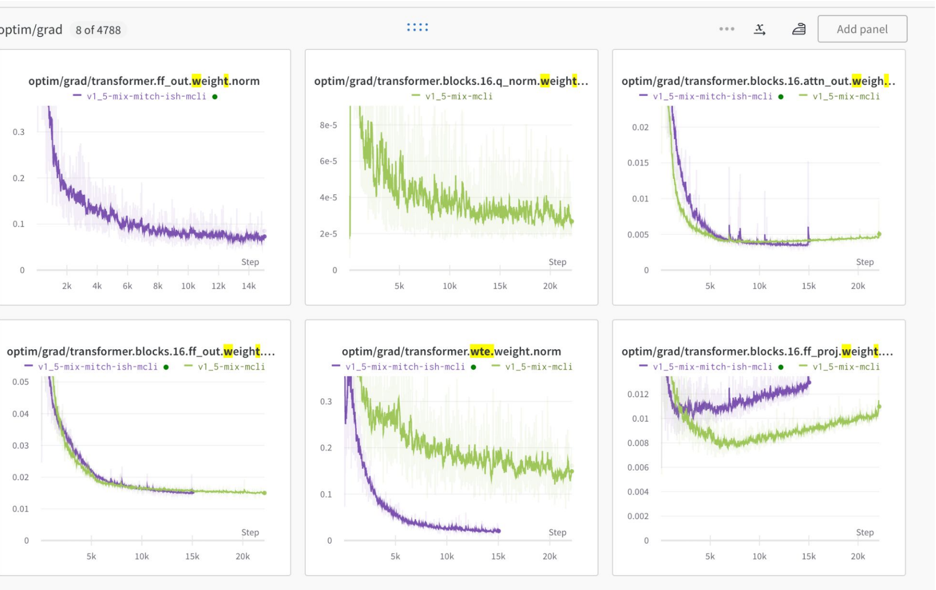
- Optimizer state
- Gradients
- Params
- Activations
- Gradient clipping
- Learning rate
- Throughput
- Total tokens

