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

Luca Soldaini [♥α] **Rodney Kinney** [♥α] **Akshita Bhagia** [♥α] **Dustin Schwenk** [♥α]

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

Kyle Lo [♥α]

[α] Allen Institute for AI [β] University of California, Berkeley [χ] Carnegie Mellon University
[σ] Spiffy AI [τ] Massachusetts Institute of Technology [ω] University 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`

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

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 (Gebru 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

[1] This 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.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15725–15788
August 11-16, 2024 ©2024 Association for Computational Linguistics

Source		Doc Type	UTF-8 bytes *(GB)*	Documents *(millions)*	Unicode words *(billions)*	Llama tokens *(billions)*
Common Crawl		web pages	9,812	3,734	1,928	2,479
GitHub		code	1,043	210	260	411
Reddit		social media	339	377	72	89
Semantic Scholar		papers	268	38.8	50	70
Project Gutenberg		books	20.4	0.056	4.0	6.0
Wikipedia, Wikibooks		encyclopedic	16.2	6.2	3.7	4.3
Total			**11,519**	**4,367**	**2,318**	**3,059**

Table 1: The Dolma corpus at-a-glance. It consists of three trillion tokens sampled from a diverse set of domains; sourced from approximately 200 TB of raw text before curation down to an 11 TB dataset. It has been extensively cleaned for language model pretraining use. Tokens calculated using the LLaMA tokenizer.

its composition influences model behavior, allowing those training models to interrogate and improve current data practices (Longpre et al., 2023; Gao, 2021; Elazar et al., 2023). Examples of this research include memorization (Carlini et al., 2022; Chang et al., 2023), deduplication (Lee et al., 2022), adversarial attacks (Wallace et al., 2021), benchmark contamination (Magar and Schwartz, 2022), and training data attribution (Hammoudeh and Lowd, 2022; Grosse et al., 2023).

To support broader participation and inquiry in these lines of research, we present **D**ata for **O**pen **L**anguage **M**odels' **A**ppetite (Dolma), an open corpus of three trillion tokens designed to support language model pretraining research. We source much of our data from sources similar to those present in past work, including a mix of web text from Common Crawl, scientific research from Semantic Scholar, code from GitHub, public domain books, social media posts from Reddit, and encyclopedic materials from Wikipedia. Compared to other publicly-available pretraining corpora, Dolma offers a larger pool of tokens at comparable quality while maintaining diverse data composition. In summary, our contributions are two-fold:

- We release the **Dolma Corpus**, a diverse, **multi-source** collection of 3T tokens[1] across over 4B documents acquired from 6 different data sources that are (*i*) commonly seen in large-scale language model pretraining and (*ii*) made accessible to the general public. Table 1 provides a high-level overview of the amount of data from each source.

- We open source the **Dolma Toolkit**, a high-performance, portable tool designed to efficiently curate large datasets for language model pretraining. Through this toolkit, practitioners can not only reproduce our dataset, but also study and improve data curation practices.

[1]We follow the definition of "token" as a subword obtained using a tokenizer (such as LLaMA's or GPT-NeoX's), which is distinct from "word", as in a unit of text as defined by the Unicode text segmentation standard.

2 Related Work

Closed data curation practices in language model pretraining research. Pretraining data practices for language model research have grown increasingly closed, both with respect to **access** to data as well as **documentation** of key details about the data itself or its curation practices that would enable reproduction efforts or further scientific study. Proprietary models (e.g., GPT-4, OpenAI, 2023; PaLM 2, Anil et al., 2023; Claude, Anthropic, 2023) disclose little to no information (not even corpus size, or data provenance), and do not share data artifacts. Despite increasing access to powerful open models, few are released alongside their training data; exceptions include T5 on C4 (Raffel et al., 2020), BLOOM (Leong et al., 2022) on ROOTS (Piktus et al., 2023), GPT-J (Wang and Komatsuzaki, 2021), GPT-NeoX (Black et al., 2022), Pythia (Biderman et al., 2023) on Pile (Gao et al., 2020), and INCITE (Together Computer, 2023c) on RedPajama v1 (Together Computer, 2023a). The most powerful open models (e.g., Llama 2 (Touvron et al., 2023b), Mistral (Jiang et al., 2023), Yi (Bai et al., 2023), Qwen (01.AI, 2023)) do not share their data nor provide sufficient details for reproduction. Among large-scale language model pretraining efforts, the ones accompanied with transparent data curation documentation include LLaMA (Touvron et al., 2023a) (*released model, unreleased data*), Gopher (Rae et al., 2021) (*unreleased model and data*), and Falcon (Almazrouei et al., 2023) (*released model, released partial data*). Appendix §C further illustrates the many unknowns of data curation practices of open and closed models, as well as recent trends away from open data practices that have motivated our work.

Open corpora for language model pretraining. We recognize prior efforts to curate, document, and release open corpora to support language model pretraining research. However, limitations in these prior corpora have motivated us to curate a new dataset:

- C4 (Raffel et al., 2020) (175B tokens) and Pile (Gao et al., 2020) (387B tokens) are high-quality datasets with demonstrated use in training language models, but are unfortunately **limited in scale**. ROOTS (Piktus et al., 2023) is large (≈400B tokens) but given its multilingual focus, its English-only portion is only 30% of the dataset and thus contributes too few tokens to train English-only models. We recognize that scale and English-only concentration do not imply a "higher-quality" dataset; rather, certain threads of research necessitate these foci, motivating our new corpus (see § 3).

- While Falcon (Almazrouei et al., 2023) (580B tokens) and RedPajama v2 (Together Computer, 2023b) (30T tokens) meet our scale criterion, they are entirely derived from Common Crawl web pages, and thus **lack source diversity** commonly targeted when curating data for the largest language models (e.g., scientific papers, code). We also note that RedPajama v2 is only **lightly-curated**, distributing content output by CCNet (Wenzek et al., 2020) mostly as-is, thus placing the onus on model developers to decide their own filtering before training.

- RedPajama v1 (Together Computer, 2023a) (≈1.2T tokens) is most similar to our effort and a source of inspiration when designing Dolma. While RedPajama v1 was a **specific** reproduction of the LLaMA (Touvron et al., 2023a) training data, we have a **broader** reproduction target which required diving into data sources that RedPajama v1 did not pursue, including larger collections of scientific papers and social media forums like Reddit (see § 3). Further, recent work has identified data quality issues suggesting significant additional cleanup of RedPajama v1 is recommended before costly language model training (Soboleva et al., 2023; Elazar et al., 2023).

While this manuscript was under review, several other open corpora for language modeling have been released, including FineWeb (Penedo et al., 2024), Zyda (Tokpanov et al., 2024), and the datasets used to train LLM360 Amber (Liu et al., 2023), LLM360 K2 (LLM360 Team, 2024), and MAP-Neo (Zhang et al., 2024) models.

3 Data Design Goals

We present the design goals of Dolma and discuss how these goals guided our decision-making during data curation. In sharing these, we hope to inform users of Dolma's strengths and limitations while also reinforcing practice around such disclosures in dataset curation research (see curation rationales in Bender and Friedman (2018) and motivation questions in Gebru et al. (2021)).

Be consistent with prior language model pretraining recipes. By matching data sources and methods used to create other language modeling corpora, to the extent they are known, we enable the broader research community to use our artifacts to study (and scrutinize) language models being developed today, even those developed behind closed doors. In this **reproduction** effort, we follow established practices to the extent they are known. Notably, this also means scoping Dolma to **English-only** text to better leverage known curation practices and maximize generalizability of scientific work on Dolma to existing language models.[2]

When in doubt, make evidence-backed decisions. Still, there remain myriad data curation decisions for which there is no single clear recipe from prior work, both when best practice isn't known as well as when implementations differ in subtle ways. In such cases, we prioritize decisions that **maximize performance** of language models trained on Dolma over a diverse suite of tasks and datasets (see §4.2).

Large scale data to train large models. Hoffmann et al. (2022) suggested that one can train compute-optimal models by maintaining a fixed ratio between language model size (in parameters) and a minimum number of training tokens. Recent works that follow these "scaling laws," such as Llama 2, show that there is still room for performance improvement by increasing the number of training tokens. We aim for a sufficiently large corpus—**2–3T tokens**—to allow further study of the relationship between model and dataset size.

Make necessary adjustments to preserve openness. A core tenet of our work is openness, which we define to mean (*i*) **sharing the data itself** and (*ii*) **documenting the process to curate it**. This requirement means we occasionally must deviate from known recipes due to additional practical, legal or ethical considerations that arise when pursuing dataset research in the open. For example, despite their use in training language models like LLaMA, we avoid sources like Books3 (Gao et al., 2020) which are the center of ongoing legal cases around AI use of copyrighted materials (Knibbs, 2023). Similarly, despite the lack of discussion around the removal of personally identifiable information in prior recipes, we perform this filtering to mitigate risks associated with data release (Subramani et al., 2023).

4 Data Curation Methodology

4.1 The Dolma Toolkit

Pretraining data curation requires defining complex pipelines that transform raw data from multiple sources into a single collection of cleaned, plain text documents (Wenzek et al., 2020; Almazrouei et al., 2023). To curate Dolma, we create and open-source a high-performance toolkit to facilitate efficient processing on

[2]Recognizing that this focus reinforces the assumption of English as the "default" language, we hope to expand Dolma to more languages in the future. We release our data curation tools to support such efforts.

hundreds of terabytes of text content. Our toolkit unifies common dataset curation steps into "filtering" and "mixing" operations:

`filtering` We unify common data transformations like language, quality or content filters into a single implementation. Given a configuration—a text unit (e.g., document, paragraph,[3] sentence, etc.), a scoring method (e.g., linear classifier, language model perplexity, regular expression matches), and a removal policy (e.g., delete, replace with string)—our toolkit parallelizes filtering operations by identifying and removing undesirable text at massive scale. For Dolma, we use these to filter non-English, "low quality" or unnatural,[4] toxicity,[5] and PII at the document and sub-document levels. In internal tests to replicate C4 recipe, our toolkit performed filtering at a rate of 122 CPU hours per TB; for reference, processing the full "raw" Dolma files totaling 200 TB on a `c6a.48xlarge` instance with 192 vCPUs would take 5 days.

`mixing` We unify common cross-file operations, like up/down-sampling, deduplication and decontamination, into a single Rust module that "mixes" content across files into a smaller set of files. For example, we can achieve up-sampling by repeatedly reading the same file paths when mixing. We also implement a Bloom filter (Bloom, 1970) compatible with our mixer which enables linear-time probabilistic detection of duplicates. We can repurpose this for test set decontamination by first seeding the Bloom filter with test examples, then flagging any detected duplicates when mixing the pretraining data.

4.2 Data Ablations

To help us make informed decisions, we conduct **data ablations** in which we train language models on a dataset following a specific data curation decision, or *intervention*, and evaluate the resulting model's performance on a range of test datasets against a *baseline* dataset. By comparing intervention and baseline results while controlling for model architecture and training, we can isolate the impact of specific dataset curation decisions on downstream models.

[3]We define a paragraph to be a span of text ending in a newline `UTF-8` character "`\n`".

[4]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 (Gururangan et al., 2022).

[5]Similar to "quality", there is no single definition for "toxicity". Rather, specific definitions vary depending on task (Vidgen and Derczynski, 2020) and dataset curators' social identities (Santy et al., 2023); annotators' beliefs also influence toxic language detection (Sap et al., 2021). Predicting toxicity remains challenging (Welbl et al., 2021; Markov et al., 2023), especially as existing methods have been shown to discriminate against minoritized groups (Xu et al., 2021).

Model training. We conduct data ablations using a 1.2 billion parameter decoder-only model from the OLMo family of open language models (Groeneveld et al., 2024). This is in line with similar model sizes that have been used for ablations in prior work (Le Scao et al., 2022). As training such models to completion is prohibitively expensive, especially when one must perform these experiments for each significant data curation decision, we only train these models up to 150 billion tokens before terminating them early. Further details of our training setup in Appendix D.1.

Tasks and test datasets. To select our evaluation tasks and datasets, we prioritize those that *(i)* have been used in prior language model pretraining evaluation, *(ii)* capture a diverse range of language model knowledge and capabilities, and *(iii)* for which we can avoid test set contamination (Dodge et al., 2021; Yang et al., 2023). We arrive at **8 datasets** in our evaluation suite (full details in Appendix §D) that have been used in prior language modeling research (e.g., LLaMA, Llama 2, etc.) and capture a range of capabilities (e.g., question answering, commonsense reasoning, etc.). Full test set contamination analysis validating our dataset choices in Appendix §L.

Evaluation. We perform evaluation of our data ablation models using zero-shot in-context prompting, casting every task as (ranked) text classification, following in-context prompt truncation from Min et al. (2022), prompts from PromptSource (Bach et al., 2022), and using an in-house evaluation harness similar to the Eleuther harness (Gao et al., 2023).

5 Curating Dolma-Web

In this section, we describe the web subset of Dolma, which consists of 2.28T tokens derived from **Common Crawl**,[6] a collection of over 250 billion pages that were crawled since 2007. Common Crawl is organized in snapshots, each corresponding to a full crawl over its seed URLs; as of Feb 2024, there are 97 snapshots. We used 25 snapshots between `2020-05` to `2023-06`.[7]

5.1 Acquisition & Language Filtering

Our web pipeline leverages CCNet (Wenzek et al., 2020) to perform language filtering and initial content deduplication. CCNet has been used to develop other language model datasets like that for LLaMA, RedPajama v1, RedPajama v2. CCNet processes each web page with a FastText (Joulin et al., 2016a) language ID model[8] to determine the primary language for each document; we keep all pages with English document score greater than or equal to 0.5 (removed 61.7% of the data, by byte size).

[6]`commoncrawl.org`

[7]To minimize storage and compute costs, we only acquired enough shards of Common Crawl to meet our target 2-3T token corpus size, assuming at least a 10x reduction from the sum of all data cleaning efforts, including CCNet (§3).

[8]`fasttext.cc/docs/en/language-identification`

Further, CCNet identifies and removes very common paragraphs by grouping shards in each snapshot into small sets and removing duplicated paragraphs in each. This step removed approximately 70% of paragraphs, primarily consisting of headers and navigation elements. Overall, CCNet pipeline filters out 84.2% of the content in Common Crawl, from 175.1 TB to 27.7 TB. More details are provided in our Datasheet §N.

