

---

# Dolma: An Open Corpus of 3 Trillion Tokens for Language Model Pretraining Research

---

Luca Soldaini<sup>α</sup> Akshita Bhagia<sup>α</sup> Rodney Kinney<sup>α</sup> Dustin Schwenk<sup>α</sup>

Russell Authur<sup>α</sup> Khyathi Chandu<sup>α</sup> Li Lucy<sup>β</sup> Xinxin Lyu<sup>α</sup> Ian Magnusson<sup>α</sup>  
Aakanksha Naik<sup>α</sup> Matthew E. Peters<sup>α</sup> Abhilasha Ravichander<sup>α</sup> Zejiang Shen<sup>τ</sup>  
Emma Strubell<sup>χ,α</sup> Nishant Subramani<sup>χ,α</sup> Oyvind Tafjord<sup>α</sup> Evan Pete Walsh<sup>α</sup>

Hannaneh Hajishirzi<sup>α,ω</sup> Noah A. Smith<sup>α,ω</sup> Luke Zettlemoyer<sup>ω</sup>  
Iz Beltagy<sup>α</sup> Jesse Dodge<sup>α</sup> Dirk Groeneveld<sup>α</sup>

Kyle Lo<sup>α</sup>

<sup>α</sup>Allen Institute for AI <sup>β</sup>University of California, Berkeley <sup>χ</sup>Carnegie Mellon University  
<sup>τ</sup>Massachusetts Institute of Technology <sup>ω</sup>University of Washington

{lucas,kylel}@allenai.org

## Information


Manuscript version: 0.1  
Corpus version: 1.0


This document is a **preliminary version** of the manuscript for Dolma.  
In its current version, it contains a data sheet [10] for this corpus.

## Abstract

This manuscript contains the data sheet for Dolma, a 3 trillion token corpus from a diverse mix of web content, academic publications, code, books, and encyclopedic materials. Dolma is openly available for download. It is licensed under the AI2 ImpACT license as a medium risk artifact.

 **Dataset** [huggingface.co/datasets/allenai/dolma](https://huggingface.co/datasets/allenai/dolma)

 **Code** [github.com/allenai/dolma](https://github.com/allenai/dolma)

 **License** [allenai.org/impact-license](https://allenai.org/impact-license)

## 1 Data Sheet

### 1.1 Motivation for Dataset Creation

#### Why was the dataset created?

Dolma was created with the primary purpose of training AI2's autoregressive language model OLMo. It is a mixture of documents from multiple data sources. Documents have been transformed using a combination of rule-based and statistical tools to extract textual content, remove layout information, and filter for English content.

Dolma contains data sourced from different domains. In particular, it contains a mixture of text obtained from a web scrape, scientific content extracted from academic PDFs and its associated metadata, code over a variety of programming languages, reference material from Wikipedia and Wikibooks, as well as public domain books from Project Gutenberg.

**What (other) tasks could the dataset be used for?**

We expect this dataset to be useful to train other language models, either in its current form or through further filtering and combining it with other datasets.

Beside language model training, this dataset could be used to study interaction between pretraining corpora and models trained on them. For example, one could study provenance of generations from the model, or perform further corpus analysis.

Specific subset of Dolma could be used to train domain specific models. For example, the code subset could be used to train an AI programming assistant.

**Are there obvious tasks for which it should not be used?**

Dolma is published under the ImpACT license [1] as a “medium-risk artifact”. ImpACT establishes risk-based use restrictions, and requires disclosing creation of derivatives to Allen Institute for AI.

Under no circumstance, this dataset should not be treated as a replacement for any of its original sources.

**Has the dataset been used for any tasks already?**

No model trained on this dataset has been publicly released yet.

**If so, where are the results so others can compare?**

A manuscript is forthcoming.

**Who funded the creation of the dataset?**

All individuals who are responsible for this dataset are employed by the Allen Institute for AI. Similarly, computing resources are provided by AI2.

**If there is an associated grant, provide the grant number.**

Compute for the OLMo project is provided by AMD and CSC, using GPUs on the LUMI supercomputer.

## 1.2 Dataset Composition

**What are the instances? Are there multiple types of instances?**

Instances are plain-text spans on English text or computer code. Each instance was obtained by processing web pages (which might include news, documents, forums, etc), academic articles, computer code from GitHub, encyclopedic content from Wikipedia, or Project Gutenberg books.

