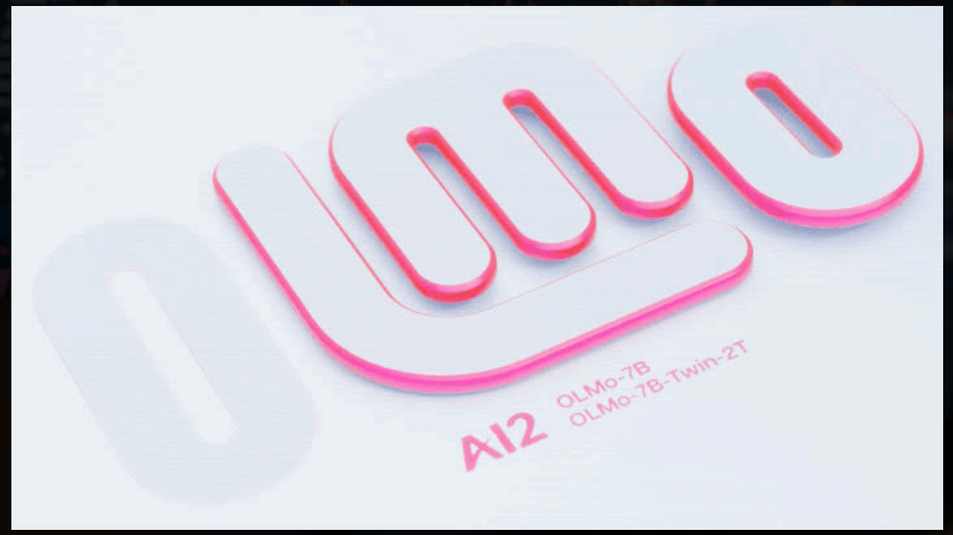
The Open Language Model (OLMo)

AI MAKERSPACE



Presented by:
Dr. Greg & The Wiz 🪄

🎯 ALIGNING OUR AIM

AI
MAKERSPACE

BY THE END OF THE SESSION

- What “truly open” means for an LLM today
- Understand the **data**, **model architecture**, and **evaluation** methods used for OLMo
- How to **fine-tune OLMo** on **open-instruct**!

OVERVIEW

- AI2 & OLMo
- Dolma, Paloma, and more!
- Instruct-Tuning OLMo
- Conclusions & QA

AI2 & OLMo

ALLEN INSTITUTE FOR AI

Non-profit research institute (ca. 2014)

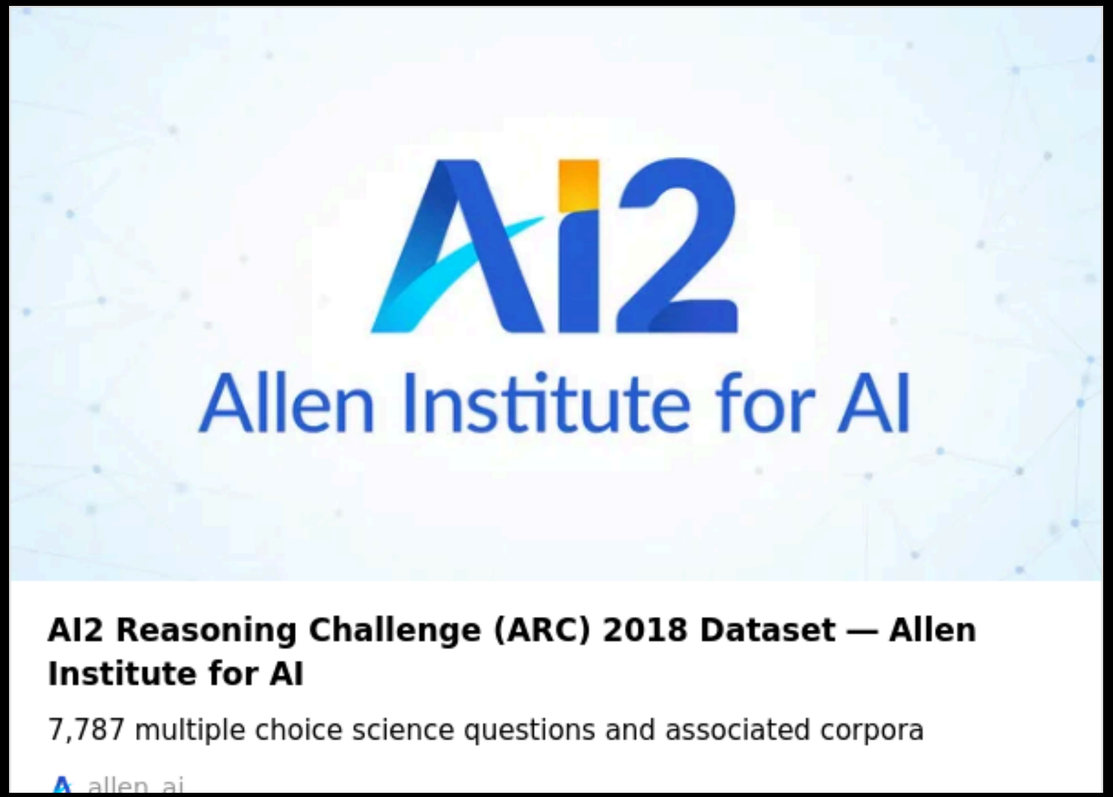
Created by Paul Allen, co-founder of Microsoft

Mission: Conduct **high-impact AI research and engineering in service of the common good.**