5.2 Quality Filtering

Web crawled data requires significant cleanup before language model training; undesirable content ranges from artifacts introduced by HTML to plain text conversion (*e.g.*, page headers, ill-formatted text) to pages lacking "prose-like" content (*e.g.*, boilerplate text, short segments). Per arguments posed in Rae et al. (2021) and Almazrouei et al. (2023) against model-based quality filters, we approach quality filtering by combining heuristics introduced by Gopher and C4. Specifically, we keep all the Gopher rules (`Gopher All`) and keep a single heuristic from C4 designed to remove paragraphs that do not end in punctuation (`C4 NoPunc`), as opposed to adopting the full set of C4 rules (`C4 All`). Implementation details of all filtering rules are provided in our Datasheet §N.

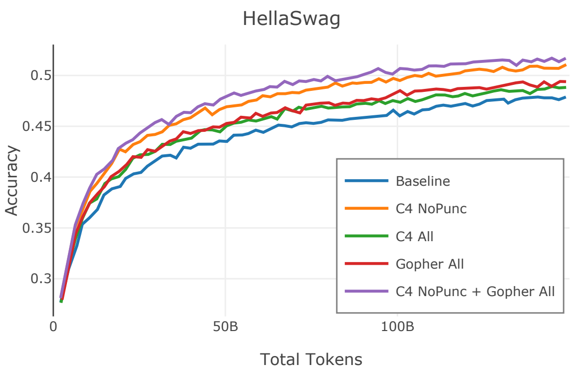


Figure 1: We find a positive effect of web data quality filters on 1.2B model performance, evaluated across training iterations, over a no-filtering baseline. We only show results on HellaSwag here; all figures for other evaluation datasets are in the Appendix §O.

Ablation results shown in §1 validate our filtering strategy: we find that `C4 NoPunc` on its own outperforms both `C4 All` as well as `Gopher All` on both perplexity and downstream tasks. Finally, combining `Gopher All` + `C4 NoPunc` offers the best performance. In all, `Gopher All` tagged 15.23% of `UTF-8` characters for removal, while `C4 NoPunc` tagged 22.73% of characters for removal.

Model and heuristic filters are orthogonal. CCNet also provides quality scores using KenLM (Heafield, 2011) perplexity that groups documents based on Wikipedia-likeness; these buckets are often interpreted as high (21.9%), medium (28.5%), or low (49.6%) quality content, in which more Wikipedia-like is often associated with higher quality. To our surprise, we found our heuristic filtering rules did not affect these proportions, suggesting that such model-based quality filters may capture other signals orthogonal to heuristic filters.

5.3 Content Filtering

Filtering Toxic Content Data sampled from the web often contains harmful or toxic content (Matic et al., 2020; Luccioni and Viviano, 2021; Birhane et al., 2023a,b). Such content is often filtered to minimize the likelihood that downstream language models are prone to toxic content generation (Anil et al., 2023; Rae et al., 2021; Thoppilan et al., 2022; Hoffmann et al., 2022; Longpre et al., 2023). To remove this content from Dolma, we train our own FastText classifiers on the Jigsaw Toxic Comments (cjadams et al., 2017) dataset, producing two models that identify "`hate`" and "`NSFW`" content, respectively. See Appendix §H for implementation details. We run these classifiers on Common Crawl sentences[9] and remove any sentence scored *above* a set threshold.

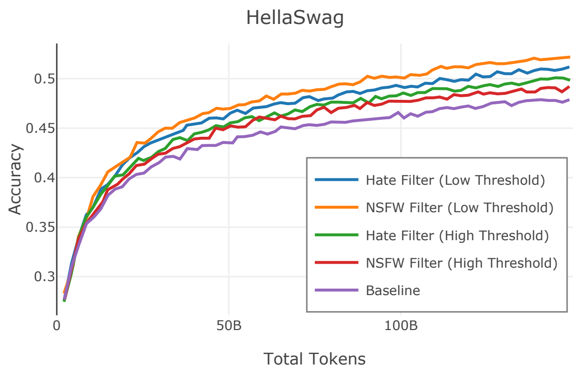


Figure 2: We find a positive effect of web data content filters on 1.2B model performance, evaluated across training iterations, over a no-filtering baseline. We only show results on HellaSwag here; all figures for other evaluation datasets are in the Appendix §O.

To understand filter thresholding effects on Dolma, we conduct a data ablation choosing two very different thresholds for these content filters (§2). We find the "High Threshold" (τ = 0.4) removes *less* content (5.5–7.3%), but generally yields lower downstream performance than the "Low Threshold" (τ = 0.0004) which removes *more* content (29.1–34.9%).[10]

Weighing the tradeoff between dataset scale ("High") and performance maximization ("Low"), we adopt the more permissive "High" threshold to ensure we meet our minimum token count requirement. The cause of this was surprising: Our quality, content, and deduplication filters overlap very little in which texts they remove

[9] Using BlingFire sentence splitter (Microsoft, 2019).

[10] Manual inspection of the distribution of sentence scores revealed a bi-modal distribution with peaks near 0.0 and 1.0 (e.g., Figure 8). As such, we chose "Low" to remove even slightly toxic data (> 0.0), and "High" to limit our max data removal amount to preserve our target dataset scale.

(Figure 9), resulting in a compounded filtering effect when combining them. In future versions of Dolma, we will start with more shards of Common Crawl and adopt stricter filter thresholds.

Filtering Personally Identifiable Information Data sampled from the web can also leak personally identifiable information (PII) of users (Luccioni and Viviano, 2021; Subramani et al., 2023). Traces of PII are abundant in large-scale datasets (Elazar et al., 2023), and language models have also been shown to reproduce PII at inference time (Carlini et al., 2022; Chen et al., 2023b). Dolma's size makes it impractical to use model-based PII detectors like Presidio (Microsoft, 2018); instead, we rely on carefully-crafted regular expressions that sacrifice some accuracy for significant speed-up. Following Subramani et al. (2023), we focus on three kinds of PII that are detectable with high precision: email addresses, IP addresses and phone numbers. For documents with 5 or fewer PII spans, we replace the span with a special token (e.g., `|||EMAIL_ADDRESS|||`); this affects 0.02% of documents. Otherwise, we remove entire documents with higher density of PII spans; this affects 0.001% of documents. In data ablation experiments, we find that execution details around PII (e.g., removal versus special token replacement) had no effect on model performance, which is expected given the tiny percentage of affected data. See Appendix §I for implementation details; all figures for results on evaluation suite are in the Appendix §O.

5.4 Deduplication

Deduplication of pretraining data has been shown to be effective for improving token efficiency during model training (Lee et al., 2022; Abbas et al., 2023; Tirumala et al., 2023); as such, it has become common practice among pretraining data recipes. In Dolma, we perform three stages of deduplication:

(i) **Exact URL dedup** filters 53.2% of documents.

(ii) **Exact document dedup** filters 14.9% of URL-deduped documents, including empty documents.

(iii) **Exact paragraph dedup** filters 18.7% of paragraphs from the URL-deduped documents, including empty paragraphs.

This multi-stage approach is designed to increase efficiency: Stage (i) is commonly used first thanks to its computational efficiency (Agarwal et al., 2009; Koppula et al., 2010; Penedo et al., 2023). Stages (i) and (ii) are designed to remove copies of the same item, such as re-crawls of the same URL and identical pages with multiple URLs (e.g., same news article in multiple online newspapers). Performing these early before any content or quality filtering greatly reduces the number of documents to process. In contrast, Stage (iii) removes common boilerplate content (e.g., the byline under all articles by the same author); as paragraph removal risks disrupting content analysis, we perform it last. We perform all three stages using the Bloom filter in §4.1.

5.5 Putting It All Together

To summarize, the Dolma web pipeline transforms the output of CCNet through URL and document-level deduplication, then quality and content filtering, and finally paragraph-level deduplication.

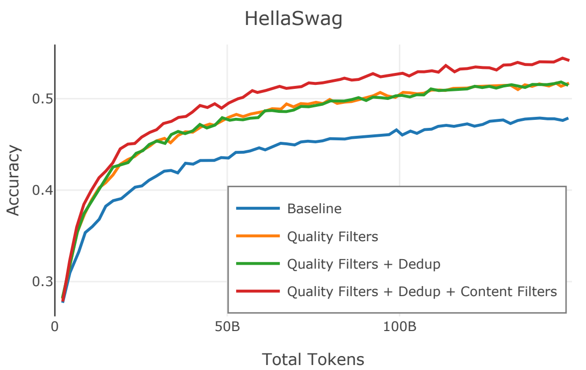


Figure 3: We find a positive compounding effect on 1.2B model performance, evaluated across training iterations, when stacking quality filtering, content filtering and paragraph-level deduplication, over a no-filtering baseline. We show results on HellaSwag here; all figures for other evaluation datasets are in the Appendix §O.

We show the positive compounding effect of all stages of our web pipeline on downstream model performance, as assessed through our data ablations §4.2. We present summary statistics in Appendix §K.

6 </> Curating Dolma-Code

In this section, we describe the code subset of Dolma, which consists of 411B tokens derived from **GitHub**.

6.1 Acquisition & Language Filtering

Like prior work in code language models (e.g., StarCoder (Li et al., 2023b)), we also acquire code through the Stack (Kocetkov et al., 2022), a deduplicated but otherwise unfiltered collection of permissively-licensed GitHub repositories. The raw version of this dataset was collected in March 2023. We filter data-heavy files with extensions such as `JSON` and `CSV`.

6.2 Quality Filtering

We apply heuristics derived from the code subset of RedPajama v1 and StarCoder. RedPajama v1 uses rules to remove repetitive file preambles, such as license statements and documents with excessively long lines or mostly numerical content. Overall, RedPajama v1 is removes files that are mostly data or generated through templates. To select high-quality code snippets, we also use rules from the StarCoder pipeline; these heuristics filter GitHub repositories with no to few stars, files with too few or too many comments, and HTML files with low code-to-text ratio. Implementation details of all filtering rules are provided in our Datasheet §N.

When conducting data ablations, we find that, compared to RedPajama v1 rules alone, RedPajama v1 and

StarCoder rules combined lead to lower perplexity on code datasets (*e.g.*, HumanEval; Chen et al., 2021) and improved performance on datasets in our evaluation suite.[11] Therefore, we chose to use this combination of the two filtering rules for this Dolma code subset.

6.3 Content Filtering

We apply the same heuristics to filter and mask PII used in the web subset (§5). Additionally, we filter any documents containing code secrets and software-specific personal information by running the `detect-secrets` library (Yelp, 2013) and removing any documents with a match.

6.4 Deduplication

We started from the already-deduplicated version of the Stack, which used the pipeline first introduced by Allal et al. (2023), which uses MinHash (Broder, 2002) and Locally Sensitive Hashing to find similar documents.

7 Curating Dolma-Social

In this section, we describe the social media subset of Dolma, which consists of 80B tokens derived from **Reddit** data.

7.1 Acquisition & Language Filtering

We derive this subset from 378M posts from December 2005 until March 2023 obtained through Pushshift (Baumgartner et al., 2020). We include both *submissions*—initial message in conversations on Reddit—and *comments*—replies to messages. The tree-like structure of Reddit threads allows for multiple possible data formats depending on how the various components of a thread are linearized for language model pretraining. To better inform this transformation, we conduct a data ablation over several approaches:

1. **Atomic Content**. Treats all comments and submissionas independent documents.

2. **Partial Threads**. Comments from the same thread combined into a multi-round dialogue between users. Submissions as separate documents.

3. **Full Threads**. Combines submissions with all child comments into one document.

See Appendix §E for implementation details. From results in Figure 4, we see treating submissions and comments as independent documents (Atomic Content) leads to better performance on our evaluation suite. We hypothesize that artificial formatting introduced when combining thread elements negatively impacts language model training; we leave further investigation to future work. Finally, we filter non-English content using the approach from §5.1.

[11] All figures for results on evaluation suite in Appendix §O.

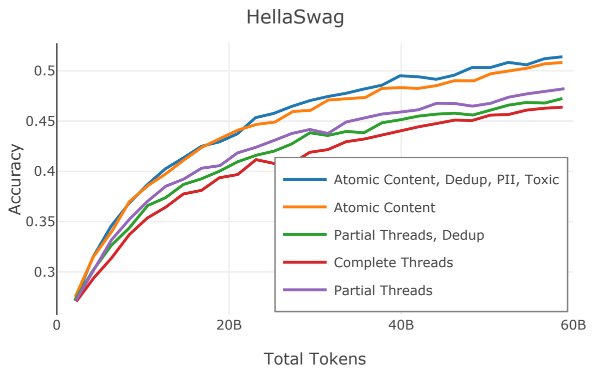


Figure 4: Experimenting with different Reddit thread linearization methods with 1.2B models, evaluated across training iterations. We only show results on HellaSwag here; all figures for other evaluation datasets are in the Appendix §O.

7.2 Quality Filtering

Like web crawled data, social media posts also require significant cleanup before language model training. We repurpose the pipeline introduced by Henderson et al. (2019) to filter submissions and comments. We remove comments shorter than 500 characters, and submissions shorter than 400 characters.[12] We also remove documents over 40,000 characters.

We remove comments with fewer than 3 votes[13], as lower scores are more likely for comments that are deeply nested in a conversational thread (Weninger et al., 2013) or content that is more likely to results in emotionally-charged discourse (Davis and Graham, 2021). Votes have been used as a signal in constructing the WebText (Radford et al., 2019) and OpenWebText (Peterson, 2020) corpora. We discard documents that have been deleted by their authors, removed by moderators, or labeled by their authors as *"over 18"*. We exclude any document originated from a set 26,123 banned or NSFW subreddits.[14]

7.3 Content Filtering

We apply the same content filtering in §5.3, except due to the short length of many Reddit documents, instead of masking PII, we fully remove the document.

7.4 Deduplication

We employ the same strategy used in the web pipeline (§5.4). Since submissions and comments are shorter than web documents, we only deduplicate at a

[12] Qualitative inspection of the data suggested that submissions are of higher quality than comments; thus, we use a more permissive minimum length.

[13] The total votes for each document are obtained by computing the difference between positive votes, also known as "upvotes", negative votes or "downvotes".

[14] Available on GitHub as part of Dolma Toolkit (see `subreddit_blocklist.txt`). The list was curated by merging several sources that tracked banned subreddits. We also include any subreddit with over 10% of posts tagged as NSFW.

document-level. This strategy is useful to reduce the incidence of "*copypasta*" (identical text repeated across comments and subreddits for comedic effect) and other repetitive information.

8 Assembling Other Data Sources

In this section, we briefly summarize additional high-quality sources that were used to derive Dolma. More details on collection and processing in Datasheet §N.