**Are relationships between instances made explicit in the data?**

Metadata for subsets of Dolma could be used to reconstruct relationships between items:

- **Common Crawl.** Each document uses the URL of the web page from which it was extracted as its identifier; therefore, it can be used to identify relationships between documents.
- **C4.** The URL of each web page from which documents were extracted is included as metadata; therefore, it can be used to identify relationships between documents.
- **peS2o.** The id of each document is the Semantic Scholar Corpus ID of its corresponding manuscript. Metadata for each manuscript can be obtained using the Semantic Scholar APIs [15].
- **The Stack.** The name of the GitHub repository each document belongs to is included as metadata.
- **Project Gutenberg.** The title of each book is included as the first line of each document.

Subset		Size		
Source	Kind	Gzip files (GB)	Documents (millions)	Tokens (billions)
<b>Common Crawl</b>				
24 shards, 2020-05 to 2023-06	web	4,197	4,600	2,415
<b>C4</b>				
[24]	web	302	364	175
[8]				
<b>peS2o</b>				
[27]	academic	150	38.8	57
<b>The Stack</b>				
[16]	code	675	236	430
<b>Project Gutenberg</b>				
	books	6.6	0.052	4.8
<b>Wikipedia, Wikibooks</b>				
(en, simple)	encyclopedic	5.8	6.1	3.6
<b>Total</b>		<b>5,334</b>	<b>5,245</b>	<b>3,084</b>

Table 1: Composition of Dolma.

- **Wikipedia, Wikibooks.** For both, metadata includes the URL corresponding to the page content was extracted from. Structure and connections between documents can be recovered through the URL.

#### How many instances of each type are there?

Summary statistics are reported in Table 1.

#### What data does each instance consist of? “Raw” data (e.g., unprocessed text or images)? Features/attributes?

For each source, raw data is not available directly but could be recovered using source-specific methods:

- **Common Crawl.** We obtain data from common crawl shards from 2020-05 to 2023-06. WARC files from common crawl can be intersected with Dolma ids to recover original HTML files.
- **C4.** We obtained this corpus from the HuggingFace Hub <sup>1</sup>. In turn, documents in C4 have been derived from a Common Crawl shard for 04/2019. URLs in C4 can be used to recover HTML files.
- **peS2o.** peS2o is derived from S2ORC [18]. Original parsed documents can be obtained from extracting documents in S2ORC that share the same ID with peS2o. Further, metadata in S2ORC can be used to obtain original PDF.
- **The Stack.** The filename and repository name, both available in metadata, can be used to recover original file contents.
- **Project Gutenberg.** The title of each book is the first line of each document.
- **Wikipedia, Wikibooks.** For both, metadata includes the URL corresponding to the page content was extracted from. Structure and connections between documents can be recovered through the URL.

#### Is there a label/target associated with instances? If the instances are related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

<sup>1</sup><https://huggingface.co/datasets/allenai/c4>

There are no labels associated with instances. Many text instances were likely created by people or groups of people, but in the vast majority of cases authorship information is unavailable let alone subpopulation metadata. we leave aggregation and reporting of these statistics to future work.

**Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guarantees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with any of the data?**

The data are derived from the web and the original resources may not persist over time. However, each source represents an archival snapshot of that data that should remain fixed and available:

- **Common Crawl.** The Common Crawl data is available on Amazon S3 as part of the Amazon Web Services' Open Data Sponsorship program and can be freely downloaded<sup>2</sup>. We followed Common Crawl terms of use<sup>3</sup>.
- **C4.** This corpus can be obtained from from the HuggingFace Hub<sup>1</sup> and is released under ODC-By 1.0 [21].
- **peS2o.** peS2o is derived from S2ORC [18]. S2ORC is released through the Semantic Scholar Public API<sup>4</sup> under ODC-By 1.0 [21].
- **The Stack.** The corpus is available on the HuggingFace Hub<sup>5</sup> and consists of code released under a variety of permissive licenses. More details including terms of use for hosting or sharing the corpus are provided in the datacard at the link above.
- **Project Gutenberg.** Project Gutenberg consists of books that are not protected under U.S. copyright law. The corpus is available at [gutenberg.org](http://gutenberg.org).
- **Wikipedia, Wikibooks.** Wikimedia data dumps are freely available<sup>6</sup> and released under CC BY-SA 4.0 license [7].

**Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)**

No. A separate evaluation suite Dolma as been decontaminated against will be released at a later date. Downstream users of this dataset could use any alternative evaluation suite.

**What experiments were initially run on this dataset? Have a summary of those results and, if available, provide the link to a paper with more information here.**

A forthcoming manuscript will detail ablations and other experiments that have been conducted to guide the creation of this dataset.

### 1.3 Data Collection Process

**How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)**

Data acquisition for each subset was performed as follows:

- **Common Crawl.** Shards were downloaded from Common Crawl's official S3 bucket<sup>7</sup> using the `cc_net` pipeline [29]. Data was obtained between March 17<sup>th</sup> and March 27<sup>th</sup>, 2023.
- **C4.** We clone C4 from the HuggingFace Hub<sup>1</sup> using Git with the Git-LFS extension. Repository cloned on May 24<sup>th</sup>, 2023.

---

<sup>2</sup><https://commoncrawl.org/the-data/get-started/>

<sup>3</sup><https://commoncrawl.org/terms-of-use/>

<sup>4</sup><https://www.semanticscholar.org/product/api>

<sup>5</sup><https://huggingface.co/datasets/bigcode/the-stack>

<sup>6</sup><https://dumps.wikimedia.org>

<sup>7</sup><s3://commoncrawl/>

- **peS2o.** We clone peS2o from the HuggingFace Hub<sup>8</sup> using Git with the Git-LFS extension. We use pes2o V2. Repository cloned on June 30<sup>th</sup>, 2023.
- **The Stack.** We clone The Stack from the HuggingFace Hub<sup>5</sup> using Git with the Git-LFS extension. Repository cloned on May 28<sup>th</sup>, 2023.
- **Project Gutenberg.** Data was downloaded directly from [gutenberg.org](http://gutenberg.org). We used GutenbergPy [2] to extract books. Website accessed on April 3<sup>rd</sup>, 2023.
- **Wikipedia, Wikibooks.** Dumps were downloaded from Wikimedia’s website<sup>6</sup>. We use the dump from March 20<sup>th</sup>, 2023.

**Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)**

Data was collected and postprocessed by full-time employees at the Allen Institute for AI. No instances in this dataset are manually annotated.

**Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?**

Please see list above.

**How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?**

Any metadata associated with each instance was obtained directly from each source.

**Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances? If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?**

Sampling for each subset was performed as follows:

- **Common Crawl.** Common Crawl is not a representative sample of the web. Summary statistics about Common Crawl are reported through the `cc-crawl-statistics` [6] project, available at [commoncrawl.github.io/cc-crawl-statistics](https://commoncrawl.github.io/cc-crawl-statistics). Dolma uses Common Crawl shards from 2020-05 to 2023-06<sup>9</sup>.
- **C4.** We use C4 in its entirety.
- **The Stack.** We use The Stack in its entirety.
- **peS2o.** We use pes2o V2 in its entirety.
- **Project Gutenberg.** We process all Gutenberg books.
- **Wikipedia, Wikibooks.** We use the *English* and *Simple* subset of Wikipedia and Wikibooks in their entirety.

**Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?**

Common Crawl is the only source we did not use in its entirety. We use only about a quarter of all shards available. This amount was deemed sufficient for the goal of the OLMo project (train an autoregressive language model with up to 70 billion parameters) given the amount of compute we have available. We decided to use the 24 most recent Common Crawl shards.

<sup>8</sup><https://huggingface.co/datasets/allenai/peS2o>

<sup>9</sup>Common Crawl shards follow naming convention `xxxx-yy`, where `xxxx` is the year the shard was finalized, and `yy` is the week (ranging from 01 to 52).

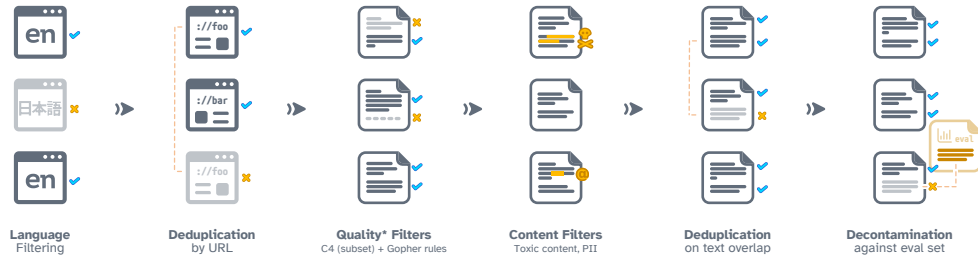


Figure 1: Steps in the web processing pipeline used for the Common Crawl subset of Dolma. The term quality filters is used in accordance with relevant literature<sup>10</sup>.