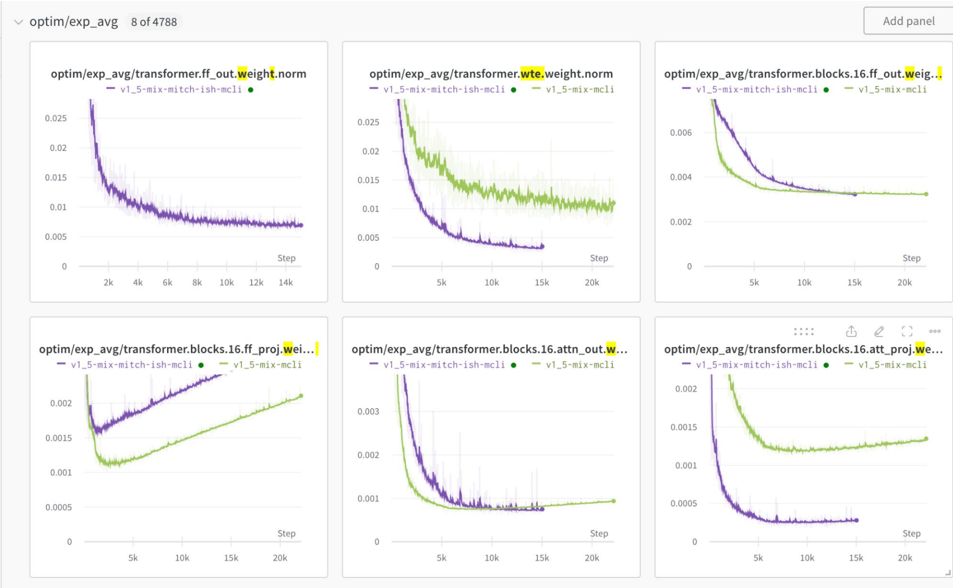
Absolutely essential for debugging

Super helpful for reproducibility

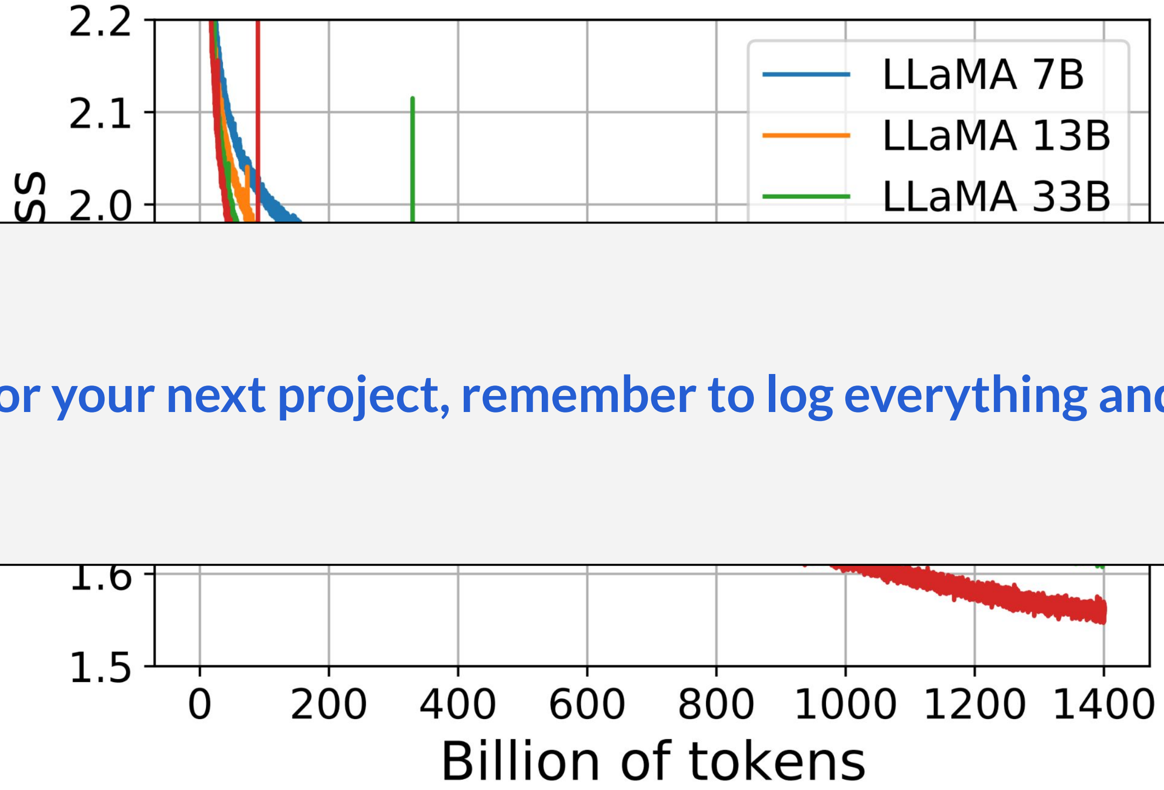
Monitoring

Goals: tooling and evaluation to **make sure pretraining is on the right track**





Monitoring



Helps finding new insights

For your next project, remember to log everything and to release your wandb log

significant underestimate

OLMo Evaluation

Goals: tooling and evaluation to **make sure pretraining is on the right track**

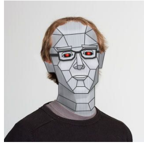
Downstream evaluation

- **New evaluation framework on Catwalk,**
- **18+ core LLM evaluation tasks**, *offline analysis*, **QA** *(MMLU, ARC,..)*, **summarization** *(SciTLDR,..)*, **misc. classification**.
- **In-the-loop evaluation,** early detection of training issues.

Perplexity evaluation

- A **new suite of perplexity tasks** for ensuring progress on core LMing task.
- New techniques for **data decontamination**, ensuring reliability.

OLMo Model Training



Isn’t LLM training a solved problem?

Llama-2 is out, can’t we just follow their setup? No, because

1) The design space is so huge. Every released model is a single datapoint in that space, but it is not the only nor the best point
2) Even if we want to blindly follow it, we can’t because it is not open source
    a) Data is not open source
    b) Training code is not open source (hids low-level but important implementation details)
    c) Training and optimization hyperparamters are not open source
    d) Model and optimizer ablations are not open source
3) And even if it was, how are we going to learn to build the next model and keep advancing the field?

Compute

- Partnering with AMD and LUMI, a supercomputer in Finland
- Total 2M GPU hours
- AMD GPUs are good, but software has a few issues
  - MI250 is comparable to A100
- LUMI can be busy, especially with the global GPU shortage

Training: Highlights

- New training code, adaptable to AMD hardware and NVIDIA
  - Built a platform that ran dozens of data ablations at LUMI.
  - 7B model is trained up to 400B model and still going
  - Results are on-par with comparable-size models

How are we doing? (downstream)

Current checkpoint
300B tokens
Training towards 2T tokens

Are we on the path towards models that can do things?