C4 for Curated Web Content Similar to data recipes for LLaMA and Llama 2, we supplement our web subset with C4 (Raffel et al., 2020). We further refine this data by reprocessing it through our full web pipeline (excluding URL deduplication) (§5) which removed additional content, including more low-quality and duplicated texts, and performed PII masking.

Semantic Scholar for Academic Literature The peS2o dataset (Soldaini and Lo, 2023) is a collection of approximately 40 million open-access academic papers that have been cleaned, filtered, deduplicated, and formatted for pretraining language models. It is derived from the Semantic Scholar Open Research Corpus (S2ORC; Lo et al., 2020). As this dataset has been created for language modeling purposes, we use it as-is.

Project Gutenberg for Books Project Gutenberg is a repository of over 70 thousand public domain books. We collected Project Gutenberg's archive in April 2023. We use English language books, which we filter using the same approach described in §5.1. We deduplicate this dataset based on book title exact match.

Wikipedia and Wikibooks for Encyclopedic Content This dataset was derived by March 2023 Wikimedia dumps. We use the "English" and "Simple" editions of Wikipedia and Wikibooks as base for the Encyclopedic subset of Dolma. Sources were processed using WikiExtractor(Attardi, 2023). We remove any document with 25 or fewer `UTF-8`-segmented words, as we found shorter pages to either be the result of short, templated pages (*e.g.*, pages containing only a few words and an information box) or XML parsing errors. By design, this dataset does not contain duplicated documents.

9 Training a Language Model on Dolma

As a final validation step of the Dolma pipeline, we train, evaluate and release a decoder-only, autoregressive language model which we call `OLMo-1B`. We present zero-shot experimental results of `OLMo-1B` on a range of downstream tasks demonstrating comparable quality to other released language models of comparable size.

9.1 Evaluating `OLMo-1B`

In Table 2 we compare `OLMo-1B` with other 1B models. We note that, while all models share a roughly comparable number of parameters, only TinyLlama was trained on roughly the same number of tokens as `OLMo-1B`. Pythia was trained on nearly 10 times fewer tokens and StableLM$_2$ was trained on 2 trillion tokens for two epochs (data composition not shared). Nevertheless, we find that `OLMo-1B` performs better on average than the most comparable model, TinyLlama, outperforming it in 4 out of 8 tasks from our evaluation suite §4.2. Though zero-shot evaluations of such tasks are often challenging for smaller 1B models, we see that performance across all tasks and models is above naive random performance.

Task	*StableLM$_2$ (1.6B)*	*Pythia (1.1B)*	*TinyLlama (1.1B)*	*OLMo-1B (1.2B)*
ARC-E	63.7	50.2	53.2	58.1
ARC-C	43.8	33.1	34.8	34.5
BoolQ	76.6	61.8	64.6	60.7
HellaSwag	68.2	44.7	58.7	62.5
OpenBookQA	45.8	37.8	43.6	46.4
PIQA	74.0	69.1	71.1	73.7
SciQ	94.7	86.0	90.5	88.1
WinoGrande	64.9	53.3	58.9	58.9
Average	***66.5***	**54.5**	**59.4**	**60.3**

Table 2: Comparison of `OLMo-1B` and other similarly-sized language models on our evaluation suite.

9.2 Measuring Domain Fit

In §3, we motivated our decision in curating Dolma to cover a diverse set of sources. In this section, we use `OLMo-1B` to assess Dolma's distribution of documents leads to pretrained language models that fit well to diverse textual domains, compared to training on other open corpora. To represent diverse domains, we use Paloma (Magnusson et al., 2023), a stratified collection of hundreds of fine-grained textual sources; thus, training on more diverse datasets should result in models with lower overall perplexity on Paloma. We repeat our data ablation methodology, training 1.2B models on 150B token samples from C4, mC4 (English-only) (Xue et al., 2020), RedPajama v1, RefinedWeb (Almazrouei et al., 2023), Pile, and Dolma.

From the results in Figure 5, we observe the following: (1) The model trained on Pile performs well as it is comprised of many diverse sources, despite its overall smaller scale. (2) Larger multi-source datasets like Dolma and, to a lesser extent, RedPajama v1 yield models with similar coverage of diverse domains to Pile. (3) Finally, training on single-source corpora like C4, mC4 (English-only), and RefinedWeb leads to models with poor fit to diverse domains as indicated by higher average perplexity.

Our controlled perplexity analysis reveals the importance of including non-web data from diverse curated sources. The metric that we use from Paloma surfaces how models fit more heterogeneous data, because it samples marked domains from each source equally rather

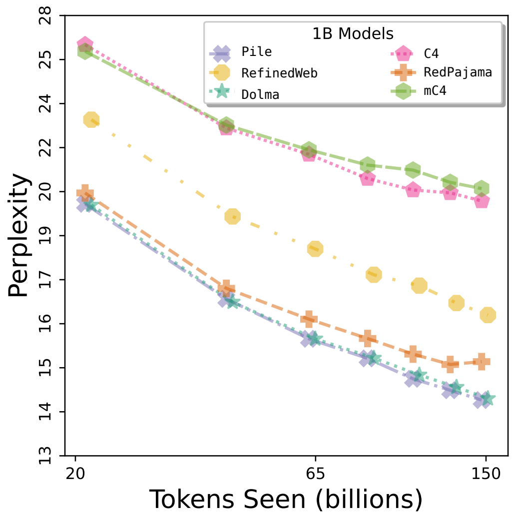


Figure 5: 1.2B parameter language models trained on 150B tokens from Dolma and other open corpora, evaluated across training iterations on perplexity over diverse domains in Paloma (Magnusson et al., 2023).

than by their unequal proportions in the source. Intuitively, the model trained on the Pile is well-fit to such data as that pretraining corpus is mostly sourced from similar smaller, hand-picked sources. But as we wish to scale the total number of tokens in a corpus, the challenge becomes how to integrate more available web data without losing sample efficiency on diverse evaluations such as Paloma. In this case, we see that `OLMo-1B` nearly matches the perplexity curve of the Pile model despite a much larger fraction of web data included.

Conclusion

In this manuscript, we introduce Dolma, a three trillion token English corpus for language model pretraining. The Dolma corpus is comprised of a diverse set of content, including web documents, scientific papers, code, public-domain books, social media, and encyclopedic materials. Building off a list of explicit desiderata, we document our data curation pipelines, providing experimental results that support our decisions. We freely release Dolma and open-source all tools we used to curate this dataset as part of the OLMo project (Groeneveld et al., 2024). Since the time of writing, we have made improvements to Dolma and have continued to make releases; for example, our follow-on release of Dolma `v.1.7` yields significant performance improvement on downstream tasks, holding the model constant.[15] We hope this line of work can promote transparency, reproducibility, and further research in the field of language modeling, as well as address the current gap in the availability of pretraining data of commercial and open language models. We release Dolma under ODC-By and our toolkit under Apache 2.0.

[15] `medium.com/p/92b43f7d269d`

Limitations

English-only corpus. Dolma was curated to contain English data. As tools for language identification may have false negatives, Dolma might contain a small percentage of non-English data. Traces of non-English data are unlikely to lead to any meaningful downstream performance on non-English tasks for any model trained on Dolma. Thus, Dolma reinforces the expectation of English being the "default" language for NLP.

Representativeness of sources in Dolma. As mentioned in §3, it is impossible to curate a corpus that is representative of all language model data curation practices. Further, many open and close language models are trained on content that cannot be acquired or redistributed, and thus could not be included in Dolma.

Single model configuration for ablations. The experimental setup we use to validate our data curation pipeline only covers a subset of model types used to create language models. For example, while many language models are in the 7 billion to 70 billion parameters range, we train 1 billion parameter models; further, we did not investigate the use of any alternative architectures to dense auto-regressive transformer models. This choice was dictated by the need to efficiently iterate over many possible configurations, but it might result in design decisions that are not relevant at larger model sizes. We expect downstream model developers to scrutinize Dolma before using it to train their language models, similar to the process we sketch in §9.

Limited tasks in evaluation suite. As detailed in §4.2, we select tasks that have been used to evaluate previous base language models, and that are not present in our training data (i.e., Dolma is not contaminated against them). As such, we can only assess a subset of tasks language models are routinely used for. For example, the effect of adding code to pretraining data cannot be fully measured until models are able to generate executable code; such capability is typically observed only after models are finetuned to follow instructions (Muennighoff et al., 2023a; Zhuo et al., 2024).

Manual inspection and evaluation of Dolma is infeasible. Given the corpus size, it is impossible to fully inspect Dolma to assess its content. While tools like WIMBD (Elazar et al., 2023) and Data Portraits (Marone and Durme, 2023) aid programmatic inspection of subsets of data, they cannot provide an assessment of all documents in a corpus. As such, we cannot fully describe the properties of Dolma in terms of data distribution, content quality, and potential harms due to the inclusion or exclusion of particular content.

Ethical Considerations

Minimize risk of harm to individuals during data curation. Curating a pretraining corpus may introduce risk to individuals, either by facilitating access to information that is present in the corpus, or by enabling training of harmful models that disclose personal information (Carlini et al., 2020) or produce toxic content (Gehman et al., 2020; Ngo et al., 2021). To minimize these risks while meeting our stated goals, we engaged with legal and ethics experts early in the project and evaluated data design decisions based on their feedback on a case-by-case basis. Broadly, we follow accepted practices when available (*e.g.*, masking of certain personal identifiable information), and take a measured approach when diverging opinions exist in the literature (*e.g.*, most effective approach to identify and remove toxic content). Further, we will provide tools to request data removal[16] We believe in compromising on desired research artifact properties like model reproducibility, performance, and extensibility in cases of significant harm to individuals.

Besides a risk-based approach, alternative frameworks for considering the ethical implications of language model data have also been proposed. Data stewardship (Jernite et al., 2022) seeks to create a framework to collect and reflect explicit interests of data owners. Data trusts (Chan et al., 2023) or data licensing (Li et al., 2023a) can also enable explicit consent in sharing data for AI training. As no current state-of-the-art model is trained on data collected through these frameworks, these approaches would limit the representativeness goal stated in §3. As these principles are adopted, we will consider them for future versions of Dolma.

Copyright and fair use considerations. At the time of writing, the landscape governing applicability of copyright law and fair use doctrine (also known as "fair dealing") and language models is largely undetermined (Cooper et al., 2023; Lee et al., 2024). In the United States, legal scholars and practitioners have suggested that training models on copyright content might constitute fair use (Balasubramaniam et al., 2023; MacKie-Mason and Li, 2023; Henderson et al., 2023), while also recognizing limitations of existing doctrine in this application (Farhadi et al., 2023). Further, legal assessments regarding the use of copyrighted data in language models vary widely depending on jurisdiction: in early 2024, Israel (Israel Ministry of Justice, 2022) and Japan (Technomancers.ai, 2023) allow copyrighted content to be used for AI training data, although the latter is currently re-considering this framework. While most datasets we used were curated with copyright and licensing in mind (e.g., open access papers in peS2o (Soldaini and Lo, 2023), open source repositories in the Stack (Kocetkov et al., 2022)) or were already permissively licensed (e.g., Wikipedia is released under a Creative Commons license), we recognize that large web crawls may also contain copyrighted material. Yet, given current tools, it's not possible to reliably or scalably detect copyrighted materials in a corpus of this size. Our decision to curate and distribute Dolma factors in several considerations, including that all our data sources were publicly available and already being used in large-scale language model pretraining (both open and closed). We recognize that the legal landscape of AI is changing rapidly, especially as it pertains to use of copyrighted materials for training models.

[16] Available at `forms.gle/FzpUXLJhE57JLJ3f8`

References

01.AI. 2023. Yi-34b.

Amro Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *ArXiv*, abs/2303.09540.

Judit Acs. 2019. Exploring BERT's Vocabulary.

Amit Agarwal, Hema Swetha Koppula, Krishna P. Leela, Krishna Prasad Chitrapura, Sachin Garg, Pavan Kumar GM, Chittaranjan Haty, Anirban Roy, and Amit Sasturkar. 2009. Url normalization for de-duplication of web pages. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, page 1987–1990, New York, NY, USA. Association for Computing Machinery.

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models.

Loubna Ben Allal, Raymond Li, Denis Kocetkov, Chenghao Mou, Christopher Akiki, Carlos Munoz Ferrandis, Niklas Muennighoff, Mayank Mishra, Alex Gu, Manan Dey, Logesh Kumar Umapathi, Carolyn Jane Anderson, Yangtian Zi, Joel Lamy Poirier, Hailey Schoelkopf, Sergey Troshin, Dmitry Abulkhanov, Manuel Romero, Michael Lappert, Francesco De Toni, Bernardo García del Río, Qian Liu, Shamik Bose, Urvashi Bhattacharyya, Terry Yue Zhuo, Ian Yu, Paulo Villegas, Marco Zocca, Sourab Mangrulkar, David Lansky, Huu Nguyen, Danish Contractor, Luis Villa, Jia Li, Dzmitry Bahdanau, Yacine Jernite, Sean Hughes, Daniel Fried, Arjun Guha, Harm de Vries, and Leandro von Werra. 2023. SantaCoder: don't reach for the stars! *arXiv [cs.SE]*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance. *TII UAE.*

Angelescu, Radu. 2013. GutenbergPy. `https://github.com/raduangelescu/gutenbergpy`. Version `0.3.5` [accessed August 2023].

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Tachard Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey,

Z. Chen, Eric Chu, J. Clark, Laurent El Shafey, Yanping Huang, Kathleen S. Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan A. Botha, James Bradbury, Siddhartha Brahma, Kevin Michael Brooks, Michele Catasta, Yongzhou Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, C Crépy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, M. C. D'iaz, Nan Du, Ethan Dyer, Vladimir Feinberg, Fan Feng, Vlad Fienber, Markus Freitag, Xavier García, Sebastian Gehrmann, Lucas González, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, An Ren Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wen Hao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Mu-Li Li, Wei Li, Yaguang Li, Jun Yu Li, Hyeontaek Lim, Han Lin, Zhong-Zhong Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alexandra Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Marie Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniela Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Ke Xu, Yunhan Xu, Lin Wu Xue, Pengcheng Yin, Jiahui Yu, Qiaoling Zhang, Steven Zheng, Ce Zheng, Wei Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report. *ArXiv*, abs/2305.10403.

Anthropic. 2023. Introducing Claude. https://www.anthropic.com/index/introducing-claude.