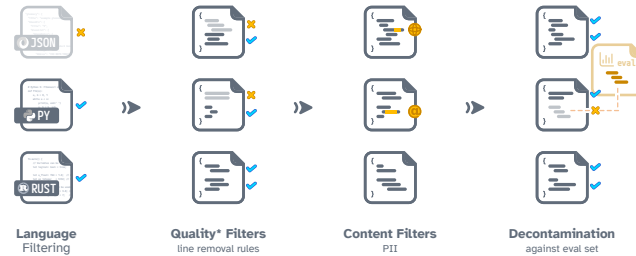


Figure 2: Steps in the web processing pipeline used for the Common Crawl subset of Dolma. The term quality filters is used in accordance with relevant literature<sup>10</sup>.

### Are there any known errors, sources of noise, or redundancies in the data?

Not that we are aware of, although a negligible portion of Common Crawl data could have been lost due to network issues with S3 storage. When accessing Common Crawl, we implemented retry mechanisms, but copy could have failed due to exceeding the retry limits.

## 1.4 Data Preprocessing

**What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)**

All data sources are filtered using FastText language identification models [13, 14] with an English threshold of 0.5.

For the **Common Crawl** subset, we use the following filters (Figure 1) that substantially modify the original data. Note that data might be tagged for removal by one or more filter.

- *As part of the Common Crawl pipeline:* Linearize all HTML into plain text files;
- **Deduplication by URL:** We deduplicate pages by URL (*53% of duplicates removed*);
- **Language identification:** remove all documents with an English score lower than 0.5, as determined by FastText language identification models [13, 14] (*approximately 45% of Common Crawl is in English [6]*);
- **Quality filter**<sup>10</sup>: As in C4 processing, remove lines that do not end up in “.”, “?”, “!”, or “”” (*22.73% of data tagged for removal*);
- **Quality filter**<sup>10</sup>: Remove any line that does not pass any of the Gopher rules [23] (*15.23% of data tagged for removal*);

<sup>10</sup>The term “quality filter”, while widely used in literature, does not appropriately describe the outcome of filtering a dataset. Quality might be perceived as a comment on the informativeness, comprehensiveness, or other characteristics valued by humans. However, the filters used in Dolma and other language models efforts select text according to criteria that are inherently ideological [12].

- **Content filter:** Remove sentences that get ranked as toxic by a FastText classifier (score above 0.4). We train a bigram classifier on the Jigsaw dataset [5] (*1.01% of data tagged for removal*);
- **Content filter:** Mask Personal Identifiable Information (PII) using regular expressions that identify emails, phone numbers, and IP addresses; pages containing 6 or more PII's are completely removed from the corpus (*0.05% tagged for masking, 0.11% tagged for removal*);
- **Deduplication:** We deduplicate the web subset at a paragraph level using a Bloom filter (*19.01% of data tagged for removal*).

For the code subset derived from The Stack, we use the following filters (Figure 2):

- **Language filtering:** Removed the following language files: `assembly`, `csv`, `json`, `json5`, `jsonld`, `jsoniq`, `svg`;
- **Quality Filter:** Removed copyright statements in code files from document preamble<sup>11</sup>;
- **Quality Filter:** Applied RedPijama [28] code filters, including: removal of documents with over 100 characters per line on average, removed documents where a single line contains 1000 characters, removal of documents with fewer than 25% of alphanumeric characters, and removal of documents with a ratio of alphabetical characters to number of tokens below 1.5 (*41.49% of data tagged for removal*);
- **Content filter:** Mask Personal Identifiable Information (PII) using regular expressions that identify emails, phone numbers, and IP addresses; pages containing 6 or more PII's are completely removed from the corpus.

