

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

Luca Soldaini Rodney Kinney Akshita Bhagia Dustin Schwenk

David Atkinson Russell Authur Ben Bogin ^ω Khyathi Chandu
Jennifer Dumas Yanai Elazar ^ω Valentin Hofmann Ananya Harsh Jha
Sachin Kumar Li Lucy ^β Xinxu Lyu ^ω Nathan Lambert Ian Magnusson
Jacob Morrison Niklas Muennighoff Aakanksha Naik Crystal Nam
Matthew E. Peters ^σ Abhilasha Ravichander Kyle Richardson Zejiang Shen ^τ
Emma Strubell ^χ Nishant Subramani ^χ Oyvind Tafjord Pete Walsh
Luke Zettlemoyer ^ω Noah A. Smith ^ω Hannaneh Hajishirzi ^ω
Iz Beltagy Dirk Groeneveld Jesse Dodge

Kyle Lo

Allen Institute for AI ^βUniversity of California, Berkeley ^χCarnegie Mellon University
^σSpiffy AI ^τMassachusetts Institute of Technology ^ωUniversity of Washington
{lucas,kylel}@allenai.org

Abstract

Information about pretraining corpora used to train the current best-performing language models is seldom discussed: commercial models rarely detail their data, and even open models are often released without accompanying training data or recipes to reproduce them. As a result, it is challenging to conduct and advance scientific research on language modeling, such as understanding how training data impacts model capabilities and limitations. To facilitate scientific research on language model pretraining, we curate and release **Dolma**, a three-trillion-token English corpus, built from a diverse mixture of web content, scientific papers, code, public-domain books, social media, and encyclopedic materials. We extensively document Dolma, including its design principles, details about its construction, and a summary of its contents. We present analyses and experimental results on intermediate states of Dolma to share what we have learned about important data curation practices. Finally, we open-source our data curation toolkit to enable reproduction of our work as well as support further research in large-scale data curation.¹

 hf.co/datasets/allenai/dolma

 github.com/allenai/dolma

1 Introduction

Language models are now central to tackling myriad natural language processing tasks, including few-shot learning, summarization, question answering, and more. Increasingly, the most powerful language models are built by a few organizations who withhold most model development details (Anthropic, 2023; OpenAI, 2023; Anil et al., 2023; Gemini Team et al., 2023). In particular, the composition of language model pretraining data is often vaguely described, even in cases where the model itself is released for public use, such as Llama 2 (Touvron et al., 2023b). This hinders understanding of the effects of pretraining corpus composition on model capabilities and limitations, with impacts on scientific progress as well as on the public who interfaces with these models. Our aim is to increase participation in scientific research of language models through open corpora:

- Data transparency helps developers and users of **applications** that rely on language models to make more informed decisions (Geburu et al., 2021). For example, models have shown to perform better on tasks that are more similar to their pretraining data (Razeghi et al., 2022; Kandpal et al., 2023), or social biases in models’ pretraining data may necessitate additional consideration when using them (Feng et al., 2023; Navigli et al., 2023; Seshadri et al., 2023).

- Open pretraining data is necessary to **analyze** how

¹This manuscript was prepared for **Dolma v. 1.6**. As our work on open data for language modeling continues, we will continue to improve Dolma. Updated versions can be found in the provided links.

[♥]Core authors. See Appendix B for list of contributions.