ARC

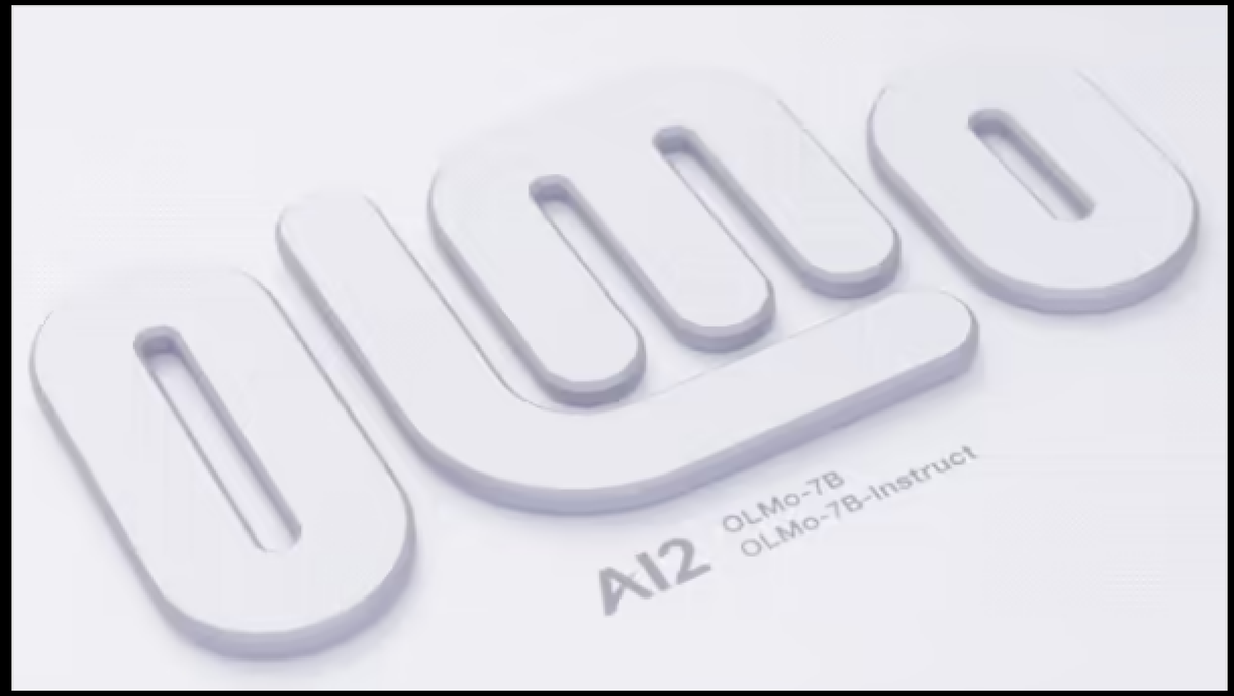
- AI2 Reasoning Challenge, 2018
- “Common Sense” Reasoning
- Still part of HF Open LLM Leaderboard

Average	ARC
80.48	76.02



OPEN LANGUAGE MODEL

“We hope this release will empower and strengthen the open research community and **inspire a new wave of innovation**.”

