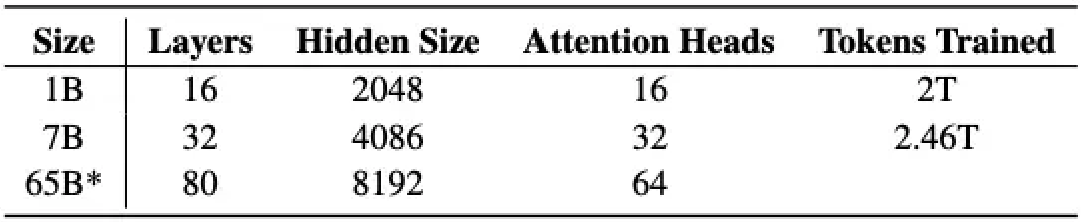
OLMo

In this guide, we provide an overview of the Open Language Mode (OLMo), including prompts and usage examples. The guide also includes tips, applications, limitations, papers, and additional reading materials related to OLMo.

Introduction to OLMo

The Allen Institute of AI has released a new open language model and framework called OLMo. This effort is meant to provide full access to data, training code, models, evaluation code so as to accelerate the study of language models collectively.

Their first release includes four variants at the 7B parameter scale and one model at the 1B scale, all trained on at least 2T tokens. This marks the first of many releases which also includes an upcoming 65B OLMo model.



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

The releases includes:

- full training data, including the code that produces the data
- full models weights, training code, logs, metrics, and inference code
- several checkpoints per model
- evaluation code
- fine-tuning code

All the code, weights, and intermediate checkpoints are released under the Apache 2.0 License.

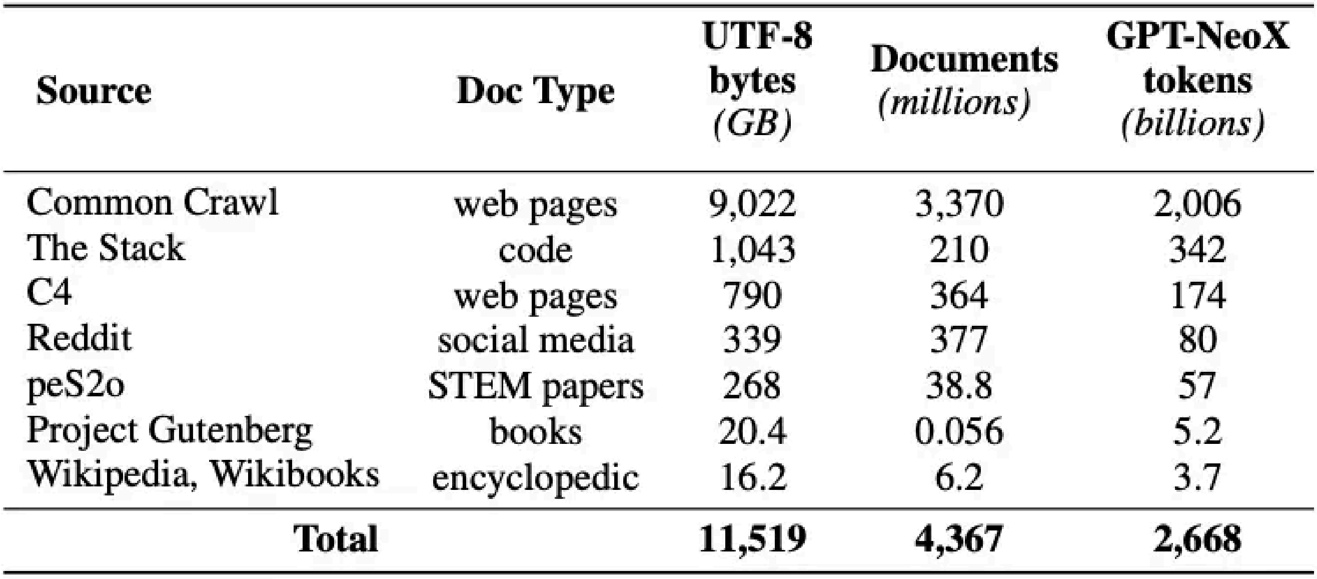
OLMo-7B

Both the OLMo-7B and OLMo-1B models adopt a decoder-only transformer architecture. It follows improvements from other models like PaLM and Llama:

- no biases
- a non-parametric layer norm
- SwiGLU activation function
- Rotary positional embeddings (RoPE)
- a vocabulary of 50,280

Dolma Dataset

This release also includes the release a pre-training dataset called Dolma -- a diverse, multi-source corpus of 3 trillion token across 5B documents acquired from 7 different data sources. The creation of Dolma involves steps like language filtering, quality filtering, content filtering, deduplication, multi-source mixing, and tokenization.



Source	**Doc Type**	**UTF-8 bytes** *(GB)*	**Documents** *(millions)*	**GPT-NeoX tokens** *(billions)*
Common Crawl	web pages	9,022	3,370	2,006
The Stack	code	1,043	210	342
C4	web pages	790	364	174
Reddit	social media	339	377	80
peS2o	STEM papers	268	38.8	57
Project Gutenberg	books	20.4	0.056	5.2
Wikipedia, Wikibooks	encyclopedic	16.2	6.2	3.7
Total		**11,519**	**4,367**	**2,668**

The training dataset includes a 2T-token sample from Dolma. The tokens are concatenated together after appending a special `EOS` token to the end of each document. The training instances include groups of consecutive chunks of 2048 tokens, which are also shuffled.

More training details and hardware specifications to train the models can be found in the paper.

Results

The models are evaluated on downstream tasks using the Catwalk. The OLMo models are compared to other several publicly available models like Falcon and Llama 2. Specifically, the model is evaluated on a set of tasks that aim to measure the model's commonsense reasoning abilities. The downstream evaluation suite includes datasets like `piqa` and `hellaswag`. The authors perform zero-shot evaluation using rank classification (i.e., completions are ranked by likelihood) and accuracy is reported. OLMo-7B outperforms all other models on 2 end-tasks and remains top-3 on 8/9 end-tasks. See a summary of the results in the chart below.

Prompting Guide for OLMo

Coming soon...

Figures source: OLMo: Accelerating the Science of Language Models

References

- OLMo: Open Language Model
- OLMo: Accelerating the Science of Language Models

Last updated on September 19, 2024