Giuseppe Attardi. 2023. Wikiextractor. https://github.com/attardi/wikiextractor/tree/8f1b434a80608e1e313d38d263ed7c79c9ee75a9. Accessed: 2024-02-15.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. ReFinED: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Alshaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenhang Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, K. Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Yu Bowen, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xing Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *ArXiv*, abs/2309.16609.

Gowri Saini Balasubramaniam, Sara Rachel Benson, Anita Say Chan, Keith Jacobs, Karen V. Jenkins, Smirity Kaushik, Jiaqi Ma, Madelyn Rose Sanfilippo, Eryclis Rodrigues Bezerra Silva, Emmy Tither, Michael Twidale, Ted E. Underwood, Yaman Yu, and Kyrie Zhou. 2023. Comment on docket document (colc-2023-0006-0001): Copyright and artificial intelligence (ai). https://www.regulations.gov/comment/COLC-2023-0006-8998. Posted by the U.S. Copyright Office. See attached file(s).

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *arXiv [cs.SI]*.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Stella Rose Biderman, Hailey Schoelkopf, Quentin G. Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *ArXiv*, abs/2304.01373.

Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. 2023a. Into the laions den: Investigating hate in multimodal datasets. *ArXiv*, abs/2311.03449.

Abeba Birhane, Vinay Uday Prabhu, Sanghyun Han, and Vishnu Naresh Boddeti. 2023b. On hate scaling laws for data-swamps. *ArXiv*, abs/2306.13141.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. PIQA: Reasoning about physical commonsense in natural language. *arXiv [cs.CL]*.

Sid Black, Stella Rose Biderman, Eric Hallahan, Quentin G. Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason

Phang, Michael Martin Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Benqi Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model. *ArXiv*, abs/2204.06745.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. 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. Association for Computational Linguistics.

Burton H Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015b. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

A Z Broder. 2002. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29. IEEE Comput. Soc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2022. Quantifying memorization across neural language models. *arXiv [cs.LG]*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. 2020. Extracting training data from large language models. *arXiv [cs.CR]*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Alan Chan, Herbie Bradley, and Nitarshan Rajkumar. 2023. Reclaiming the digital commons: A public data trust for training data. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 855–868, New York, NY, USA. Association for Computing Machinery.

Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to chatgpt/gpt-4. *ArXiv*, abs/2305.00118.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. 2023a. Symbolic discovery of optimization algorithms.

Yang Chen, Ethan Mendes, Sauvik Das, Wei Xu, and Alan Ritter. 2023b. Can language models be instructed to protect personal information? *arXiv [cs.CL]*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. 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. Association for Computational Linguistics.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168.

Common Crawl. 2016. `cc-crawl-statistics`. `https://github.com/commoncrawl/cc-crawl-statistics`. [accessed August 2023].

A. Feder Cooper, Katherine Lee, James Grimmelmann, Daphne Ippolito, Christopher Callison-Burch, Christopher A. Choquette-Choo, Niloofar Mireshghallah, Miles Brundage, David Mimno, Madiha Zahrah Choksi, Jack M. Balkin, Nicholas Carlini, Christopher De Sa, Jonathan Frankle, Deep Ganguli, Bryant Gipson, Andres Guadamuz, Swee Leng Harris, Abigail Jacobs, Elizabeth E. Joh, Gautam Kamath, Mark A. Lemley, Cass Matthews, Christine McLeavey, Corynne McSherry, Milad Nasr, Paul Ohm, Adam Roberts, Tom Rubin, Pamela Samuelson, Ludwig Schubert, Kristen Vaccaro, Luis Villa, Felix T. Wu, and Elana Zeide. 2023. Report of the 1st Workshop on Generative AI and Law. Available at SSRN: https://ssrn.com/abstract=4634513.

Creative Commons. 2013. Attribution-ShareAlike 4.0 International. `https://creativecommons.org/licenses/by-sa/4.0/legalcode`. [accessed August 2023].

Jenny L Davis and Timothy Graham. 2021. Emotional consequences and attention rewards: the social effects of ratings on reddit. *Information, communication and society*, 24(5):649–666.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. 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. Association for Computational Linguistics.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

Ali Farhadi, David Atkinson, Chris Callison-Burch, Nicole DeCario, Jennifer Dumas, Kyle Lo, Crystal Nam, and Luca Soldaini. 2023. AI2 Response to Notice of Inquiry and Request for Comments.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Leo Gao. 2021. An empirical exploration in quality filtering of text data. *CoRR*, abs/2109.00698.

Leo Gao, Stella Rose Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv*, abs/2101.00027.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association* *for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M R Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo-Yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Yaguang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian Lin, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, Mohammadhossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A Choquette-Choo, Yunjie Li, T J Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen

Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, Xianghai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, M K Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models. *arXiv [cs.CL]*.

Sidney Greenbaum. 1991. Ice: The international corpus of english. *English Today*, 7(4):3–7.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. *arXiv [cs.CL]*.

Roger Baker Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamil.e Lukovsiut.e, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Sam Bowman. 2023. Studying large language model generalization with influence functions.

Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. 2022. 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. Association for Computational Linguistics.

Zayd Hammoudeh and Daniel Lowd. 2022. Training data influence analysis and estimation: A survey. *ArXiv*, abs/2212.04612.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the Workshop on NLP for Conversational AI*. Data available at github.com/PolyAI-LDN/conversational-datasets.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. 2023. Foundation models and fair use. *ArXiv*, abs/2303.15715.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. 2022. Training compute-optimal large language models. *ArXiv*, abs/2203.15556.

Jimin Hong, TaeHee Kim, Hyesu Lim, and Jaegul Choo. 2021. AVocaDo: Strategy for adapting vocabulary to downstream domain. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4692–4700, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Israel Ministry of Justice. 2022. Opinion: Uses of copyrighted materials for machine learning. Accessed: 2024-02-15.

Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Gérard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Isaac Johnson, Dragomir R. Radev, So maieh Nikpoor, Jorg Frohberg, Aaron Gokaslan, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data governance in the age of large-scale data-driven language technology. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15696–15707. PMLR.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld Weld. 2023. The Semantic Scholar Open Data Platform. *arXiv preprint arXiv:2301.10140*.

John Kirk and Gerald Nelson. 2018. The international corpus of english project: A progress report. *World Englishes*.

Kate Knibbs. 2023. The battle over books3 could change ai forever.

Denis Kocetkov, Raymond Li, Loubna Ben Allal, Jia Li, Chenghao Mou, Carlos Muñoz Ferrandis, Yacine Jernite, Margaret Mitchell, Sean Hughes, Thomas Wolf, et al. 2022. The Stack: 3 TB of permissively licensed source code. *arXiv preprint arXiv:2211.15533*.

Hema Swetha Koppula, Krishna P. Leela, Amit Agarwal, Krishna Prasad Chitrapura, Sachin Garg, and Amit Sasturkar. 2010. Learning url patterns for webpage de-duplication. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, page 381–390, New York, NY, USA. Association for Computing Machinery.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Hugo Laurenccon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo Gonz'alez Ponferrada, Huu Nguyen, Jorg Frohberg, Mario vSavsko, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Rose Biderman, Anna Rogers, Loubna Ben Allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, S. Longpre, Sebastian Nagel, Leon Weber, Manuel Sevilla Muñoz, Jian Zhu, Daniel Alexander van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa Etxabe, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Trung Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset. *ArXiv*, abs/2303.03915.

Teven Le Scao, Thomas Wang, Daniel Hesslow, Stas Bekman, M Saiful Bari, Stella Biderman, Hady Elsahar, Niklas Muennighoff, Jason Phang, Ofir Press, Colin Raffel, Victor Sanh, Sheng Shen, Lintang Sutawika, Jaesung Tae, Zheng Xin Yong, Julien Launay, and Iz Beltagy. 2022. What language model to train if you have one million GPU hours? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 765–782, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Katherine Lee, A. Feder Cooper, and James Grimmelmann. 2024. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain. *Journal of the Copyright Society*. Forthcoming.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. 2022. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Quentin Lhoest, Albert Villanova del Moral, Patrick von Platen, Thomas Wolf, Mario Šaško, Yacine Jernite, Abhishek Thakur, Lewis Tunstall, Suraj Patil, Mariama Drame, Julien Chaumond, Julien Plu, Joe Davison, Simon Brandeis, Victor Sanh, Teven Le Scao, Kevin Canwen Xu, Nicolas Patry, Steven Liu, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Nathan Raw, Sylvain Lesage, Anton Lozhkov, Matthew Carrigan, Théo Matussière, Leandro von Werra, Lysandre Debut, Stas Bekman, and Clément Delangue. 2021. 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.

Hanlin Li, Nicholas Vincent, Yacine Jernite, Nick Merrill, Jesse Josua Benjamin, and Alek Tarkowski. 2023a. Can licensing mitigate the negative implications of commercial web scraping? In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '23 Companion, page 553–555, New York, NY, USA. Association for Computing Machinery.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nourhan Fahmy, Urvashi Bhattacharyya, W. Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jana Ebert, Tri Dao, Mayank Mishra, Alexander Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean M. Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023b. Starcoder: may the source be with you! *ArXiv*, abs/2305.06161.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, Richard Fan, Yi Gu, Victor Miller, Yonghao Zhuang, Guowei He, Haonan Li, Fajri Koto, Liping Tang, Nikhil Ranjan, Zhiqiang Shen, Xuguang Ren, Roberto Iriondo, Cun Mu, Zhiting Hu, Mark Schulze, Preslav Nakov, Tim Baldwin, and Eric P. Xing. 2023. Llm360: Towards fully transparent open-source llms.

LLM360 Team. 2024. Llm360 k2-65b: Scaling up open and transparent language models.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

S. Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David M. Mimno, and Daphne Ippolito. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *ArXiv*, abs/2305.13169.

Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 182–189, Online. Association for Computational Linguistics.

Jeffrey MacKie-Mason and Haipeng Li. 2023. Re: Notice of inquiry ("noi") and request for comments, artificial intelligence and copyright, docket no. 2023-6. `https://www.regulations.gov/comment/COLC-2023-0006-8194`. Posted by the U.S. Copyright Office.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1384–1403, Abu

Dhabi, United Arab Emirates. Association for Computational Linguistics.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.

Ian Magnusson, Akshita Bhagia, Valentin Hofmann, Luca Soldaini, Ananya Harsh Jha, Oyvind Tafjord, Dustin Schwenk, Evan Pete Walsh, Yanai Elazar, Kyle Lo, Dirk Groeneveld, Iz Beltagy, Hannaneh Hajishirzi, Noah A Smith, Kyle Richardson, and Jesse Dodge. 2023. Paloma: A benchmark for evaluating language model fit. *arXiv [cs.CL]*.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. A holistic approach to undesired content detection in the real world. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Marc Marone and Benjamin Van Durme. 2023. Data portraits: Recording foundation model training data. *ArXiv*, abs/2303.03919.

Srdjan Matic, Costas Iordanou, Georgios Smaragdakis, and Nikolaos Laoutaris. 2020. Identifying sensitive urls at web-scale. *Proceedings of the ACM Internet Measurement Conference*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. *arXiv preprint arXiv:1609.07843*.

Microsoft. 2018. Presidio - data protection and de-identification sdk.

Microsoft. 2019. Blingfire: A lightning fast Finite State machine and REgular expression manipulation library. https://github.com/microsoft/BlingFire.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv [cs.CL]*.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Niklas Muennighoff, Qian Liu, Armel Zebaze, Qinkai Zheng, Binyuan Hui, Terry Yue Zhuo, Swayam Singh, Xiangru Tang, Leandro Von Werra, and Shayne Longpre. 2023a. Octopack: Instruction tuning code large language models. *arXiv preprint arXiv:2308.07124*.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023b. Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Helen Ngo, Cooper Raterink, João G M Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration.

Ofir Press, Noah A Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation.

Open Data Commons. 2010. Open Data Commons Attribution License (ODC-By) v1.0. https://opendatacommons.org/licenses/by/1-0/. Announcement. [accessed August 2023].

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Antonis Papasavva, Savvas Zannettou, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2020. 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*.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Colin Raffel, Leandro Werra, and Thomas Wolf. 2024. FineWeb: decanting the web for the finest text data at scale. https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra-Aimée Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *ArXiv*, abs/2306.01116.

Joshua Peterson. 2020. openwebtext: Open clone of OpenAI's unreleased WebText dataset scraper. this version uses pushshift.io files instead of the API for speed.

Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The ROOTS search tool: Data transparency for LLMs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 304–314, Toronto, Canada. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nathan McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of Massive Datasets*. Cambridge University Press, USA.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. 2022. 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. Association for Computational Linguistics.

Manoel Horta Ribeiro, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, Summer Long, Stephanie Greenberg, and Savvas Zannettou. 2021. 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.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An adversarial winograd schema challenge at scale. *arXiv [cs.CL]*.

Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. *arXiv [cs.CL]*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv [cs.CL]*.

Teven Le Scao, Angela Fan, Christopher Akiki, Elizabeth-Jane Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, Franccois Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Rose Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu

Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine L. Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad Ali Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla A. Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Olusola Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emily Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Macedo Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, M. K. K. Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zachary Kyle Nguyen, Abhinav Ramesh Kashyap, A. Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, R. Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo L. Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, T. A. Laud, Th'eo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. *ArXiv*, abs/2211.05100.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. 2023. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*.

Noam Shazeer. 2020. GLU variants improve transformer.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Luca Soldaini and Kyle Lo. 2023. peS2o (Pretraining Efficiently on S2ORC) Dataset. `https://github.com/allenai/peS2o`.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti

Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Nishant Subramani, Sasha Luccioni, Jesse Dodge, and Margaret Mitchell. 2023. Detecting personal information in training corpora: an analysis. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 208–220, Toronto, Canada. Association for Computational Linguistics.

Technomancers.ai. 2023. Japan goes all in: Copyright doesn't apply to AI training. Accessed: 2024-2-15.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, Yaguang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language models for dialog applications. *arXiv [cs.CL]*.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. 2023. D4: Improving llm pretraining via document de-duplication and diversification. *ArXiv*, abs/2308.12284.

Together Computer. 2023a. Redpajama-data-1t.

Together Computer. 2023b. Redpajama-data-v2.

Together Computer. 2023c. Redpajama-incite-base-3b-v1.

Yury Tokpanov, Beren Millidge, Paolo Glorioso, Jonathan Pilault, Adam Ibrahim, James Whittington, and Quentin Anthony. 2024. Zyda: A 1.3t dataset for open language modeling.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PloS one*, 15(12):e0243300.