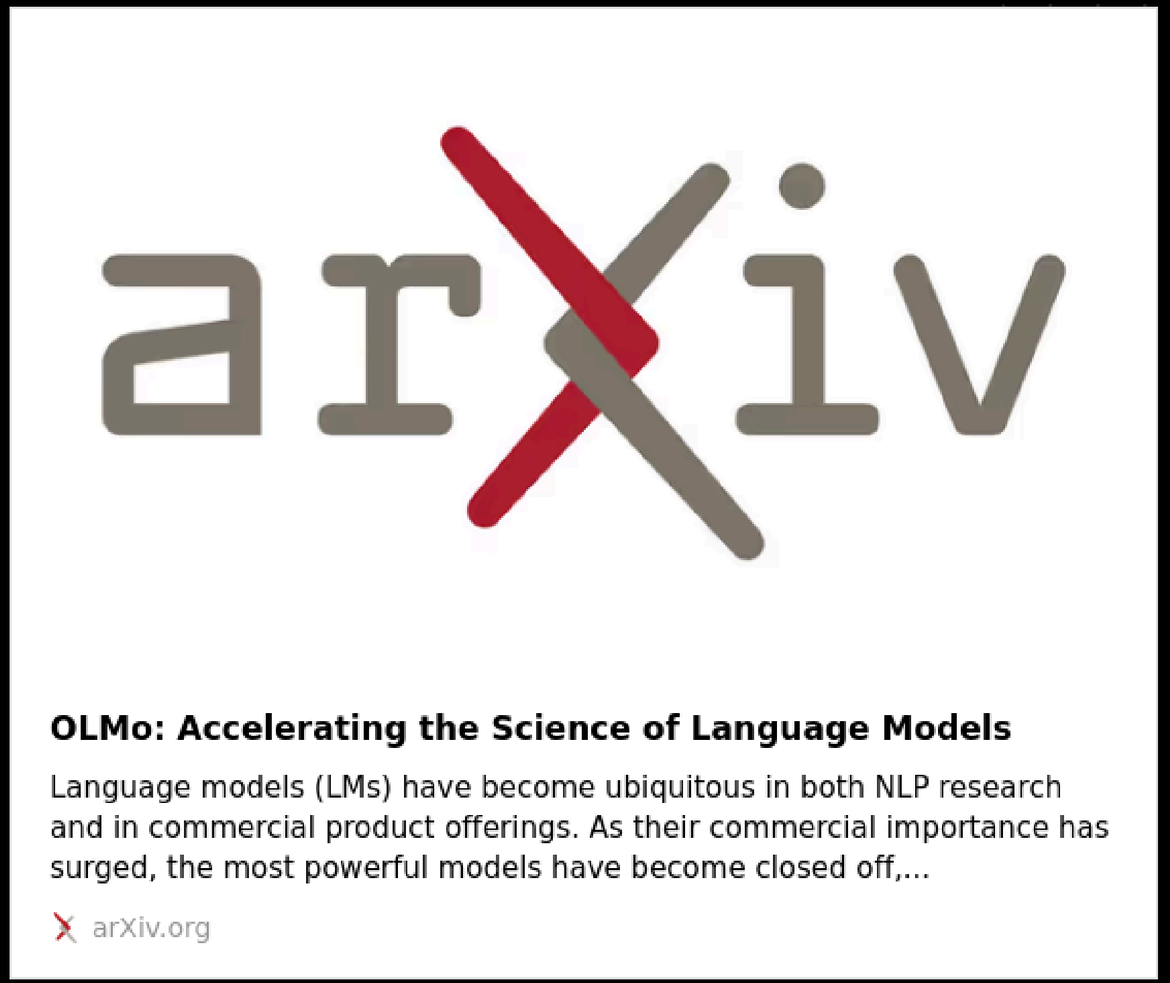


OLMo

OLMo: Accelerating the Science of Language Models

Provides access to:

- Data
- Training code
- Models
- Evaluation code



Recent “Open” LMs

Recent LM releases have varied in their degree of openness. For example, Mistral 8x7B provided model weights and a brief report (Jiang et al., 2024), while LLaMA came with in-depth adaptation training instructions (Touvron et al., 2023b), and Mosaic Pretrained Transformer came with many details, including the dataset distribution, though not the data itself (MosaicML NLP Team, 2023). Falcon’s pretraining data was partially released (Almazrouei et al., 2023), and the most open models—the Pythia suite (Biderman et al., 2023) and BLOOM (BigScience et al., 2022)—released training code, model checkpoints, training data and more.

Where OLMo Fits

“Narrowing the gap from open-source to Llama 2”

Community-Driven AGI via Open-Source LLMs

Explore how LLM360 is democratizing AGI with open-source large language models, fostering an inclusive and ethical AI development environment.

LLM360 /

With OLMo, we release the whole framework from data to training to evaluation tools: multiple training checkpoints across multiple hardware types, training logs, and exact datasets used, with a permissive license. We are not the only team to do this; recent work from LLM360 targets similar goals (Liu et al., 2023). OLMo narrows the gap from their models to state-of-the-art capabilities of models like LLaMA2. This project has benefited from lessons learned from all of these previous efforts with their varying degrees of openness, and we believe that a large, diverse population of open models is the best hope for scientific progress on understanding language models and engineering progress on improving their utility.

Recent “Open” LMs

Recent LM releases have varied in their degree of openness. For example, Mistral 8x7B provided model weights and a brief report (Jiang et al., 2024), while LLaMA came with in-depth adaptation training instructions (Touvron et al., 2023b), and Mosaic Pretrained Transformer came with many details, including the dataset distribution, though not the data itself (MosaicML NLP Team, 2023). Falcon’s pretraining data was partially released (Almazrouei et al., 2023), and the most open models—the Pythia suite (Biderman et al., 2023) and BLOOM (BigScience et al., 2022)—released training code, model checkpoints, training data and more.

"Truly Open"

The OLMo framework encompasses the tools and resources required for building and researching language models. For training and modeling, it includes full model weights, training code, training logs, ablations, training metrics in the form of Weights & Biases logs, and inference code. This first release includes four variants of our language model at the 7B scale corresponding to different architectures, optimizers, and training hardware, and one model at the 1B scale, all trained on at least 2T tokens. We are also releasing hundreds of intermediate checkpoints available as revisions on HuggingFace. For dataset building and analysis, it includes the full training data used for these models, including code that produces the training data, from AI2's Dolma (Soldaini et al., 2024), and WIMBD (Elazar et al., 2023) for analyzing pretraining data. For evaluation, it includes AI2's Catwalk (Groeneveld et al., 2023) for downstream evaluation and Paloma (Magnusson et al., 2023) for perplexity-based evaluation. For instruction-tuning, we released Open Instruct (Ivison et al., 2023; Wang et al., 2023), and we are currently using it to produce an adapted (instruction-tuned and RLHFed) version of OLMo, which we will release soon. Finally, all code and weights are released under the Apache 2.0 License.[1]

*“This is the **first step** in a long series of planned releases, continuing with **larger** models, **instruction-tuned** models, and **more modalities** and **variants** down the line.”*