We perform decontamination for all subsets of Dolma. In particular, we remove paragraphs that are shared with documents in our perplexity evaluation suite. Documents in the suite are sampled across several datasets: C4 and mC4 [24, 8], The Pile [9], WikiText 103 [20], Penn Tree Bank [19], S2ORC and Wiki subsets of M2D2 [25], C4 100 domains [4], ICE [11], Twitter AAE [3], Manosphere [26], Gab [30], and 4chan [22]. Overall, only 0.003% of our dataset is removed due to contamination with this evaluation set. A comprehensive manuscript on the design and representativeness of our evaluation suite is to be released at a later date.

**Was the “raw” data saved in addition to the preprocessed/cleaned data? (e.g., to support unanticipated future uses)**

Raw data is available for all subsets except Common Crawl. Due to space constraints, we only keep linearized version of Common Crawl shards, filtered by Language ID as described above.

Raw data is immediately available for download outside Allen Institute for AI. Interested individuals may contact authors of this manuscript if interested in the raw data or establishing a research collaboration.

**Is the preprocessing software available?**

Yes, all preprocessing software is available on GitHub at [github.com/allenai/dolma](https://github.com/allenai/dolma) and on PyPI<sup>12</sup>.

**Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?**

Yes, it does.

**1.5 Dataset Distribution**

**How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)**

Dolma is distributed via the HuggingFace Hub, which offers access via the `datasets` [17] Python package, direct download, and Git using the Git-LFS extension. Additionally, a copy is stored on the cloud storage of the Allen Institute for AI.

<sup>11</sup>Code license and provenance is still tracked in metadata.

<sup>12</sup><https://pypi.org/project/dolma/>

**When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)**

The dataset is available now. This manuscript serves as a reference for the dataset.

**What license (if any) is it distributed under? Are there any copyrights on the data?**

Dolma is published under the ImpACT license [1] as a “medium-risk artifact”. Users must agree to all terms and restrictions of the license before accessing or using the dataset.

**Are there any fees or access/export restrictions?**

The dataset is distributed for free.

The ImpACT license contains restrictions on the distribution of Dolma and any derivatives. Briefly, Dolma cannot be redistributed as-is; however, users are allowed to create derivatives and distributing them. Distribution of derivatives must comply with the following rules<sup>13</sup>:

- **Flow Down Use-Based Restrictions.** The license of derivative must contain Use-Based Restrictions from the ImpACT license for all downstream use and/or further distribution. Use-Based Restrictions should be enforceable through a legal agreement.
- **Attribution.** Derivative must include the applicable attribution notice with your distribution as provided in the legal text of the respective AI2 ImpACT Licenses. This attribution should continue to run downstream.
- **Notices.** Retain all other copyright, IP, and attribution notices that come with the artifact.
- **Derivative Impact Reports.** When creating a derivative, users must submit an impact report using a submission form available from [allenai.org/impact-license](https://allenai.org/impact-license). Your completed Derivative Impact Reports should be published, posted, or otherwise made available to the general public without any requirements or barriers to access.

## 1.6 Dataset Maintenance

**Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?**

The Allen Institute for AI maintains the dataset. For support questions, users are invited to open an issue on GitHub<sup>14</sup> or on the community tab of dataset page<sup>15</sup> (the former being preferred over the latter). Any other inquiry should be sent to [ai2-info@allenai.org](mailto:ai2-info@allenai.org).

**Will the dataset be updated? How often and by whom? How will updates/revisions be documented and communicated (e.g., mailing list, GitHub)? Is there an erratum?**

Dataset will be uploaded on a need-to basis by maintainers at the Allen Institute for AI. Newer version of the dataset will be labeled accordingly. The latest version of the dataset, as well as a changelog, will be made available starting from the first revision.

**If the dataset becomes obsolete how will this be communicated? Is there a repository to link to any/all papers/systems that use this dataset?**

Users should keep track of the version of the dataset in use. In exceptional cases, we might email signatories of the ImpACT license to notify of critical Dolma updates.

Dolma users should cite this manuscript when using this data.

**If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?**

<sup>13</sup>Text presented here is paraphrased from the ImpACT license primer on the website of the Allen Institute for AI: <https://allenai.org/impact-license>

<sup>14</sup><https://github.com/allenai/dolma/issues>

<sup>15</sup><https://huggingface.co/datasets/allenai/dolma/discussions>



Creation and distribution of derivatives is described above. In case contributors want to flow their improvement back to future Dolma releases, they should contact corresponding authors of this manuscript.