Eric Wallace, Tony Zhao, Shi Feng, and Sameer Singh. 2021. Concealed data poisoning attacks on NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 139–150, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://stability.ai/news/introducing-stable-lm-2`.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv [cs.HC]*.

Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An exploration of discussion threads in social news sites. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, New York, NY, USA. ACM.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv: Artificial Intelligence*.

Albert Xu, Eshaan Pathak, Eric Wallace, Suchin Gururangan, Maarten Sap, and Dan Klein. 2021. Detoxifying language models risks marginalizing minority voices. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2390–2397, Online. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv [cs.CL]*.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023. Rethinking benchmark and contamination for language models with rephrased samples. *ArXiv*, abs/2311.04850.

Yelp. 2013. Detect secrets. `https://github.com/Yelp/detect-secrets`. V1.4.0.

Savvas Zannettou, Barry Bradlyn, Emiliano De Cristofaro, Haewoon Kwak, Michael Sirivianos, Gianluca Stringini, and Jeremy Blackburn. 2018. 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. International World Wide Web Conferences Steering Committee.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, Raven Yuan, Tuney Zheng, Wei Pang, Xinrun Du, Yiming Liang, Yinghao Ma, Yizhi Li, Ziyang Ma, Bill Lin, Emmanouil Benetos, Huan Yang, Junting Zhou, Kaijing Ma, Minghao Liu, Morry Niu, Noah Wang, Quehry Que, Ruibo Liu, Sine Liu, Shawn Guo, Soren Gao, Wangchunshu Zhou, Xinyue Zhang, Yizhi Zhou, Yubo Wang, Yuelin Bai, Yuhan Zhang, Yuxiang Zhang, Zenith Wang, Zhenzhu Yang, Zijian Zhao, Jiajun Zhang, Wanli Ouyang, Wenhao Huang, and Wenhu Chen. 2024. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv: 2405.19327*.

Hao Zhang. 2022. Language model decomposition: Quantifying the dependency and correlation of language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2508–2517, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. 2024. Astraios: Parameter-efficient instruction tuning code large language models. *arXiv preprint arXiv:2401.00788*.

A Acknowledgements

Dolma would not have been possible without the support of many individuals and institutions. The experimental components of this work were made possible through

a partnership with AMD and CSC, enabling use of the LUMI supercomputer. We thank Jonathan Frankle, Cody Blakeney, Matthew Leavitt and Daniel King and the rest of the MosaicML team for sharing findings from experiments on preliminary versions of our data. We thank Vitaliy Chiley for messaging us on Twitter with a suggestion for resolving a random number generator bug that was affecting our data shuffling. We thank Erfan Al-Hossami, Shayne Longpre, and Gregory Yauney for sharing findings from their own large-scale pretraining data experiments. We thank Ce Zhang and Maurice Weber of Together AI for thoughtful discussion on open datasets and data distribution format. We thank Stella Biderman and Aviya Skowron for discussions around data licensing and data processing framework. We thank our teammates at AI2 Nicole DeCario, Matt Latzke, Darrell Plessas, Kelsey MacMillan, Carissa Schoenick, Sam Skjonsberg, and Michael Schmitz for their help with the website, design, internal and external communications, budgeting, and other activities that supported smooth progress on this project. Finally, we also express gratitude for the helpful discussions and feedback from our teammates at AI2 and close collaborators, including Prithviraj (Raj) Ammanabrolu, Maria Antoniak, Chris Callison-Burch, Peter Clark, Pradeep Dasigi, Nicole DeCario, Doug Downey, Ali Farhadi, Suchin Gururangan, Sydney Levine, Maarten Sap, Ludwig Schmidt, Will Smith, Yulia Tsvetkov, and Daniel S. Weld.

B Author Contributions

Dolma would not be possible without the help of our many teammates and collaborators. Weekly project meetings, messaging apps and documentation were accessible for anyone at AI2. Major decisions about Dolma were often made in these channels, with exception for certain topics (e.g., legal, funding). While many were involved in the Dolma effort (see Acknowledgements §A), the authors of this paper were those who owned and delivered a critical piece of the puzzle. We detail their contributions below (authors in alphabetical order):

Contributors to **data acquisition and source-specific data processing** include Akshita Bhagia, Dirk Groeneveld, Rodney Kinney, Kyle Lo, Dustin Schwenk, and Luca Soldaini. Everyone contributed to literature review on available sources and best practices and decisions around sources to pursue. Akshita Bhagia, Rodney Kinney, Dustin Schwenk, and Luca Soldaini handled the bulk of data acquisition and processing and ablation experiments with 1B models for source-specific design decisions. Kyle Lo and Luca Soldaini handled discussions with legal to inform our choice of sources.

Contributors to **infrastructure and tooling** include Russell Authur, Dirk Groeneveld, Rodney Kinney, Kyle Lo, and Luca Soldaini. Rodney Kinney, Kyle Lo, and Luca Soldaini designed and implemented the shared toolkit used for processing our corpus at scale. Dirk Groeneveld wrote the Bloom filter for deduplication and decontamination. Russell Authur wrote a toolkit for acquisition and storage of Common Crawl data.

Contributors to **source-agnostic data processing** include Khyathi Chandu, Yanai Elazar, Rodney Kinney, Kyle Lo, Xinxi Lyu, Ian Magnusson, Aakanksha Naik, Abhilasha Ravichander, Zejiang Shen, and Luca Soldaini. Khyathi Chandu, and Aakanksha Naik developed the toxic text filter. Kyle Lo, and Xinxi Lyu helped evaluate it. Luca Soldaini developed the language filtering approach. Rodney Kinney, Zejiang Shen, and Luca Soldaini developed the "quality" filter. Yanai Elazar identified repeating n-gram sequences. Abhilasha Ravichander, Kyle Lo, and Luca Soldaini developed the PII filter. Jesse Dodge and Ian Magnusson developed the evaluation set decontamination approach.

Contributors to **ablation experiments** include Iz Beltagy, Akshita Bhagia, Jesse Dodge, Dirk Groeneveld, Rodney Kinney, Kyle Lo, Ian Magnusson, Matthew Peters, Kyle Richardson, Dustin Schwenk, Luca Soldaini, Nishant Subramani, Oyvind Tafjord, and Pete Walsh. This work included designing and prioritizing experiments given compute constraints, implementing and running the 1B model experiments, and interpreting results. In particular, Oyvind Tafjord's work on the evaluation toolkit and Pete Walsh's work on the model implementation were critical.

Contributors to **posthoc experiments and analysis** on the final Dolma artifacts. Ben Bogin led the probing experiments on 1B model weights to assess impact of differing code mixtures with support from Kyle Lo and Niklas Muennighoff. Yanai Elazar ran the data analysis tool to summarize and document Dolma's composition. Valentin Hofmann led the tokenization fertility analysis with support from Kyle Lo. Ananya Harsh Jha and Ian Magnusson performed experiments training and evaluating baseline 1B models on other open datasets with support from Luca Soldaini. Sachin Kumar and Jacob Morrison performed analysis of systematic issues in our choice of language identification and toxicity classifiers with support from Kyle Lo. Niklas Muennighoff led analysis of correlation between different filters employed on Common Crawl data with support from Kyle Lo and Luca Soldaini.

Contributors to **licensing and release policy** include David Atkinson, Jesse Dodge, Jennifer Dumas, Nathan Lambert, Kyle Lo, Crystal Nam, and Luca Soldaini. David Atkinson, Jesse Dodge, Jennifer Dumas, and Crystal Nam led the bulk of this, including research into data licenses, risk-level determination for pretraining data, and defining the release policy. Kyle Lo and Luca Soldaini provided feedback throughout this process and handled technical details needed for the release. Nathan Lambert provided feedback on release process and handled the actual release strategy, particularly around external communication.

All of the contributors above helped with **documentation and writing** of their respective components. In particular, Li Lucy provided an extensive literature review of language models, open corpora and pretraining corpus creation practices. Emma Strubell gave valuable feedback on our manuscript. Nathan Lambert helped with feedback on the blog post and other forms of external-facing communication about Dolma.

Hannaneh Hajishirzi, Noah Smith, and Luke Zettlemoyer **advised** on the project, including broad strategy, writing, recruiting and providing resources. As OLMo project leads, Iz Beltagy, Jesse Dodge, and Dirk Groeneveld helped with **visibility and coordination** with other critical OLMo project workstreams. Notably, we credit Noah Smith for coming up with the name Dolma.

Finally, Kyle Lo and Luca Soldaini **led** the overall Dolma project and were involved in all aspects, including project management, planning and design, discussions with legal and ethics committees, data and compute partnerships, infrastructure, tooling, implementation, experiments, writing/documentation, etc.

C (Lack of) details about pretraining data curation for both open and closed language models

We provide a high-level overview of the pretraining data curation practices (or lack of reporting therof) of the largest, most performant language models (in no particular order) to illustrate the need for clear documentation and transparency around dataset curation.

C.1 PaLM 2 (Anil et al., 2023)

Anil et al. (2023) provides limited information on pretraining data used for PaLM 2; we summarize what we could from gather from their manuscript's Sections 3 and D1:

1. **Corpus size**. Unreported other than it's larger than what was used to train PaLM (Chowdhery et al., 2022)

2. **Data provenance**. Unreported other than they use web documents, books, code, mathematics, and conversational data.

3. **PII**. Reported as performed filtering, but without further details.

4. **Toxicity**. Toxic text identified using Perspective API but lacking details needed for reproduction (i.e., text unit, threshold). No details on removal. They did report tackling toxicity through the use of control tokens, but do not provide enough details on this method.

5. **Language ID**. Reports the most frequent languages included as well as their frequencies. Lacking details needed for reproduction (i.e., text unit, tools used, threshold).

6. **Quality**. Reported as performed filtering, but without further details.

7. **Deduplication**. Reported as performed filtering, but without further details.

8. **Decontamination**. N/A.

9. **Other**. Anil et al. (2023) report aggregated statistics of how often certain demographic identities are represented (or not) in the data. Such statistics include identities (e.g., American) or English pronouns. These were identified using tools such as KnowYourData or those available on GoogleCloud, but the manuscript lacks specifics necessary for reproduction.

C.2 GPT-4 (OpenAI, 2023)

OpenAI (2023) provides limited information on pretraining data used for GPT-4; we summarize what we could from gather from their manuscript's Section 2, Appendix C and D, footnotes 5, 6, 10 and 27, and Sections 1.1 and 3.1 in the System Card:

1. **Corpus size**. N/A

2. **Data provenance**. N/A aside from reporting that (1) data was sourced from both the Internet as well as third-party providers, (2) data was sourced mainly before September 2021 with trace amounts of more recent data, and (3) they included GSM-8K (Cobbe et al., 2021) as a tiny fraction of the total pretraining mix.

3. **PII**. N/A.

4. **Toxicity**. Removed documents that violate their usage policies from pretraining, including "erotic content," using a combination of lexicon-based heuristics and bespoke classifiers following Markov et al. (2023).

5. **Language ID**. N/A aside from reporting that the majority of pretraining data is in English.

6. **Quality**. N/A.

7. **Deduplication**. N/A.

8. **Decontamination**. No discussion of decontamination procedures, but instead reported post-hoc statistics measuring extent of contamination on professional and academic exams, as well as several academic benchmarks. Method for identifying contamination based on exact substring match (after removing whitespaces) of a test example against a pretraining data example. They reported some contamination with BIG-Bench (Srivastava et al., 2023).

9. **Other**. There are myraid works performing "data archeology" on GPT-4 that is, attempting to glean information about the pretraining data used in GPT-4 through probes for memorization. For example, Chang et al. (2023) show GPT-4 can generate sequences from copyrighted books. We do not attempt to survey all of these investigative works.

C.3 Claude (Anthropic, 2023)

Unfortunately, we know next to nothing about the pretraining data used for Claude.

C.4 Llama 2 (Touvron et al., 2023b)

Touvron et al. (2023b) provides limited information on pretraining data used for Llama 2; we summarize what we could from gather from their manuscript's Sections 2.1, 4.1, and A.6:

1. **Corpus size**. 2T tokens.

2. **Data provenance**. N/A aside from they avoided using Meta user data.

3. **PII**. Reported as excluded data from certain websites known to contain high volumes of PII, though what these sites are was not disclosed.

4. **Toxicity**. Not explicitly discussed, but appears to not have performed toxicity filtering, opting instead to handle toxic text generation in a later training stage. They do report results from a post hoc analysis in which they used a HateBERT (Caselli et al., 2021) classifier finetuned on ToxiGen (Hartvigsen et al., 2022) to score each document line (and averaged to produce a document-level score).

5. **Language ID**. Not stated as used in pretraining data curation, but they provide a post hoc analysis of the pretraining dataset using FastText Language ID with a 0.5 threshold for detected language. We assume this is likely the same protocol they used for pretraining data curation as it is also seen in the CCNet library (Wenzek et al., 2020), which was used for Llama (Touvron et al., 2023a).

6. **Quality**. N/A.

7. **Deduplication**. N/A.

8. **Decontamination**. They provide extensive reporting on their deduplication method, which relies on a modified version of the ngram deduplication tool from Lee et al. (2022).

9. **Other**. Reported upsampling certain sources, but without further details. They also report a similar analysis as in PaLM 2 (Anil et al., 2023) on aggregate statistics about demographic identities and English pronouns.

C.5 LLaMA (Touvron et al., 2023a)

Touvron et al. (2023a) provides some information on pretraining data used for training LLaMA; we summarize what we could gather from their manuscript's Section 2.1.

1. **Corpus size**. 1.4T tokens.

2. **Data provenance**. LLaMA used data with known provenance, including five shards of CommonCrawl between 2017 and 2020, C4 (Raffel et al., 2020), GitHub code from Google BigQuery public datasets (restricted to Apache, BSD and MIT licenses), Wikipedia dumps from June to August 2022, Project Gutenberg books, Books3 from The Pile (Gao et al., 2020), LaTeX files from arXiv, and StackExchange pages.

3. **PII**. N/A.

4. **Toxicity**. N/A. Reports evaluation on the RealToxicityPrompts (Gehman et al., 2020) benchmark.

5. **Language ID**. Reports use of the CCNet library (Wenzek et al., 2020), which employs FastText (Joulin et al., 2016a) classifiers to remove non-English text (below a 0.5 threshold). No additional language ID reported for C4, GitHub, Books, arXiv, and StackExchange sets. For Wikipedia, reported restriction of pages to those using Latin or Cyrillic scripts: bg, ca, cs, da, de, en, es, fr, hr, hu, it, nl, pl, pt, ro, ru, sl, sr, sv, uk.