OLMo Architecture

RECALL GPT 1, 2, 3

GPT ARCHITECTURE

- **12-layer** (block)
- **12-head**
- **117M** params
- **4.5** GB text

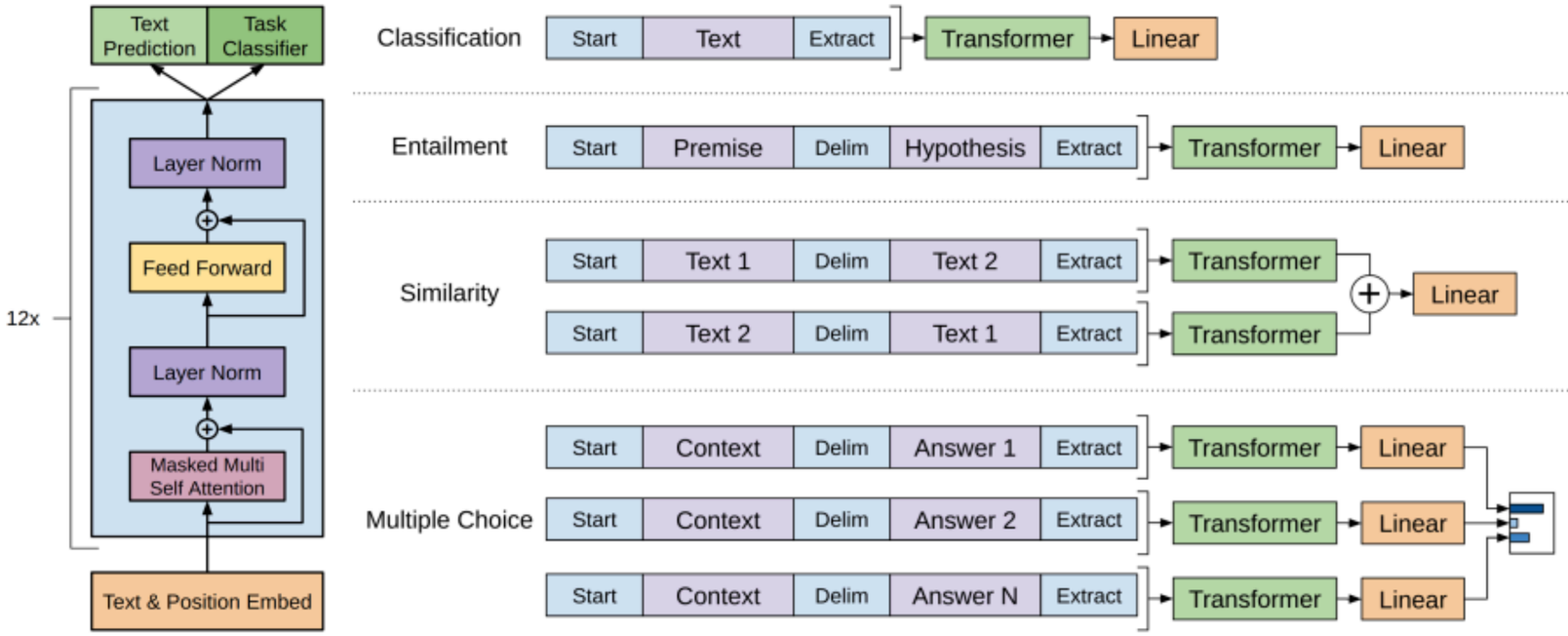
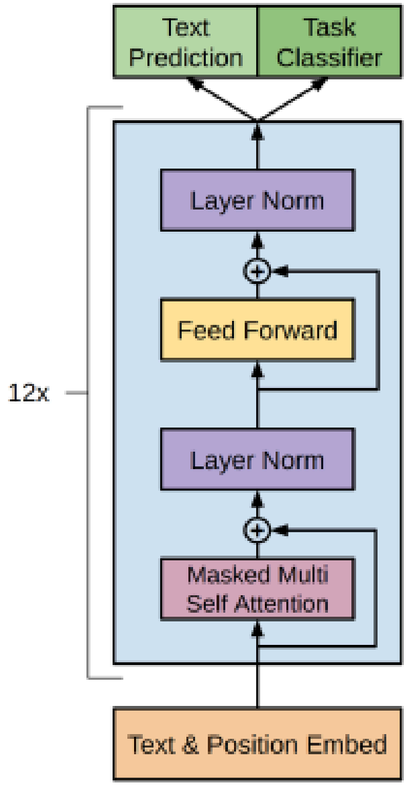


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

GPT-2 ARCHITECTURE

- **48-layer**

  (block)
- ??-head
- **1.5B** params
- **40 GB** text



Parameters	Layers	d_{model}
117M	12	768
345M	24	1024
762M	36	1280
1542M	48	1600

Table 2. Architecture hyperparameters for the 4 model sizes.

GPT-3 ARCHITECTURE

- **96-layer** (block)
- **96-head**
- **175B** params
- **570 GB** text

Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

Table 2.1: Sizes, architectures, and learning hyper-parameters (batch size in tokens and learning rate) of the models which we trained. All models were trained for a total of 300 billion tokens.

OLMo Numbers!

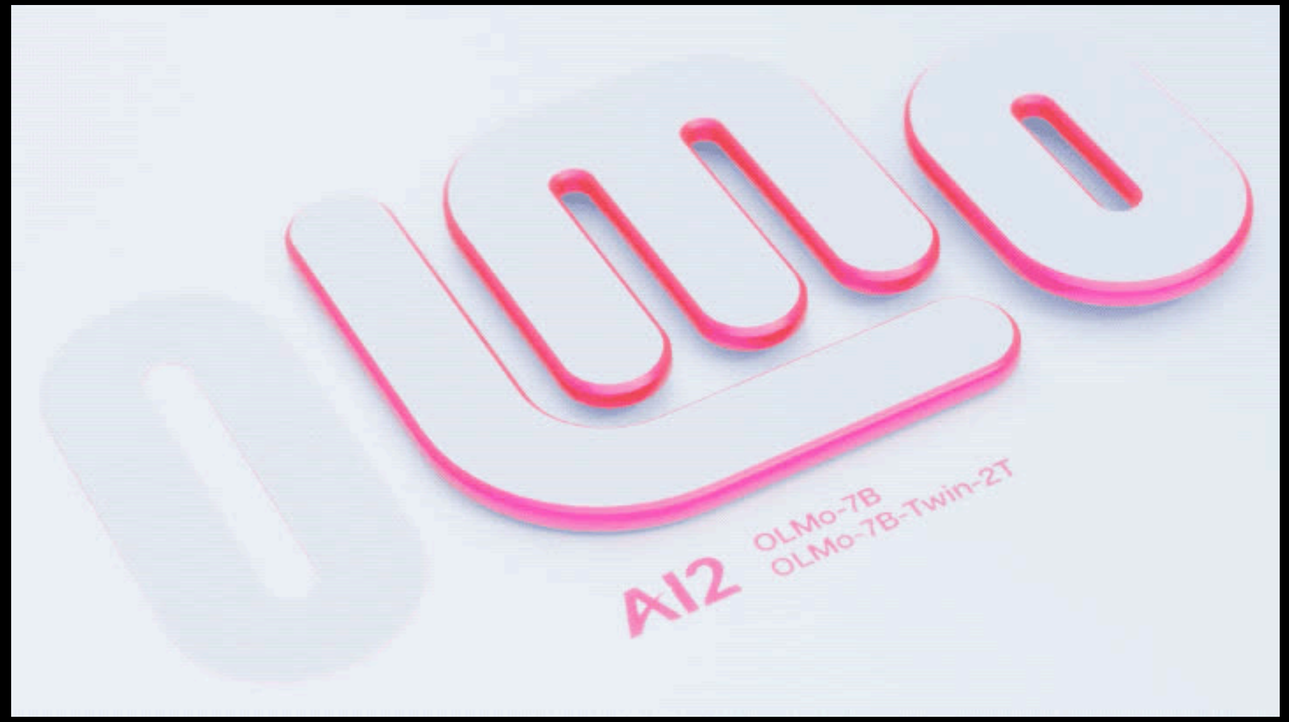
Size	Layers	Hidden Size	Attention Heads	Tokens Trained
1B	16	2048	16	2T
7B	32	4086	32	2.46T
65B*	80	8192	64	

