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

1. Introduction

The paper introduces Dolma, a large, open-source text corpus containing 3 trillion tokens from various sources including web content, code, books, social media, academic papers, and encyclopedic data. The motivation is to increase transparency and accessibility in language model research, addressing the secrecy around datasets used in commercial LLMs like GPT-4 or PaLM 2. It also aims to provide a free and open-source dataset to train LLMs which is free from abusive or controversial contents and can be used for various use-cases.

2. Related Work

This section highlights the closed nature of data curation in commercial LLMs. In contrast, some open-source models (like T5, BLOOM, GPT-J) share their datasets. Dolma aims to extend this openness by releasing both the data and a toolkit, enabling others to reproduce and analyze the dataset pipeline.

3. Data Design Goals

Dolma is designed with four key goals:

- Match prior language modeling recipes for comparability
- Make evidence-based decisions when best practices are unclear
- Scale to large data volumes for training large models
- Maintain openness by releasing data and clearly documenting curation processes

4. Data Curation Methodology

The Dolma Toolkit helps transform raw multilingual data into a cleaned, English-only, pretraining-ready format. It includes filtering (language, quality, PII, toxicity) and mixing operations (deduplication, decontamination). Filtering is parallelized and optimized for massive datasets, and Bloom filters are used for efficient duplication checks.

5. Curating Dolma-Web (Common Crawl)

This section explains how 2.28T tokens from Common Crawl were processed:

- Acquired from 25 web snapshots (2020–2023)
- Applied URL deduplication, document filtering, content filtering, and paragraph-level deduplication
- Applied quality filters (e.g., high perplexity or non-English) and removed PII and toxic content

6. Curating Dolma-Code (GitHub)

The code subset (411B tokens) comes from GitHub via “The Stack” and underwent:

- Language filtering (non-source files like JSON/CSV removed)
- Quality filtering using heuristics from RedPajama and StarCoder
- Content filtering (e.g., removing secrets using detect-secrets)

- Deduplication using MinHash and LSH

7. Curating Dolma-Social (Reddit)

This subset (89B tokens) was created from 378M Reddit posts/comments using:

- Vote thresholds and NSFW subreddit filtering
- Format variants tested via data ablations
- PII filtering and aggressive deduplication to remove low-quality or repeated content

8. Assembling Other Data Sources

Additional curated sources include:

- C4: Web texts used in T5
- Semantic Scholar: 40M academic papers from S2ORC
- Project Gutenberg: Public domain books
- Wikipedia/Wikibooks: English and Simple English editions

9. Training a Language Model on Dolma

To validate Dolma, the authors trained OLMo-1B, a 1.2B parameter decoder-only model. It showed competitive performance against other 1B models like TinyLlama and Pythia, particularly on reasoning benchmarks. Fine-tuning on program-aided outputs improved performance on tough tasks like GSM8k.

10. Data Ablation Studies

Researchers tested the effects of filtering and mixing decisions. Quality filters, PII removal, and content cleaning significantly improved downstream task performance. Including code in the pretraining mix enhances reasoning skills in language models, especially in multi-step logic tasks.

11. Ethical and Legal Considerations

To minimize harm:

- PII and toxic content are filtered
- Legal experts were consulted throughout
- Copyright concerns are addressed; permissively licensed/open-access sources were preferred. The team acknowledges legal uncertainties and commits to evolving ethical standards

12. Conclusion

Dolma offers a transparent, diverse, and massive dataset for language model research, supporting openness and reproducibility. Its tools and documentation serve as a model for future dataset releases. The team hopes Dolma will empower the community to scrutinize and improve LLM development.

13. References

[1] Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research

Summary By:
Kartikey Vijayakumar Hebbar
Northeastern University, Boston
Intro to AI Agents - Summer 2025