## 1.7 Legal & Ethical Considerations

**If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)**

Subsets of Dolma derived from web data are likely created by people or groups of people, however authorship information is often unavailable.

Authors were not directly informed about the data collection. For encyclopedic and web content, logs of web servers will contain records of spiders ran by Common Crawl. For academic content, the peS2o subset [27] is derived from manuscripts that are licensed for permissive distribution by their authors. Finally, the Allen Institute for AI did not contact Project Gutenberg.

**If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)**

Due to the nature of and size of Dolma, it is impossible to determine which obligations, if any, are appropriate.

**If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications) If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?**

The OLMo project includes Ethics committee comprised of internal and external members to the Allen Institute for AI. Plans for the creation of Dolma were reviewed with the committee, and we incorporated their recommendations.

Following practices established in similar efforts, no consent was collected from individuals who might be represented in the dataset. We make available a form<sup>16</sup> for individuals who wish to be removed from the dataset.

**If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?**

Dolma contains text instances that have been derived from web pages Common Crawl crawled from the web. Content might contain sensitive information including personal information, or financial information users of the web chose to put publicly online. This data is taken only from public places, so the same data is or has been accessible via browsing the web. We have measured a variety of types of personal information, and built tools specifically to remove some types of sensitive information, and through our license we restrict what users can do with this data.

We recommend individuals to submit a request using through our form<sup>16</sup> if they wish their information to be removed.

**If it relates to people, does it unfairly advantage or disadvantage a particular social group? In what ways? How was this mitigated?**

Dolma is not a representative sample of none of its sources. It might underrepresent or overrepresent some communities on the internet; further, papers in the peS2o subset are skewed towards STEM disciplines; books in the Gutenberg library are mostly from the public domain (at the time of publication, books published before 1927); finally, the English and Simple subset of Wikipedia and Wikibooks might be biased towards events and people from the global north.

We did not attempt to alter distribution of social groups in Dolma. Large-scale interventions to correct societal biases in large datasets remain challenging, and are left to future work.

<sup>16</sup><https://forms.gle/q4BNUUxUxKwKkfdT6>

**If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?**

This datasets contains text that was derived from web paged scraped by Common Crawl from the web. For much of that data it's not possible identify the authors. In many instances, creators purposely choose to post anonymously online, so aiming to infer authorship can be ethically fraught. We provide access to our data, and encourage any creators that would likely to have data from or about them removed to reach out.

**Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?**

We created this dataset in aggregate, not separately identifying any individual's content or information. We took reasonable steps to remove types of personal information that were possible to reliably detect. We restrict who has access to the data, and we release this under a license that prohibits uses that might be deemed discriminatory. We also provide an avenue for any person to contact us to have text from or about them removed from our corpus<sup>16</sup>.

**Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information) Does the dataset contain information that might be considered inappropriate or offensive?**

This datasets contains text that was derived from web paged scraped by Common Crawl from the web. Therefore, it can contain text posted on public websites by creators on the internet. If an author publicly posted personal information or offensive content, it could be included in this dataset. We took reasonable steps to remove types of personal information that were possible to reliably detect. We also removed documents that contained sentences that were classified as being toxic.

## References

- [1] Allen Insitute for AI. AI2 ImpACT Licenses. <https://allenai.org/impact-license>, 2023. [accessed August 2023].
- [2] Angelescu, Radu. GutenbergPy. <https://github.com/raduangelescu/gutenbergpy>, 2013. Version 0.3.5 [accessed August 2023].
- [3] S. L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas, Nov. 2016. Association for Computational Linguistics.
- [4] A. Chronopoulou, M. Peters, and J. Dodge. Efficient hierarchical domain adaptation for pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1351, Seattle, United States, July 2022. Association for Computational Linguistics.
- [5] cjadams, J. Sorensen, J. Elliott, L. Dixon, M. McDonald, nithum, and W. Cukierski. Toxic comment classification challenge, 2017.
- [6] Common Crawl. cc-crawl-statistics. <https://github.com/commoncrawl/cc-crawl-statistics>, 2016. [accessed August 2023].
- [7] Creative Commons. Attribution-ShareAlike 4.0 International. <https://creativecommons.org/licenses/by-sa/4.0/legalcode>, 2013. [accessed August 2023].
- [8] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, Nov. 2021. Association for Computational Linguistics.