Table 1: OLMo model sizes and the maximum number of tokens trained to.
* *At the time of writing our 65B model is still training.*

Dolma, Paloma, and more!

Puzzle Pieces

- **Data**
  - Dolma, WIMBD
- **Evaluation**
  - Catwalk, Paloma
- **Instruction-Tuning**
  - open-instruct

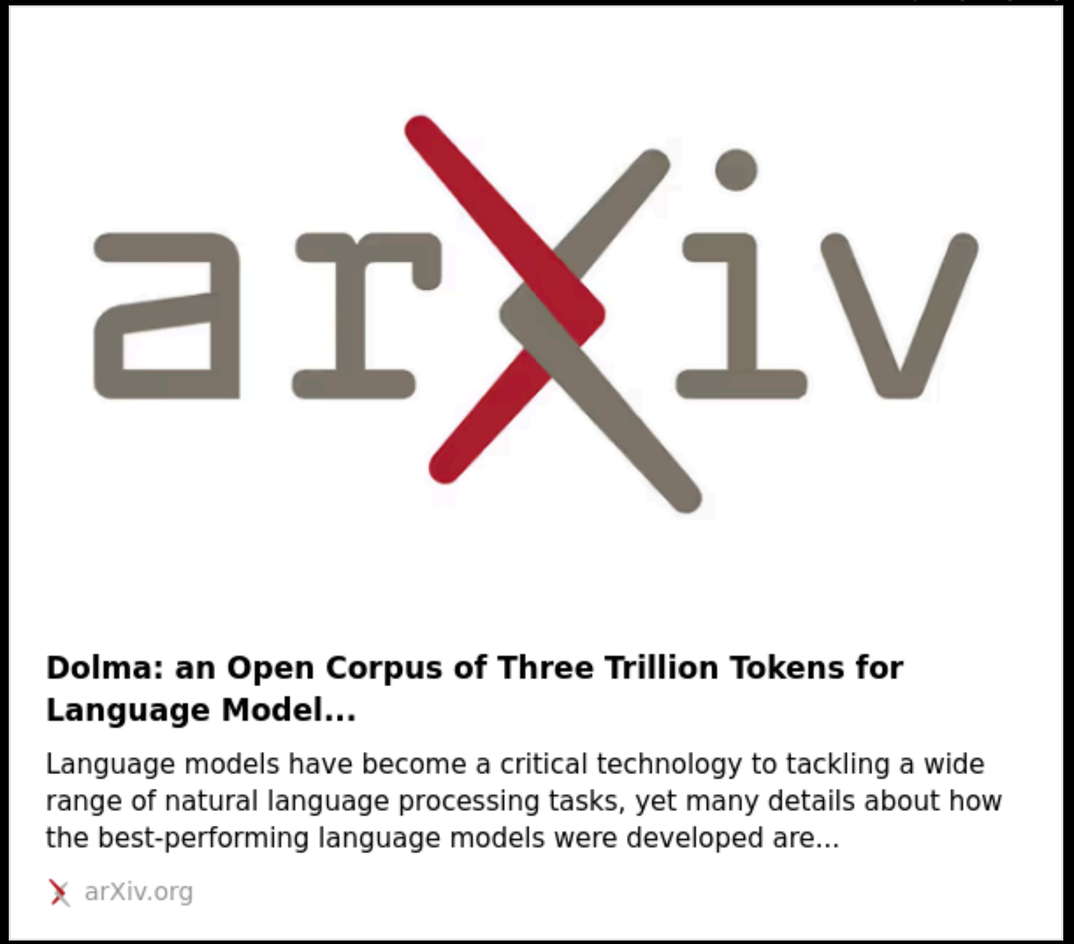


Dolma

dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research

3T Tokens

- web content
- scientific papers
- code
- public-domain books
- social media
- encyclopedic materials



Source	Doc Type	UTF-8 bytes *(GB)*	Documents *(millions)*	Unicode words *(billions)*	Llama tokens *(billions)*
Common Crawl	web pages	9,022	3,370	1,775	2,281
The Stack	code	1,043	210	260	411
C4	web pages	790	364	153	198
Reddit	social media	339	377	72	89
PeS2o	STEM 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**

Table 1: The Dolma corpus at-a-glance. It consists of three trillion tokens sampled from a diverse set of domains sourced from approximately 200 TB of raw text. It has been extensively cleaned for language model pretraining use.