6. **Quality**. Reports use of the CCNet library (Wenzek et al., 2020) to remove low-quality content from CommonCrawl; CCNet uses KenLM (Heafield, 2011), an n-gram language model to score perplexity of text as a measure of similarity to Wikipedia text. They do not report their chosen threshold for filtering. They also report use of a linear model trained to classify pages as Wikipedia Reference-like or not. They also report light heuristic filtering of boilerplate content for GitHub and Wikipedia subsets.

7. **Deduplication**. Reports use of the CCNet library (Wenzek et al., 2020) to identify duplicated lines for Common Crawl texts, file-level exact match deduplication for GitHub code, and deduplicating books with over 90% for Gutenberg and Books3 subsets.

8. **Decontamination**. N/A.

9. **Mixture**. The manuscript reports a mixture of 67% CommonCrawl, 15% C4, 4.5% GitHub, 4.5% Wikipedia, 4.5% Books, 2.5% arXiv, and 2.0% StackExchange. Model training was a single epoch over this mixture except for an upsampling of Wikipedia and Books (2 epochs).

C.6 OPT (Zhang, 2022)

From Zhang (2022)'s manuscript and provided datasheet (Gebru et al., 2021), we summarize the following:

The OPT model was trained on **180B tokens** from data sources with known **provenance**: the datasets used for RoBERTa (Liu et al., 2019), a subset of the Pile (Gao et al., 2020), and the Pushshift Reddit Dataset (Baumgartner et al., 2020) as processed by (Roller et al., 2021). They made several notable changes to these sources:

1. *RoBERTa*. Reports updated the CC-News collection up to September 2021.

2. *Pile*. Reports restricted to the following collections: CommonCrawl, DM Mathematics, Project Gutenberg, HackerNews, OpenSubtitles, OpenWebText2, USPTO and Wikipedia. (Zhang, 2022) report omission of other Pile subsets due to gradient norm spikes at the 1B model scale.

3. *Pushshift Reddit*. Reports restricted to only the longest chain of comments in each thread; an operation that reportedly reduced the dataset by 66%.

Also describes: (1) **deduplication** using MinHashLSH (Rajaraman and Ullman, 2011) with a Jaccard similarity threshold of 0.95, and (2) **language ID** filtering to English-only text, though they do not describe the method used.

They do not discuss whether they do (or do not) perform any processing for **PII**, **toxicity**, **quality**, or **decontamination**.

D Experimental Setup

D.1 Ablation Setup

For all data ablations described in this section, we train a 1B parameter model on up to 150B tokens. We follow model architecture and training from OLMo (Groeneveld et al., 2024); we summarize key details here, but direct the reader to the manuscript for further details. Each model is an decoder-only transformer model with 16 layers, 16 attention heads, and 2048 dimensionality. We use ALiBi positional embeddings (Ofir Press et al., 2021), SwiGLU activation (Shazeer, 2020), and mixed precision; model context size is set to 2048 tokens. We use EleutherAI's GPT NeoX tokenizer (Black et al., 2022). The model is trained using the LionW optimizer (Chen et al., 2023a) with 1e-4 peak learning rate, warm-up of 2000 steps, cosine decay, and 1e-2 weight decay. Batch size was set to 1024. While we set our max number of steps to 95k (which is approximately 200B tokens), we conclude our experiments at 150B tokens.

We use 64 AMD Instinct MI250X accelerators. Each MI250X accelerator contains two logical nodes; therefore, from the point of view of our training code, our experiments ran on 128 compute units grouped in 16 nodes. Per each logical unit, we use a micro-batch size of 8. We implement our experiments using the `anonymized` codebase.

D.2 Perplexity Evaluation Suite

For data ablations, we keep track of language model perplexity using Paloma (Magnusson et al., 2023). Datasets included:

- **C4** (Raffel et al., 2020; Dodge et al., 2021): Standard contemporary LM pretraining corpus automatically filtered from the April 2019 Common Crawl scrape.

- **mC4** (Xue et al., 2020); *English subset*: the English language portion of a pretraining corpus automatically filtered from 71 Common Crawl scrapes.

- **Pile** (Gao et al., 2020), *validation set*: widely-used language modeling pretraining corpus; contains documents curated from multiple sources including several non-web sources.

- **WikiText 103** (Merity et al., 2016): a standard collection of verified "Good" and "Featured" articles on Wikipedia.

- **Penn Tree Bank** (Marcus et al., 1994): widely-used NLP corpus derived from Wall Street Journal articles.

- **M2D2** (Reid et al., 2022), *S2ORC subset*: papers from Semantic Scholar (Lo et al., 2020) grouped by hierarchical academic field categories.

- **M2D2** (Reid et al., 2022), *Wiki subset*: Wikipedia articles grouped by hierarchical categories in the Wikipedia ontology

- **C4 100 domains** (Chronopoulou et al., 2022): balanced samples of the top 100 domains in C4.

- **Gab** (Zannettou et al., 2018): data from 2016-2018 from an alt-right, free-speech-oriented social media platform that has been shown to contain more hate speech than mainstream platforms.

- **ICE** (Greenbaum, 1991): English from around the world curated by local experts, with subsets for Canada, East Africa, Hong Kong, India, Ireland, Jamaica, Philippines, Singapore, and the USA.

- **Twitter AAE** (Blodgett et al., 2016): balanced sets of tweets labeled as African American or white-aligned English.

- **Manosphere** (Ribeiro et al., 2021): sample of 9 forums where a set of related masculinist ideologies developed over the past decade.

- **4chan** (Papasavva et al., 2020): data from 2016-2019 politics subsection of an anonymity-focused forum found shown to contain high rates of toxic content.

We also curated held-out sets from other open language model corpora to augment Paloma:

- **Dolma** (this work), *uniform sample*: A sample 8,358 documents from the Dolma corpus across all of its subsets (13 from books, 1,642 from Common Crawl web pages, 4,545 Reddit submissions, 450 scientific articles, 1,708 Wikipedia and Wikibooks entries).

- **RedPajama v1** (Together Computer, 2023b): 1 trillion tokens replication of the LLaMA 1 (Touvron et al., 2023a) pretraining corpus.

- **Falcon RefinedWeb** (Penedo et al., 2023): A corpus of English sampled from all Common Crawl scrapes until June 2023, more aggressively filtered and deduplicated than C4 and mC4-en.
- **Dolma 100 Subreddits** (this work): Balanced samples of the top 100 subreddits by number of posts, sourced from the Dolma Reddit subset.
- **Dolma 100 Programming Languages** (this work): Balanced samples of the top 100 programming languages by number of tokens, sourced from the Dolma Stack subset.

D.3 Downstream Evaluation Suite

We primarily base our data ablation decisions on the performance of models on this evaluation suite:

- **AI2 Reasoning Challenge** (Clark et al., 2018): A science question-answering dataset broken into *easy* and *challenge* subsets. Only the easy subset was used in online evaluations. The challenge subset was, however, included in offline evaluations.
- **BoolQ** (Clark et al., 2019): A reading comprehension dataset consisting of naturally occurring yes/no boolean questions and background contexts.
- **HellaSwag** (Zellers et al., 2019): A multiple-choice question-answering dataset that tests situational understanding and commonsense.
- **OpenBookQA** (Mihaylov et al., 2018): A multiple-choice question-answering dataset modeled on open-book science exams.
- **Physical Interaction: Question Answering (PIQA)** (Bisk et al., 2019): A multiple-choice question-answering dataset that focuses on physical commonsense and naive physics.
- **SciQ** (Welbl et al., 2017): A crowdsourced multiple-choice question-answering dataset consisting of everyday questions about physics, chemistry and biology, among other areas of science.
- **WinoGrande** (Sakaguchi et al., 2019): A dataset of pronoun resolution problems involving various forms of commonsense. Modeled after the Winograd challenge from Levesque et al. (2012).

D.4 Training Setup for `OLMo-1B`

For `OLMo-1B`, we follow the experimental setup outlined for dataset ablation experiments in Appendix D, with the following differences:

- We set the max number of steps to 739,328 (which is roughly 3.1T tokens).
- We double the batch size to 2048 and do so by scaling up to 256 compute units (double what we used for data ablations).
- Due to instabilities we found in the LionW optimizer, we switched to using AdamW.

E Construction of Conversational Threads in Forums Data

Content comes from Reddit's data API in two separate but linked forms: *submissions* and *comments*. *Submissions* are either "link posts" to external content (e.g. news articles, blogs, or even multimedia content) or "self posts" (submissions written by the poster meant to initiate a discussion thread on a topic). *Comments* are user replies to either the initiating post (top level comments) or to another user's comment. Posts, top-level comments, and replies to comments form a nested conversational thread with a submission post at it's root and comments branching out into multiple possible dialogue trees.

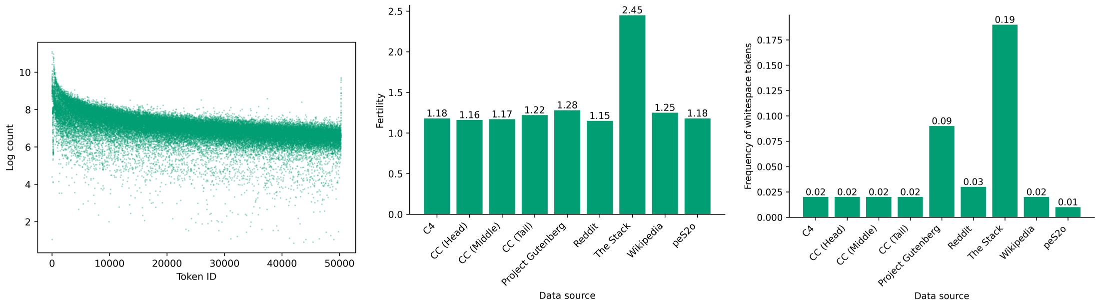
The tree-like structure of Reddit threads allows for multiple possible data formats depending on how the various components of a thread are combined. We investigate three formats for their potential as LM pretraining data:

- **Atomic content**. This simple format treats all comments and submissions as independent documents without any structure or connection to the thread they appear in.
- **Partial threads**. This format assembles comments from the same thread into a structured, multi-round dialogue between users. Submissions are left as separate documents. Assembled dialogues are limited to a maximum parent depth, and the resulting documents are only snippets of a their originating thread (which are spread across several documents).
- **Full threads**. This complex format combines a given submission and all of its child comments into a single document encompassing an entire thread. Code-like indentation is used to indicate the depth of a comment in the thread's hierarchy.

We experimentally evaluated these strategies for assembling documents in Figure 4. We found that, for language modeling purposes, treating comments and submissions as atomic units leads to better downstream performance compared to partial and full threads. We hypothesize that the more complex formatting required to handle dialogues might introduce undesirable content for language modeling, such as short and repeated comments. We leave the study of better formatting for forum content for language modeling to future work.

F Tokenization Analysis

The first step of processing text with LMs is *tokenization*, i.e., mapping the text to a sequence of tokens with corresponding input embeddings (Sennrich et al., 2016; Kudo, 2018; Kudo and Richardson, 2018). Recently, there has been a growing interest in the question of how well LM tokenizers fit different data sources (e.g., data in different languages; Ahia et al., 2023; Petrov et al., 2023) Inspired by this emerging line of work,



(a) Count analysis

(b) Fertility analysis

(c) Whitespace analysis

Figure 6: Tokenization analysis. Tokens with small IDs, which have a high count in the tokenizer training data, also tend to have a high count in Dolma (a). The Stack has a substantially higher fertility compared to the other data sources (b), which can be explained by the higher relative frequency of whitespace characters such "\n" and "\t" (c). See text for more details.

we conduct an explorative analysis of the GPTNeoX tokenizer (Black et al., 2022) applied to Dolma, which provides a first picture of how challenging the different data sources comprised by Dolma are for current LM tokenizers.

We start by taking a global look at the tokenizer's fit to Dolma. Out of the 50,280 tokens in the tokenizer vocabulary, 50,057 are present in the tokenized text of Dolma. In other words, 223 tokens are never used, amounting to roughly 0.4% of the tokenizer vocabulary. The 223 tokens mostly consist of combinations of whitespace characters (e.g., "\n\n ", two newline characters followed by two blank space characters). Note that when training an LM with the examined tokenizer on Dolma, the input embeddings corresponding to these tokens would not be updated. In terms of the count distribution of tokens, we find that tokens with smaller IDs tend to have higher counts in Dolma (see Figure 6a), which is also reflected by a strong Spearman's correlation between (i) the ranking of tokens based on their counts in Dolma and (ii) the token IDs (r = 0.638, $p <$ 0.001). Given how the tokenizer was trained (Sennrich et al., 2016; Black et al., 2022), smaller IDs correspond to byte pairs merged earlier and hence tokens occurring more frequently in the tokenizer training data Overall, these results suggest a good fit of the GPTNeoX tokenizer to Dolma.

Does the tokenizer fit all data sources included in Dolma equally well? To examine this question, we analyze fertility, which is defined as the average number of tokens per word generated by a tokenizer (Acs, 2019; Scao et al., 2022), in our case measured on a specific data source. We find that fertility is similar for most data sources, ranging between 1.15 (conversational forum subset) and 1.28 (books subset), with the exception of the code subset, which has a substantially higher fertility of 2.45 (see Figure 6b). This means that the costs of processing the code subset — be they computational or financial in nature (Petrov et al., 2023) — are more than twice as high compared to the other data sources.

What causes this discrepancy? We find that in the code subset (which mostly contains code), words are often preceded by whitespace characters *other than* a blank space (e.g., newline, tab, return). Crucially, while a blank space before a word is tokenized as part of that word (e.g., *I love you* → "I", " love", " you"), other whitespace characters yield separate tokens (e.g., *I	love	you* → "I", "\t", "love", "\t", "you"). This can also be seen by plotting the relative frequency of tokens representing whitespace characters by data source, which is one order of magnitude higher for The Stack compared to most other data sources (see Figure 6c). When training LMs on The Stack (or code more generally), it thus might be advisable to add special tokens to the tokenizer (e.g., "\nif"; Hong et al., 2021). It is important to notice that this observation applies to most tokenizers in use today (e.g., the tokenizer used by GPT-4), which tend to lack tokens such as "\nif".