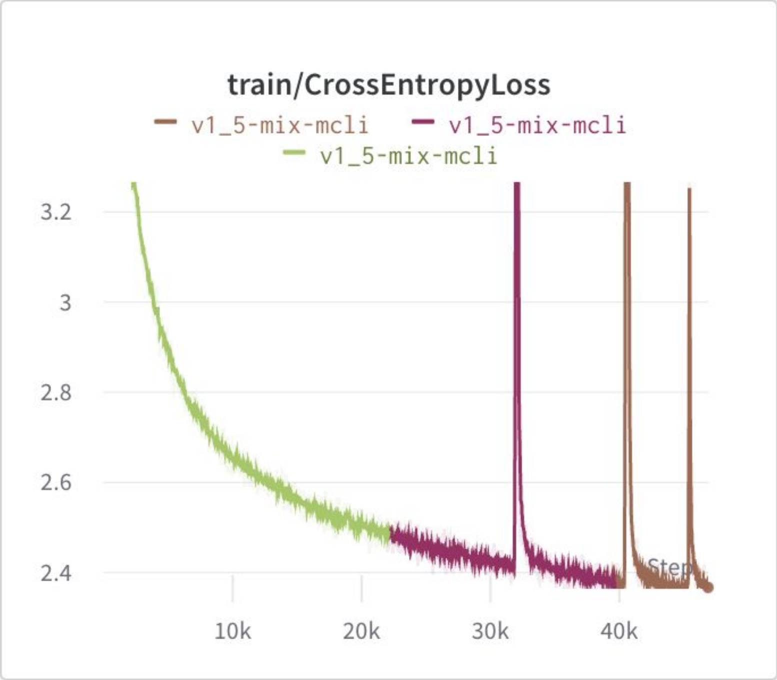
More training tokens

task	num inst	random	Pythia-6.9b step80k	OLMo-medium v1-mix-step70k	Pythia-6.9b step140k	MPT-7b	Llama-7b	XGen-7b 4k-base	Falcon-7b
arc_challenge	299	25	38.8	43.8	44.2	46.5	44.5	45.8	47.5
arc_easy	570	25	58.8	61.1	61.9	70.5	57.0	67.0	70.4
boolq	1000	50	63.2	64.6	61.1	74.2	73.1	73.6	74.6
copa	100	50	77.0	85.0	84.0	85.0	85.0	80.0	86.0
hellaswag	1000	25	59.9	70.4	63.8	77.6	74.5	67.2	75.9
openbookqa	500	25	43.8	48.4	45.0	48.6	49.8	46.4	53.0
piqa	1000	50	73.7	76.0	75.1	77.3	76.3	74.5	78.5
rte	277	50	52.4	49.5	60.7	62.8	53.1	57.8	61.7
sciq	1000	25	90.0	88.4	91.1	93.7	89.5	92.6	93.9
sst	872	50	52.2	54.1	62.3	75.8	53.0	56.0	49.1
winogrande	1000	50	61.5	63.9	62.0	69.9	68.2	68.3	68.9
wnli	71	50	50.7	46.5	38.0	47.9	56.3	52.1	47.9
Average		**39.6**	**60.2**	**62.6**	**62.4**	**69.2**	**65.0**	**65.1**	**67.3**

Same # of tokens

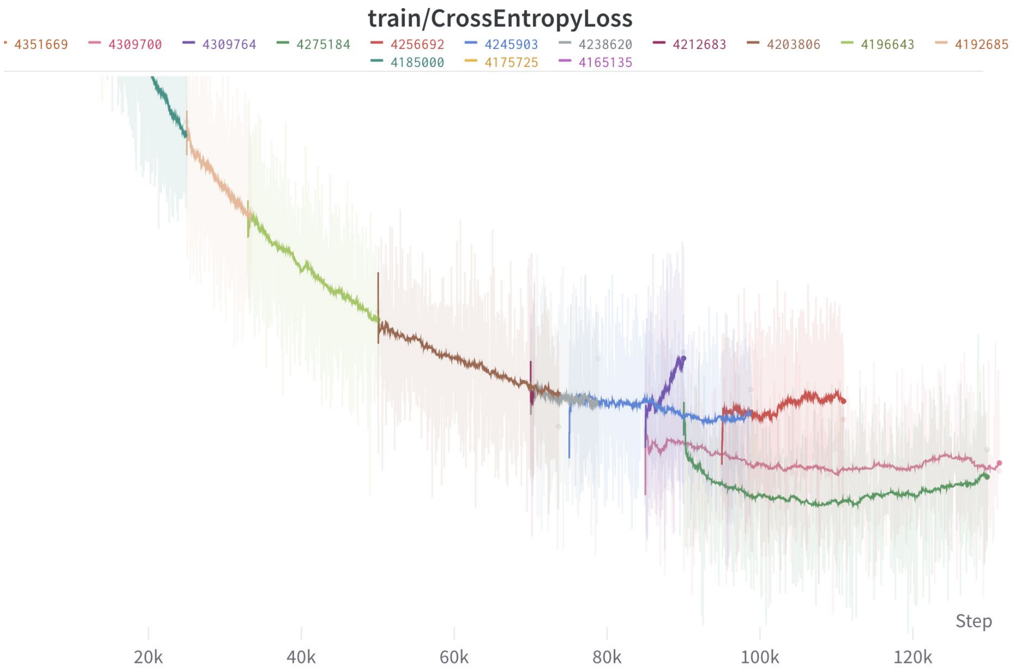
Challenges - loss spikes

- Reduces model performance if it recovered
- Various mitigation strategies
- If configured correctly, 7B shouldn’t have any
- Causes are plenty
  - Noisy data
  - Loss of precision
  - Model learning something new
  - … others