Dolma Toolkit

3.1 Web Pipeline

Figure 1: Overview of the web processing pipeline in Dolma.

WIMBD

What's In My Big Data?

Large text corpora are the backbone of language models. However, we have a limited understanding of the content of these corpora, including general statistics, quality, social factors, and inclusion of evaluation data (contamination). In this work, we propose What's In My Big Data? (WIMBD), a platform and a set of sixteen analyses that allow us to reveal and compare the contents of large text corpora. WIMBD builds on two basic capabilities—count and search—*at scale*, which allows us to analyze more than 35 terabytes on a standard compute node. We apply WIMBD to ten different corpora used to train popular language models, including *C4*, *The Pile*, and *RedPajama*. Our analysis uncovers several surprising and previously undocumented findings about these corpora, including the high prevalence of duplicate, synthetic, and low-quality content, personally identifiable information, toxic language, and benchmark contamination. For instance, we find that about 50% of the documents in *RedPajama* and *LAION-2B-en* are duplicates. In addition, several datasets used for benchmarking models trained on such corpora are contaminated with respect to important benchmarks, including the Winograd Schema Challenge and parts of GLUE and SuperGLUE. We open-source WIMBD's code and artifacts to provide a standard set of evaluations for new text-based corpora and to encourage more analyses and transparency around them: `github.com/allenai/wimbd`.

What's In My Big Data?

Large text corpora are the backbone of language models. However, we have a limited understanding of the content of these corpora, including general statistics, quality, social factors, and...

arXiv.org

Catwalk

Catwalk: A Unified Language Model Evaluation Framework for Many Datasets

Problem

It's **hard** to do **comparisons** across LLMs and benchmarks **at scale**

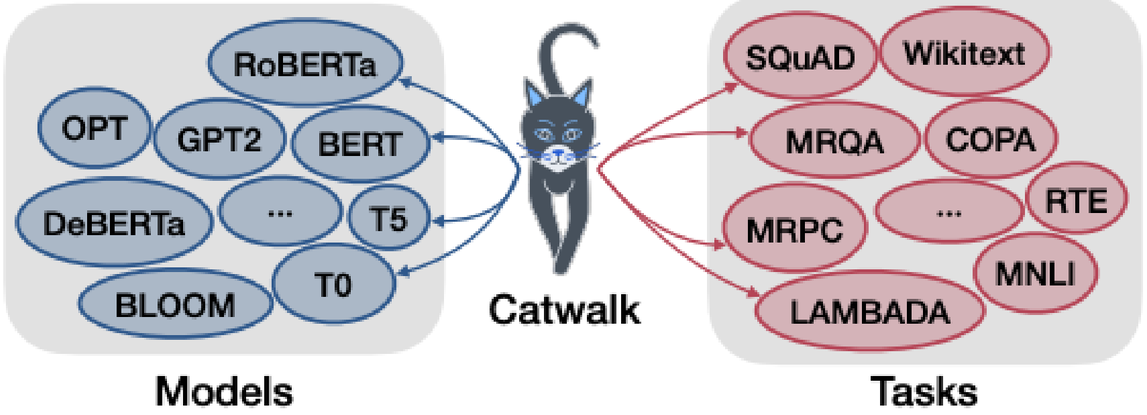
Catwalk: A Unified Language Model Evaluation Framework for Many Datasets

The success of large language models has shifted the evaluation paradigms in natural language processing (NLP). The community's interest has drifted towards comparing NLP models across many tasks,...

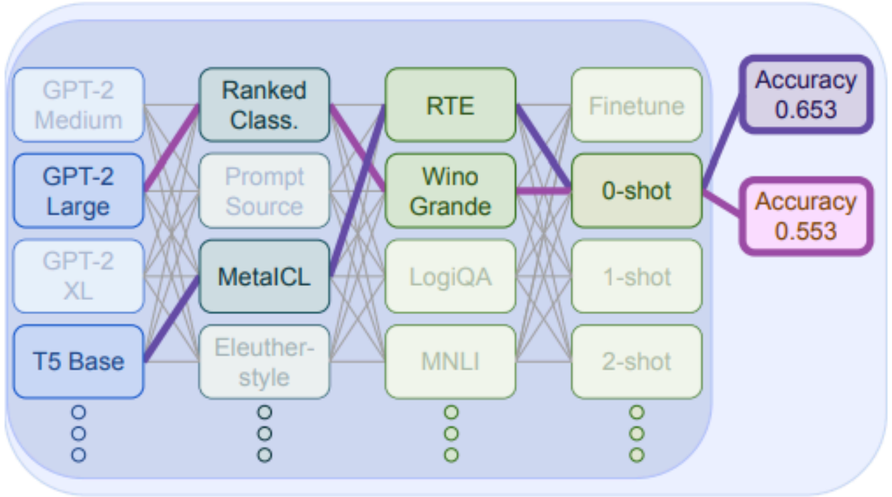
arXiv.org

Catwalk

Catwalk: A Unified Language Model Evaluation Framework for Many Datasets



(a) Catwalk provides a unified interfaces to connect models and datasets. This compares n models on m tasks with only $n + m$ custom implementations, substantially reducing the implementation workload compared to current practice where nm implementations are needed.



(b) Illustration of some decisions a user of Catwalk makes to get a result. For example, choosing GPT-2 Large, with Ranked Classification, on WinoGrande gives an accuracy of 0.553, while choosing T5-Base on RTE gives an accuracy of 0.653.

scale. For example, we finetuned and evaluated over 64 models on over 86 datasets with a single command, *without writing any code*.