G Auditing our Language Filter

To analyze the impact of the FastText language identification classifier, we ran an external audit on the International Corpus of English (ICE) (Kirk and Nelson, 2018), a dataset containing spoken and written English from nine countries around the world. We ran our language ID tool on all documents in the ICE dataset to estimate how many documents from each region would have been erroneously filtered. The ground truth in this analysis is that every document is in English, and should be classified as such. Interestingly, we found that at our fairly permissive threshold (keeping documents with at least a 0.5 score for English) correctly identified all English-language documents in ICE each as English, no matter the region it was from.

H Details on Toxicity Filters

Implementation. To remove toxic content from Dolma, we used the Jigsaw Toxic Comments dataset (cjadams et al., 2017), which contains forum comments tagged with (multilabel) categories "`toxic`", "`severe toxic`", "`threat`", "`insult`", "`obscene`",

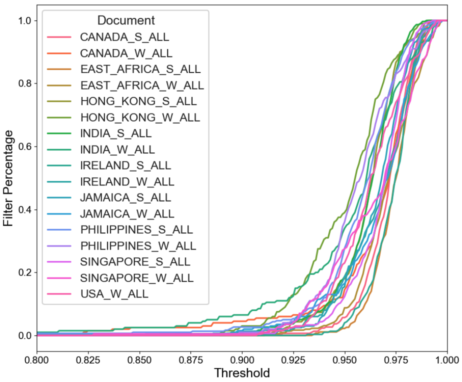


Figure 7: Percentage of English-language documents in the International Corpus of English (ICE) (Kirk and Nelson, 2018) that would be misidentified as non-English as a result of thresholding the FastText classifier's predicted English score. We find a majority of English documents in ICE remain identified as English even with a threshold of 0.90.

and/or "`identity hate`" alongside unlabeled comments, to train two FastText classifiers—a binary "`hate`" detector and a binary "`NSFW`" detector:

1. For our "`hate`" detector, we group all unlabeled comments and "`obscene`"-only comments as negatives and leave remaining comments as positives.
2. For our "`NSFW`" detector, we take all comments tagged as "`obscene`" as positives and leave other remaining comments as negatives. It is important to note this detector only filters *toxic content* that mentions sexual or obscene topics, not sexual content in general.

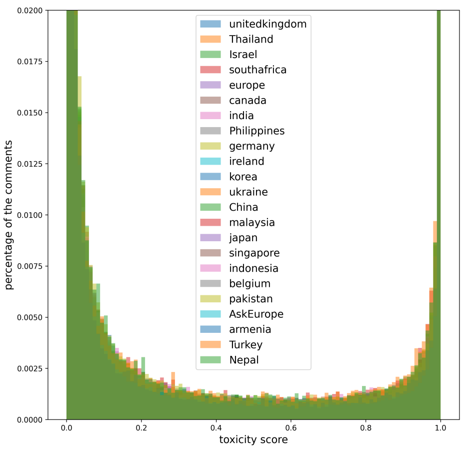


Figure 8: Distribution of Reddit comments labeled as toxic by English variation.

Analysis of resulting classifier. To measure dialectal biases in the FastText toxicity classifier, we analyze its proclivity to predict English variations spoken in different countries as toxic. Starting with the unfiltered Reddit corpus, we create a dataset of comments from location-based subreddits,[17] filtering for country-specific subreddits with more than 50K comments. This dataset serves as a crude proxy for different dialects of English, assuming most commenters live in the respective locations and speak the variation. We further assume the fraction of actually toxic comments in each of these subreddits to be roughly the same. We compute the toxicity score for each comment in this dataset using the FastText classifier and report the percentage of comments marked as toxic against different classifier thresholds in Figure 8. For all thresholds, for any two locations, we find <5% difference in the fraction of comments marked as toxic suggesting little to no bias. Further, we plot the distribution of toxicity scores for comments in each subreddit and find that scores assigned to the comments often fall at the extremes (close to 0 or close to 1), suggesting that any reasonable threshold (lying between 0.1 to 0.9) to predict toxicity will lead to similar outcomes.

I Details on PII Filters

Filter implementation. The Common Crawl, C4, Reddit, and GitHub subsets used the same regular expressions for identifying PII. We refer the reader to our GitHub for exact implementations of our regular expressions for each of the PII types — email address, phone number, and IP address. Once spans are tagged, we employ different processing strategies based on the their density on each document:

- *5 or fewer PII spans detected*: we replace all spans on a page with special tokens |||EMAIL_ADDRESS|||, |||PHONE_NUMBER|||, and |||IP_ADDRESS||| for email addresses, phone numbers, and IP addresses respectively.[18] In total, we find that 0.02% of documents in the 25 Common Crawl snapshots match this filter.

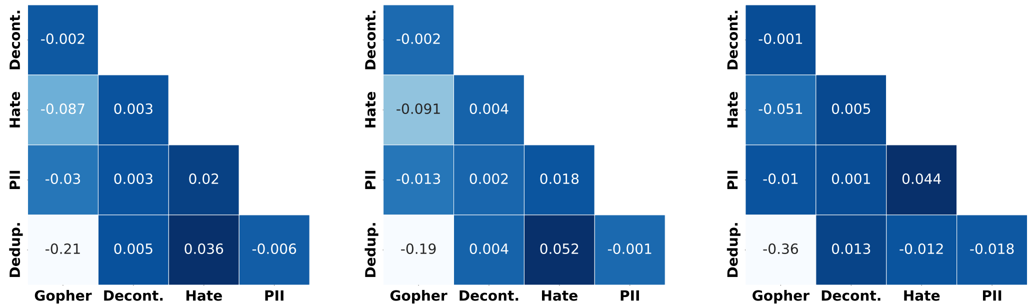
- *6 or more PII spans detected*: we remove any document that contains 6 or more matching PII spans. We use this approach because pages containing abundant phone numbers and email addresses are likely to pose a greater risk of disclosing other PII classes. 0.001% of documents in the 25 Common Crawl snapshots match this filter.

J Do quality and content filters have similar effects?

In order to further understand how filters described in § 5.2, § 5.3, and § 5.4 interact with each other, we perform a correlation analysis on a subset of documents sampled from our pipeline. The correlation among the documents flagged for removal by our Common Crawl filters is depicted in Figure 9. Overall, we find that correlations are generally low, thus our filters select fairly different documents and are not redundant.

[17] `reddit.com/r/LocationReddits/wiki/index`

[18] When training models on Dolma, we add these special tokens to the tokenizer vocabulary.



(a) High (b) Medium (c) Low

Figure 9: Pearson Correlation of various Dolma filters on the High, Medium, and Low buckets of our Common Crawl data, computed over 24M, 20M, and 43M documents, respectively. The filters are Gopher=Gopher rules from Rae et al. (2021), Dedup.=Deduplication, PII=Personally Identifiable Information, Hate=Toxicity and Decont.=Decontamination. Calculated at the document-level: two filters contribute to positive correlation when any span in a document is tagged by both filters. We find our various filters remove different documents and are not redundant.

There is some positive correlation between our PII (Personal Identifiable Information) filters and filters removing hate speech. This is likely because hate speech is often directed at people. The Gopher filtering rules correlate negatively with our deduplication, especially for the high-perplexity tail part of our data. This is due to the Gopher rules removing many high-perplexity documents such as random strings, which are not caught by deduplication due to their randomness. As these random strings likely do not contribute to a better understanding of language, it is important to filter them out and thus rely on filters beyond deduplication.

K Dolma data distribution figures using WIMBD

We use the tool from Elazar et al. (2023) to inspect the final data composition in Figure 10. In particular, we analyze web domain, year, and language distributions.

We note that Dolma contains documents from a broad set of internet domains, mostly from 2020, 2022, and 2021. The most common internet domains in Dolma, per token, are `patents.google.com`, followed by `www.nature.com` and `www.frontiersin.org`. In fact, similar to other corpora reported in Elazar et al. (2023), 63.6% of Dolma's web documents are from '.com' sites (followed then by '.org' and '.co.uk' sites). Finally, as all language identification tools are imperfect, we summarize what languages are remaining post English-only filtering: We find the most common language after English is not well identified ('un') with 0.86% of the documents, followed by 0.06% of the documents identified as Chinese.

L Test Set Contamination in Dolma

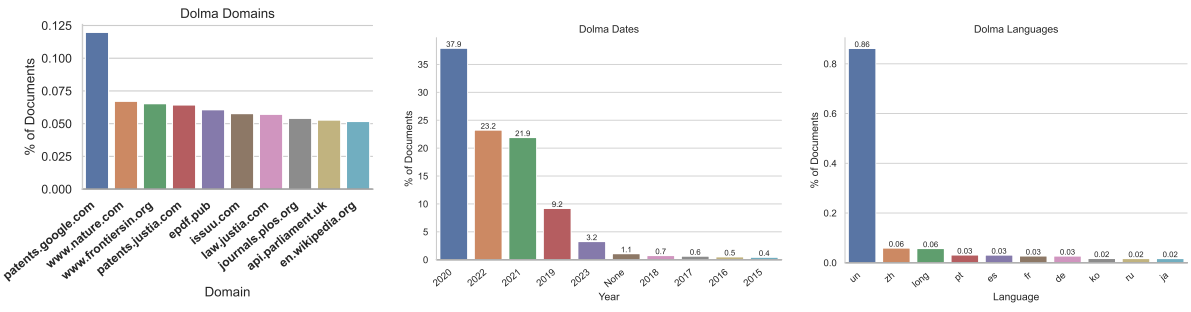
Decontamination for perplexity evaluation. Using the paragraph deduplication tools described in §5.4, we mark any paragraph in Dolma as contaminated if (*i*) it is longer than 13 Unicode-segmented tokens[19] and (*ii*) it appears in any of the documents in Paloma.

To train `OLMo-1B`, we remove any document with at least one paragraph marked as contaminated. This approach, while prone to false positives, has a negligible impact on the final removal rate ($\leq 0.001\%$ characters in Dolma contaminated, $\leq 0.02\%$ of documents removed.), and reduces likelihood of false negatives.

Decontamination of downstream tasks. Using WIMBD (Elazar et al., 2023), we analyze test set contamination in Dolma. We find contamination of entire datasets from popular benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019), and evaluation datasets like SNLI (Bowman et al., 2015b) and the Winograd Schema Challenge (Levesque et al., 2012). Further analysis reveals that many of these sets are contaminated in our code subset, as public repositories in GitHub often contains copies of these datasets. We report the top contaminated datasets in Figure 11.

Results indicate that portion of datasets in Promptsource appear in Dolma. Six datasets are completely contaminated (100%): the Winograd Schema Challenge (Levesque et al., 2012), Sick (Marelli et al., 2014), AX from GLUE (Wang et al., 2018), SemEval (specifically, Task 1 from 2014), COPA from SuperGLUE (Roemmele et al., 2011), and AX_b (the diagnostic task) from SuperGLUE (Wang et al., 2019). In addition, other datasets are mostly contaminated, with over 90% of their test sets appearing in Dolma documents: OpenAI HumanEval (Chen et al., 2021), WIC from SuperGLUE (Pilehvar and Camacho-Collados, 2019), ESNLI (Camburu et al., 2018), and SNLI (Bowman et al., 2015a). We note that the contaminated datasets have been excluded from the downstream tasks we use for model evaluation (*c.r.f.* Appendix D).

[19] Like in Elazar et al. (2023), we only consider paragraphs of sufficient length to avoid false positive matches.



(a) Web (URL) domains (b) Dates of documents (c) Non-English languages

Figure 10: Frequencies over different document metadata as computed using the WIMBD tool from Elazar et al. (2023). In subfigure (c), `un` denotes documents whose language could not be identified; `long` indicates documents that are too long to be processed with the tool's language ID module.

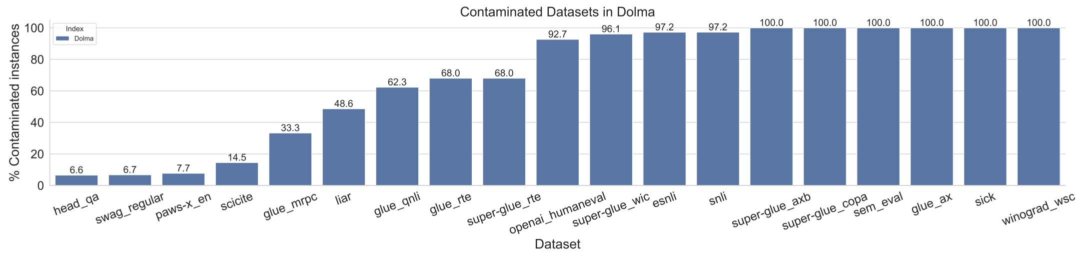


Figure 11: Contamination percentages of datasets from PromptSource (Bach et al., 2022).

M Strategies for Subsets Mixing and Upsampling with Dolma

Like the pretraining corpora of nearly every large-scale language model, Dolma is a multi-source dataset. Training on Dolma thus requires a mixing strategy that determines how much data from each source to include, and potentially which sources to upsample. Like other multi-source corpora (e.g., ROOTS (Laurenccon et al., 2023), the Pile (Gao et al., 2020), RedPajama v1 (Together Computer, 2023a)),[20] Dolma does not prescribe a single mixing strategy. We refer the reader to Rae et al. (2021) for an example of how one might programmatically search over mixing configurations to maximize performance. Here, we perform mixing experiments as an opportunity to answer some research questions about how different data sources interact. We use the same ablation setup described in §4.

How much code is important for pretraining? It is common practice for language models to be pretrained on some amount of code, even if code generation is not the intended task. Some research has suggested that mixing code into training over plain text documents improves performance on reasoning tasks (Madaan et al., 2022). We investigate whether this observation holds for models trained on Dolma, and if so, how much code is needed?

We create three mixtures from the C4 and Stack subsets containing 0%, 5% and 15% of code data. On each, we train a 1B model. We evaluate these models on three different reasoning tasks: bAbI (Weston et al., 2015), WebNLG (Gardent et al., 2017) and GSM8k (Cobbe et al., 2021). For the first two tasks, we follow the experimental setup of Muennighoff et al. (2023b) and evaluate each model in an ICL setup with a changing number of demonstrations (0-5) across 5 random seeds. Muennighoff et al. (2023b) show that adding code to pre-training data improves ICL performance on bAbI and WebNLG and they suggest that code improves long-range state-tracking capabilities. Our experiments, as shown in Table 3, corroborate these findings: while the C4-only model fails on all bAbI tasks, adding code improves performance, with a similar trend for WebNLG.