- [9] L. Gao, S. R. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, and C. Leahy. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv*, abs/2101.00027, 2020.
- [10] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021.
- [11] S. Greenbaum. Ice: The international corpus of english. *English Today*, 7(4):3–7, 1991.
- [12] S. Gururangan, D. Card, S. Dreier, E. Gade, L. Wang, Z. Wang, L. Zettlemoyer, and N. A. Smith. Whose language counts as high quality? measuring language ideologies in text data selection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [13] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. Mikolov. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.
- [14] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.
- [15] R. Kinney, C. Anastasiades, R. Authur, I. Beltagy, J. Bragg, A. Buraczynski, I. Cachola, S. Candra, Y. Chandrasekhar, A. Cohan, M. Crawford, D. Downey, J. Dunkelberger, O. Etzioni, R. Evans, S. Feldman, J. Gorney, D. Graham, F. Hu, R. Huff, D. King, S. Kohlmeier, B. Kuehl, M. Langan, D. Lin, H. Liu, K. Lo, J. Lochner, K. MacMillan, T. Murray, C. Newell, S. Rao, S. Rohatgi, P. Sayre, Z. Shen, A. Singh, L. Soldaini, S. Subramanian, A. Tanaka, A. D. Wade, L. Wagner, L. L. Wang, C. Wilhelm, C. Wu, J. Yang, A. Zamarron, M. Van Zuylen, and D. S. W. Weld. The Semantic Scholar Open Data Platform. *arXiv preprint arXiv:2301.10140*, 2023.
- [16] D. Kocetkov, R. Li, L. B. Allal, J. Li, C. Mou, C. M. Ferrandis, Y. Jernite, M. Mitchell, S. Hughes, T. Wolf, et al. The Stack: 3 TB of permissively licensed source code. *arXiv preprint arXiv:2211.15533*, 2022.
- [17] Q. Lhoest, A. Villanova del Moral, P. von Platen, T. Wolf, M. Šaško, Y. Jernite, A. Thakur, L. Tunstall, S. Patil, M. Drame, J. Chaumond, J. Plu, J. Davison, S. Brandeis, V. Sanh, T. Le Scao, K. Canwen Xu, N. Patry, S. Liu, A. McMillan-Major, P. Schmid, S. Gugger, N. Raw, S. Lesage, A. Lozhkov, M. Carrigan, T. Matussièrre, L. von Werra, L. Debut, S. Bekman, and C. Delangue. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184. Association for Computational Linguistics, Nov. 2021.
- [18] K. Lo, L. L. Wang, M. Neumann, R. Kinney, and D. Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics.
- [19] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The Penn Treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.
- [20] S. Merity, C. Xiong, J. Bradbury, and R. Socher. Pointer Sentinel Mixture Models. *arXiv preprint arXiv:1609.07843*, 2016.
- [21] Open Data Commons. Open Data Commons Attribution License (ODC-By) v1.0. <https://opendatacommons.org/licenses/by/1-0/>, 2010. Announcement. [accessed August 2023].
- [22] A. Pappasavva, S. Zannettou, E. De Cristofaro, G. Stringhini, and J. Blackburn. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. *14th International AAAI Conference On Web And Social Media (ICWSM)*, 2020, 2020.

- [23] J. W. Rae, S. Borgeaud, T. Cai, K. Millican, J. Hoffmann, F. Song, J. Aslanides, S. Henderson, R. Ring, S. Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.
- [24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- [25] M. Reid, V. Zhong, S. Gururangan, and L. Zettlemoyer. M2D2: A massively multi-domain language modeling dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 964–975, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.
- [26] M. H. Ribeiro, J. Blackburn, B. Bradlyn, E. De Cristofaro, G. Stringhini, S. Long, S. Greenberg, and S. Zannettou. The evolution of the manosphere across the web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 196–207, 2021.
- [27] L. Soldaini and K. Lo. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI, 2023. ODC-By, <https://github.com/allenai/pes2o>.
- [28] Together Computer. Toxic comment classification challenge, 4 2020.
- [29] G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and É. Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4003–4012, 2020.
- [30] S. Zannettou, B. Bradlyn, E. De Cristofaro, H. Kwak, M. Sirivianos, G. Stringini, and J. Blackburn. What is gab: A bastion of free speech or an alt-right echo chamber. In *Companion Proceedings of the The Web Conference 2018, WWW '18*, page 1007–1014, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.