Perplexity

- **Measures**: probability model prediction
- **GPT** = predicts probability of next token

- *Evaluates the “uncertainty” an LLM has in predicting the next token*

Paloma

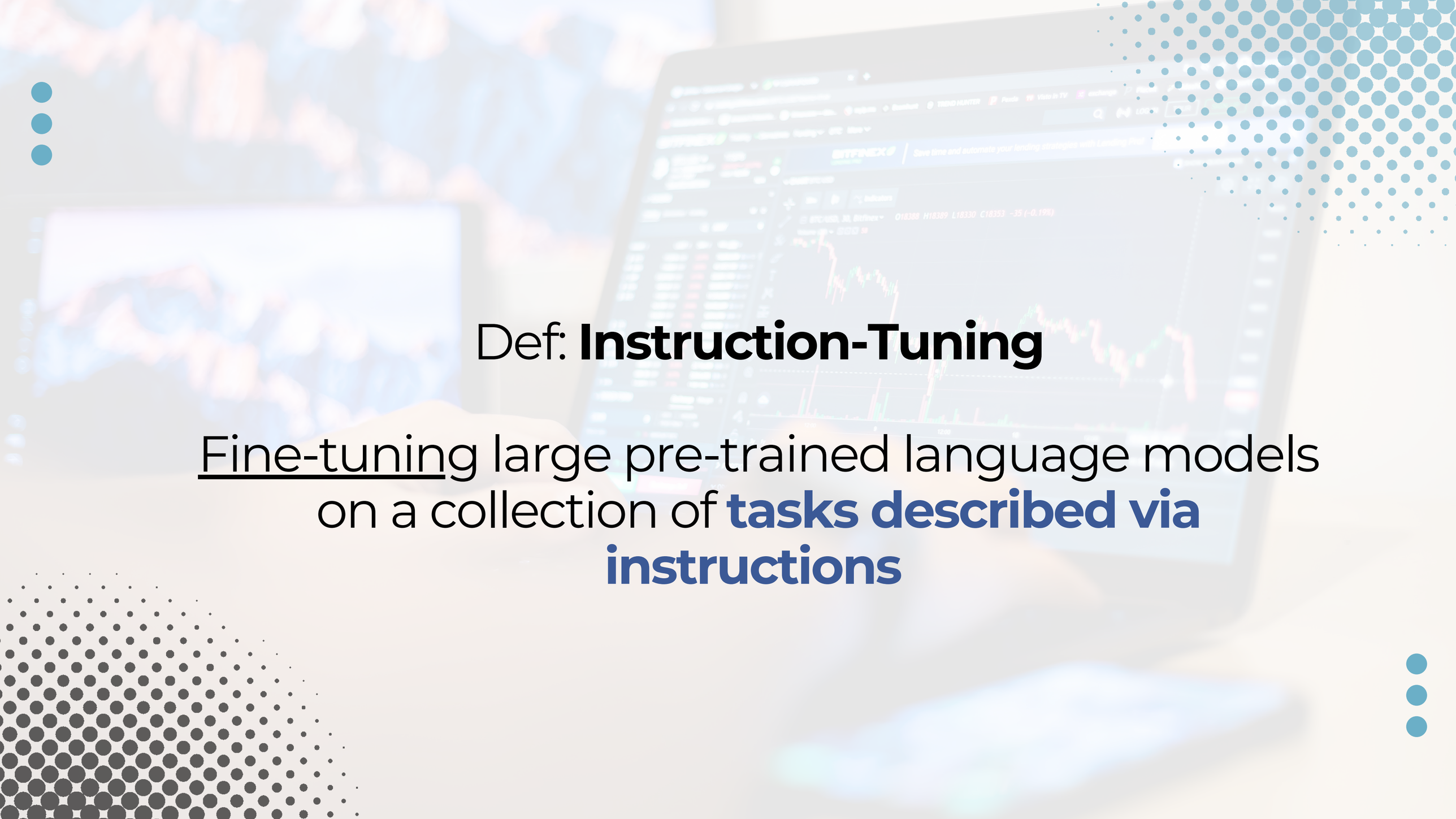
PALOMA: A BENCHMARK FOR EVALUATING LANGUAGE MODEL FIT

Language models (LMs) commonly report perplexity on monolithic data held out from training. Implicitly or explicitly, this data is composed of domains—varying distributions of language. Rather than assuming perplexity on one distribution extrapolates to others, PERPLEXITY ANALYSIS FOR LANGUAGE MODEL ASSESSMENT (PALOMA),[1] measures LM fit to 585 text domains, ranging from *nytimes.com* to *r/depression* on Reddit. We invite submissions to our benchmark and organize results by comparability based on compliance with guidelines such as removal of benchmark contamination from pretraining. Submissions can also record parameter and training token count to make comparisons of Pareto efficiency for performance as a function of these measures of cost. We populate our benchmark with results from 6 baselines pretrained on popular corpora. In case studies, we demonstrate analyses that are possible with PALOMA, such as finding that pretraining without data beyond Common Crawl leads to inconsistent fit to many domains.

Paloma: A Benchmark for Evaluating Language Model Fit

Language models (LMs) commonly report perplexity on monolithic data held out from training. Implicitly or explicitly, this data is composed of domains$\unicode{x2013}$varying distributions of...

arXiv.org



Def: **Instruction-Tuning**

Fine-tuning large pre-trained language models on a collection of **tasks described via instructions**

Open-Instruct

How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources

In this work we explore recent advances in instruction-tuning language models on a range of open instruction-following datasets. Despite recent claims that open models can be on par with state-of-the-art proprietary models, these claims are often accompanied by limited evaluation, making it difficult to compare models across the board and determine the utility of various resources. We provide a large set of instruction-tuned models from 6.7B to 65B parameters in size, trained on 12 instruction datasets ranging from manually curated (e.g., OpenAssistant) to synthetic and distilled (e.g., Alpaca) and systematically evaluate them on their factual knowledge, reasoning, multilinguality, coding, safety, and open-ended instruction following abilities through a collection of automatic, model-based, and human-based metrics. We further introduce TÜLU 🐫, our best performing instruction-tuned model suite finetuned on a combination of high-quality open resources.