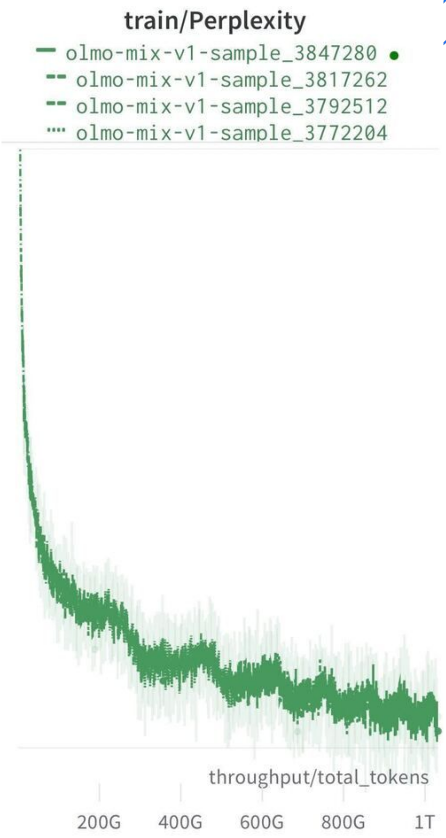
Challenges - slow loss increase

- Not reported in the literature
- Known hyperparams for 300B-tokens params don't necessarily work for 2T params



Challenges - torch.permute is not random

- Weird waves in the training loss
- This means data is not IID
- Turns out torch.permute is not very random

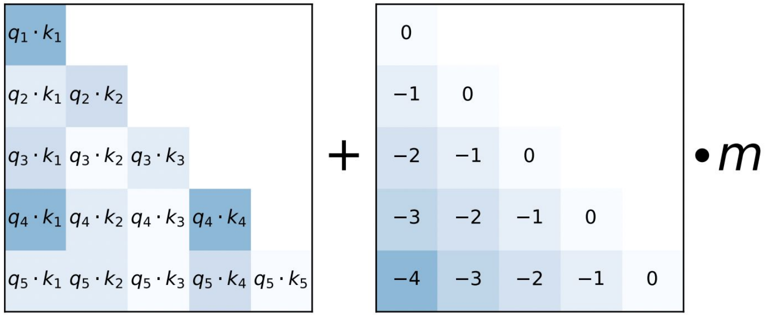


Challenges - software issues with AMD GPUs

- `torch.compile` diverges on AMD
- `torch.nn.LayerNorm(bias=None)` sigfaults
- Triton is not supported

Challenges - numerical stability

- For speed, we use bf16
- Bf16 is better than fp16, but still suffers from loss of precision compared to fp32
- e.g: Alibi bias matrix
- Which parts of the model should run in fp32?
    - Torch autocast handles a lot but not enough
    - e.g. torch.all_reduce should be in fp32



Challenges - position embedding

- We still need position embedding that can extrapolate
- But doesn’t inject position information in attention matrix as in Alibi

$$
\begin{bmatrix}
q_1 \cdot k_1 & & & & \\
q_2 \cdot k_1 & q_2 \cdot k_2 & & & \\
q_3 \cdot k_1 & q_3 \cdot k_2 & q_3 \cdot k_3 & & \\
q_4 \cdot k_1 & q_4 \cdot k_2 & q_4 \cdot k_3 & q_4 \cdot k_4 & \\
q_5 \cdot k_1 & q_5 \cdot k_2 & q_5 \cdot k_3 & q_5 \cdot k_4 & q_5 \cdot k_5
\end{bmatrix}
+
\begin{bmatrix}
0 & & & & \\
-1 & 0 & & & \\
-2 & -1 & 0 & & \\
-3 & -2 & -1 & 0 & \\
-4 & -3 & -2 & -1 & 0
\end{bmatrix}
\bullet m
$$

Wandb demo

Wandb demo

Summary

Summary

- We need Open LLMs
- Building LLMs is still challenging with lots of open questions
- Pretraining Dataset
- Evaluation
- Training
- Alignment, human feedback, continual learning

Thanks!