[Submitted on 10 Oct 2023]

OpenWebMath: An Open Dataset of High-Quality Mathematical Web Text

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, Jimmy Ba

There is growing evidence that pretraining on high quality, carefully thought-out tokens such as code or mathematics plays an important role in improving the reasoning abilities of large language models. For example, Minerva, a PaLM model finetuned on billions of tokens of mathematical documents from arXiv and the web, reported dramatically improved performance on problems that require quantitative reasoning. However, because all known open source web datasets employ preprocessing that does not faithfully preserve mathematical notation, the benefits of large scale training on quantitive web documents are unavailable to the research community. We introduce OpenWebMath, an open dataset inspired by these works containing 14.7B tokens of mathematical webpages from Common Crawl. We describe in detail our method for extracting text and LaTeX content and removing boilerplate from HTML documents, as well as our methods for quality filtering and deduplication. Additionally, we run small-scale experiments by training 1.4B parameter language models on OpenWebMath, showing that models trained on 14.7B tokens of our dataset surpass the performance of models trained on over 20x the amount of general language data. We hope that our dataset, openly released on the Hugging Face Hub, will help spur advances in the reasoning abilities of large language models.

Subjects: **Artificial Intelligence (cs.AI)**; Computation and Language (cs.CL); Machine Learning (cs.LG)

Cite as: arXiv:2310.06786 **[cs.AI]**
(or arXiv:2310.06786v1 **[cs.AI]** for this version)
https://doi.org/10.48550/arXiv.2310.06786

Submission history

From: Keiran Paster [view email]
[v1] Tue, 10 Oct 2023 16:57:28 UTC (1,127 KB)

Bibliographic Tools

Bibliographic and Citation Tools

Bibliographic Explorer (What is the Explorer?)

Litmaps (What is Litmaps?)

scite Smart Citations (What are Smart Citations?)

Code, Data, Media

Demos

Related Papers

About arXivLabs

Which authors of this paper are endorsers? | Disable MathJax (What is MathJax?)

Access Paper:

- Download PDF
- PostScript

(view license, view other formats)

Current browse context:
cs.AI
< prev | next >
new | recent | 2310

Change to browse by:
cs
cs.CL
cs.LG

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

Export BibTeX Citation

Bookmark