How Far Can Camels Go? Exploring the State of Instruction Tuning...

In this work we explore recent advances in instruction-tuning language models on a range of open instruction-following datasets. Despite recent claims that open models can be on par with...

arXiv.org

Open-Instruct

Our experiments show that different instruction-tuning datasets can uncover or enhance specific skills, while no single dataset (or combination) provides the best performance across all evaluations. Interestingly, we find that model and human preference-based evaluations fail to reflect differences in model capabilities exposed by benchmark-based evaluations, suggesting the need for the type of systemic evaluation performed in this work. Our evaluations show that the best model in any given evaluation reaches on average 87% of ChatGPT performance, and 73% of GPT-4 performance, suggesting that further investment in building better base models and instruction-tuning data is required to close the gap. We release our instruction-tuned models, including a fully finetuned 65B TÜLU, along with our code, data, and evaluation framework to facilitate future research.[2]

Open-Instruct

Camels in a Changing Climate: Enhancing LM Adaptation with TÜLU 2

Since the release of TÜLU [Wang et al., 2023b], open resources for instruction tuning have developed quickly, from better base models to new finetuning techniques. We test and incorporate a number of these advances into TÜLU, resulting in TÜLU 2, a suite of improved TÜLU models for advancing the understanding and best practices of adapting pretrained language models to downstream tasks and user preferences. Concretely, we release: (1) **TÜLU-V2-mix**, an improved collection of high-quality instruction datasets; (2) **TÜLU 2**, LLAMA-2 models finetuned on the V2 mixture; (3) **TÜLU 2+DPO**, TÜLU 2 models trained with direct preference optimization (DPO), including the largest DPO-trained model to date (**TÜLU 2+DPO 70B**); (4) **CODE TÜLU 2**, CODE LLAMA models finetuned on our V2 mix that outperform CODE LLAMA and its instruction-tuned variant, CODE LLAMA-Instruct. Our evaluation from multiple perspectives shows that the TÜLU 2 suite achieves state-of-the-art performance among open models and matches or exceeds the performance of GPT-3.5-turbo-0301 on several benchmarks. We release all the checkpoints, data, training and evaluation code to facilitate future open efforts on adapting large language models.

Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2

Since the release of TÜLU [Wang et al., 2023b], open resources for instruction tuning have developed quickly, from better base models to new finetuning techniques. We test and incorporate a...

arXiv.org

WHAT'S IN MY BIG DATA?

Tülu v2

Open instruction & RLHF models

olmo : Accelerating the Science of Language Mo

dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research

Catwalk: A Unified Language Model Evaluation Framework for Many Datasets

paloma : A BENCHMARK FOR EVALUATING LANGUAGE MODEL FIT

Today’s Build

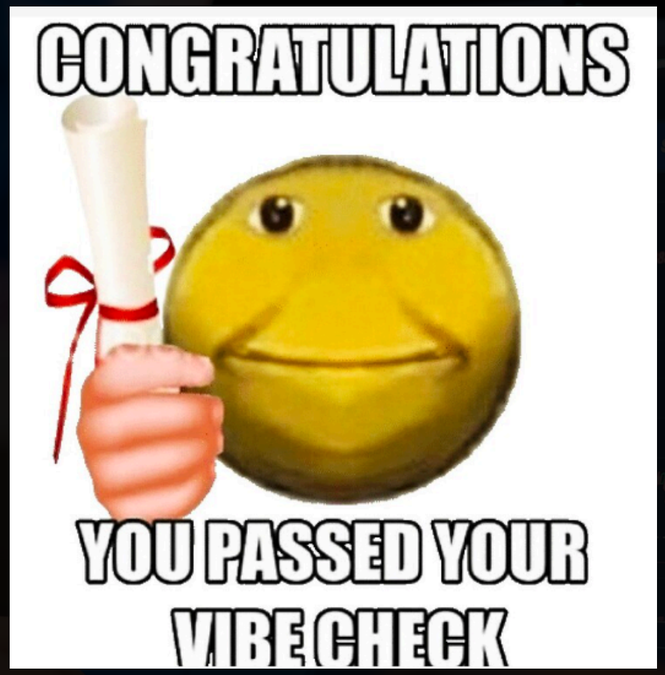
MODEL & DATA

Base LLM

- OLMo-1B

Dataset

- ~~open-instruct~~
- tulu-v2-sft-mixture



Instruct-Tuning OLMo!

Presented by
Chris Alexiuk, LLM Wizard 🪄

Sequence of Work!

- June 23: How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources
- Oct 23: What's In My Big Data?
- Nov 23: Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2
- Dec 23: Catwalk: A Unified Language Model Evaluation Framework for Many Datasets
- Dec 23: Paloma: A Benchmark for Evaluating Language Model Fit
- Jan 24: Dolma: an Open Corpus of Three Trillion Tokens for Language Model Pretraining Research
- Feb 24: OLMo: Accelerating the Science of Language Models

CONCLUSION

- **AI2** OLMo brings together a ton of stuff!
  - OLMo, Dolma, WIMBD, Catwalk, Paloma, Tulu
- OLMo **Modeling**
  - **1B, 7B** parameters
  - **2T, 2.46T** tokens
- Instruct-tuning with open-instruct is now quite implemented!

QUESTIONS?

Thank you!