On the more difficult GSM8k benchmark, all models failed to get any correct answer in an ICL setup, and even when fine-tuning the models on the entire training set. However, we find that by fine-tuning on program-aided output, where questions are solved by writing Python snippets as described in (Gao et al., 2022), code models outperform the C4-only model. These results show that models pre-trained on code can leverage code generation to answer challenging reasoning tasks even when the original task does not directly involve code.

Evaluating mixing strategies for pretraining on Dolma While Dolma does not prescribe a specific source mixture, we analyze some commonly used strate-

[20] RedPajama v1 was a reproduction of the multi-source corpus used in LLaMA (Touvron et al., 2023a). RedPajama v2 (Together Computer, 2023b) focuses solely on Common Crawl and is thus single-source.

Dataset	0% Code	5% Code	15% Code
bAbI (ICL)	0.0 ± 0.0	8.8 ± 0.9	10.1 ± 2.8
WebNLG (ICL)	16.8 ± 1.1	19.3 ± 1.1	22.0 ± 1.3
GSM8K (FT)	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
GSM8K+PAL (FT)	11.8 ± 0.8	14.2 ± 1.3	14.7 ± 0.9

Table 3: Performance of three models pre-trained with increasing amounts of code on three datasets, across 5 random seeds. We measure exact match for bAbI and GSM8K, and Rouge-2 for WebNLG.

gies[21] and compare their effect using the Paloma evaluation suite (Magnusson et al., 2023). Specifically, we present and evaluate four possible data mixtures in Table 4.

We show results of mixtures in Figure 12. Overall, we observe that the different mixtures have an effect on the ability of resulting models to capture specific subdomains. All mixtures show similar perplexity scores on pages sampled from 100 domains from C4 (Figure 12, left), indicating their general effectiveness at modeling web documents. On the other hand, we note how models struggle to model specialized domains unless they are exposed to them. As an example, a model trained on the *Web-only* mix struggles to represent data in the code domain (Figure 12, center, HumanEval). Finally, we use results on the S2ORC subset of M2D2, which consists of academic papers, to illustrate how different data mixtures affect perplexity. As is it the case with code, *Web-only* model exhibits higer perplexity due to domain mismatch. On the other hand, models trained on *Reference+* and *Gopher-like* mixes achieve lower perplexity than the model trained on the *Naïve* mix, due to more in-domain content. However, we note that, despite significant differences in the amount of academic papers between *Reference+* and *Gopher-like* (4.9% vs 24.2%), they achieve nearly identical results, suggesting that even a relatively small percentage of in-domain data is sufficient to achieve good domain fit.

N Datasheet

Following the template by Gebru et al. (2021), we provide a Datasheet for Dolma.

N.1 Motivation for Dataset Creation

Why was the dataset created?

Dolma was created with the primary purpose of training OLMo autoregressive language model. 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?

Due to the myriad transformations applied to the original source materials to derive our dataset, we believe it is ill-suited as a replacement for users seeking to directly consume the original content. We refer users of our dataset to our license and terms on the Hugging Face Hub `huggingface.co/datasets/allenai/dolma` which detail any use restrictions.

Has the dataset been used for any tasks already?

The OLMo (Groeneveld et al., 2024) model family is trained on this dataset.

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

Experimental results are detailed in this paper and in the OLMo (Groeneveld et al., 2024) manuscript.

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.

N.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, docu-

[21] We did not include any social data in these mixes as it was not ready at the time of this experiment.

Mix Name	Description	Sampling	Proportion
Naïve	Sample each source in Table 1 equally.	Web 100% Code 100% Ref. 100% Books 100%	Web 83.5% Code 13.8% Ref. 2.5% Books 0.2%
Web Only	Similar to Ayoola et al. (2022), we test a mixture that only uses web data.	Web 100% Code 0% Ref. 0% Books 0%	Web 100% Code 0% Ref. 0% Books 0%
Reference+	It is common practice to upsamole knowledge-intensive documents when composing training mixture. In our case, we upsample the PeS2o papers, Wikipedia, Wikibooks, and Gutenberg books subsets by 2x.	Web 100% Code 100% Ref. 200% Books 200%	Web 81.2% Code 13.5% Ref. 4.9% Books 0.4%
Gopher-like	Following Rae et al. (2021), we create a mix that is heavily biased towards reference material. As we do not have access to the same sources, an exact replication of their mix is not possible.	Web 17% Code 8% Ref. 200% Books 200%	Web 68.4% Code 5.4% Ref. 24.2% Books 2.0%

Table 4: Overview of the mixtures and their composition.

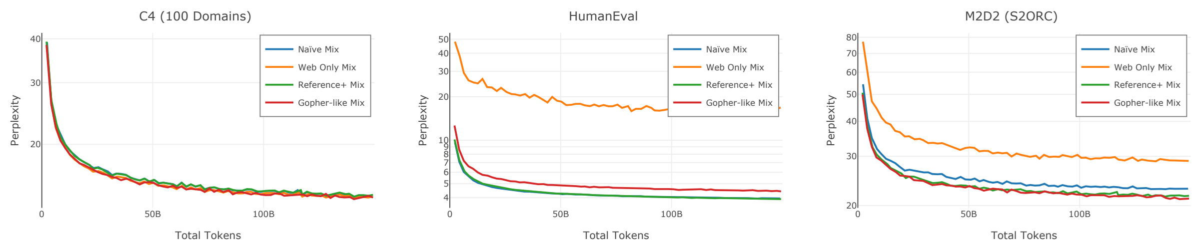


Figure 12: 1B model ablations for different proportions of Dolma data. All mixture perform similarly on web data (left), while excluding code increases perplexity on code datasets (center). Finally, increasing reference material by upsampling papers and Wikipedia yields lower perplexity on S2ORC (right). Overall, source distribution is linked to downstream capabilities; thus, Dolma users should sample subsets according to their needs.

ments, 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.
- **Reddit**. The originating subreddits and thread ids of documents are included in the metadata.
- **Semantic Scholar**. 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 (Kinney et al., 2023).
- **GitHub**. 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.
- **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 snapshots 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 Hugging Face

Hub[22]. In turn, documents in C4 have been derived from a Common Crawl snapshot for 04/2019. URLs in C4 can be used to recover HTML files.

- **Reddit**. The complete set of monthly data dumps used in this work are no longer distributed by Pushshift, however they can still be obtained through torrents and some public web archives.
- **Semantic Scholar**. peS2o is derived from S2ORC (Lo et al., 2020). 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.
- **GitHub**. 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?

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[23]. We followed Common Crawl terms of use[24].
- **C4**. This corpus can be obtained from from the Hugging Face Hub[22] and is released under ODC-By 1.0 (Open Data Commons, 2010).
- **Reddit**. Pushshift no longer distributes this dataset due to changes to the Reddit API's terms. Unofficial copies of the data might be be available through torrents and some public web archives. Pushshift data dumps inherit[25] the Terms of use of the Reddit API at the time of their collection (March 2023).
- **Semantic Scholar**. peS2o is derived from S2ORC (Lo et al., 2020). S2ORC is released through the Semantic Scholar Public API[26] under ODC-By 1.0 (Open Data Commons, 2010).
- **GitHub**. The corpus is available on the Hugging Face Hub[27] 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`.
- **Wikipedia**, **Wikibooks**. Wikimedia data dumps are freely available[28] and released under CC BY-SA 4.0 license (Creative Commons, 2013).

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

No. See current manuscript Section §4.2.

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.

See current manuscript Section §4.2 for description of data ablation methodology, and remainder of paper for full set of experiments. Every experimental result is available through links provided in the manuscript.

N.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**. snapshots were downloaded from Common Crawl's official S3 bucket[29] using the `cc_net` pipeline (Wenzek et al., 2020). Data was obtained between March 17th and March 27th, 2023.

[22] `hf.co/datasets/allenai/c4`
[23] `commoncrawl.org/the-data/get-started`
[24] `commoncrawl.org/terms-of-use`
[25] `reddit.com/r/pushshift/comments/d6luj5/comment/f0ugpqp`
[26] `semanticscholar.org/product/api`
[27] `hf.co/datasets/bigcode/the-stack-dedup`
[28] `dumps.wikimedia.org`
[29] `s3://commoncrawl/`

- **C4**. We clone C4 from the Hugging Face Hub[22] using Git with the Git-LFS extension. Repository cloned on May 24th, 2023.

- **Reddit**. Reddit was acquired in the form of monthly data dumps of comments and submissions collected and distributed by the Pushshift project[30]. We used the complete set of 422 publicly available dumps (208 comments, 214 submissions) spanning a period from 06/2005–03/2023. The majority of Dumps were acquired in March, 2023 with the last dumps downloaded in May of 2023.

- **Semantic Scholar**. We clone peS2o from the Hugging Face Hub[31] using Git with the Git-LFS extension. We use pes2o V2. Repository cloned on June 30th, 2023.

- **GitHub**. We clone The Stack (deduplicated) from the Hugging Face Hub[27] using Git with the Git-LFS extension. Repository cloned on May 28th, 2023.

- **Project Gutenberg**. Data was downloaded directly from `gutenberg.org`. We used `GutenbergPy` (Angelescu, Radu, 2013) to extract books. Website accessed on April 3rd, 2023.

- **Wikipedia**, **Wikibooks**. Dumps were downloaded from Wikimedia's website[28]. We use the dump from March 20th, 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` (Common Crawl, 2016) project, available at `commoncrawl.github.io/cc-crawl-statistics`. Dolma uses Common Crawl snapshots from `2020-05` to `2023-06`[32].

- **C4**. We use C4 in its entirety.

- **Reddit**. We use all available Reddit content from from 06/2005–03/2023.

- **GitHub**. We use The Stack (deduplicated) in its entirety.

- **Semantic Scholar**. 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 snapshots available. This amount was deemed sufficient for the goal of the Dolma project. We decided to use the 24 most recent Common Crawl snapshots at the time.

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.

N.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.)

[30] `files.pushshift.io/reddit/submissions` and `files.pushshift.io/reddit/comments`

[31] `hf.co/datasets/allenai/peS2o`

[32] Common Crawl snapshots follow naming convention `xxxx-yy`, where `xxxx` is the year the snapshot was finalized, and `yy` is the week, ranging from `01` to `52`.

All data sources are filtered using FastText language identification models (Joulin et al., 2016a,b) with an English threshold of 0.5.

For the **Common Crawl** and **C4** subsets, we use the following filters that substantially modify the original data. Note that data might be tagged for removal by one or more filter.

- ***Only Common Crawl, as part of their distribution pipeline***: Linearize all HTML into plain text files (`WET` files generation[24]);
- ***Only Common Crawl, as part of CCNet pipeline***: We remove frequently occurring paragraph in Common Crawl by identifying repeated paragraphs on small subsets of each snapshots. This step gets rid of headers that are shared across many pages, such as navigational headers. Removal is operationalized as follows: given $1 \ldots, n, \ldots, N$ shards each snapshot is comprised to, group shards in sets $S = \{n - k, n\}$; then, remove exact duplicates of paragraphs in S. Paragraphs are defined as newline-separated slices of documents, and compared using their SHA1. We choose k such that each set is at most 20GB[33]. (*approximately 70% of paragraph removed*);
- ***Only Common Crawl, 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 (Joulin et al., 2016a,b) (*removed 61.69% of web pages by size*);
- **Quality filter**[34]: Remove documents with more than half of their line not ending in ".", "?", "!", or """. (*22.73% of characters tagged for removal*);
- **Quality filter**[34]: Remove any document that does not pass any of the Gopher rules (Rae et al., 2021) (*15.23% of characters tagged for removal*);
  - Fraction of characters in most common ngram greater than a threshold[35]
  - Fraction of characters in duplicate ngrams greater than a threshold[36]
  - Contains fewer than 50 or more than 100K words
  - Median word length is less than 3 or greater than 10
  - Symbol to word ratio greater than 0.10
  - Fraction of words with alpha character less than 0.80
  - Contains fewer than 2 of a set of required words[37]
  - Fraction of lines in document starting with bullet point greater than 0.90
  - Fraction of lines in document ending with ellipsis greater than 0.30
  - Fraction of lines in document that are duplicated greater than 0.30
  - Fraction of characters in duplicated lines greater than 0.30
- **Quality filter**[34]: Remove any document that contains a token or sequence of tokens repeating over 100 times[38] (*0.003% of characters tagged for removal*);
- **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 (cjadams et al., 2017) (*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 PIIs are completely removed from the corpus (*0.05% tagged for masking, 0.11% tagged for removal*);
- **Exact document deduplication**: duplicate documents the same text. No punctuation or whitespace is removed. Empty documents count as duplicates (*14.9% of documents tagged for removal*).
- ***Only Common Crawl, deduplication by paragraph***: We deduplicate the web subset at a paragraph level using a Bloom filter (*19.1% of UTF-8 characters tagged for removal*).

For the **Reddit** subset, we use the following filters that substantially reduce the original data.

- **Language identification**: remove all documents with an English score lower than 0.5, as determined by a FastText language identification model.

[33]This is a slight modification of the original CCNet pipeline, where k is chose so that each set is 2% of snapshot. We chose to use a fixed shard size, rather an a percentage of the corpus, because fixed size is more predictable in terms of resource usage, leading to less-error prone code. Conceptually it's equivalent to putting a threshold on the absolute probability of a paragraph occurring

[34]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 (Gururangan et al., 2022).

[35]For bigrams, threshold of 0.20. For trigrams, 0.18. For 4-grams, 0.16.

[36]For 5-grams, 0.15. For 6-grams, 0.14. For 7-grams, 0.13. For 8-grams, 0.12. For 9-grams, 0.11. For 10-grams, 0.10.

[37]"the", "be", "to", "of", "and", "that", "have", "with"

[38]We use `allenai/gpt-neox-olmo-dolma-v1\_5` to obtain tokens.

- **Quality filter**[34]: Remove comments and submissions shorter than 500 characters in length.
- **Quality filter**[34]: Remove user comments with fewer than three upvotes (Reddit users vote on the quality of submissions and comments).
- **Content filter**[34]: Remove comments and submissions from banned, toxic, or NSFW subreddits.
- **Content filter**[34]: Remove sentences that get ranked as toxic or as hatespeech by a FastText classifier (score above 0.4).
- **Content filter**: Mask Personal Identifiable Information (PII) using regular expressions that identify emails, phone numbers, and IP addresses
- **Deduplication**: We deduplicate comments and submissions (jointly) at a paragraph level using a Bloom filter.

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

- **Language filter**: Removed files associated with the following programming languages:
  - Data or numerical content: *csv, json, json5, jsonld, jsoniq, svg*
  - Assembly code: *assembly*
- **Quality filter**[34]: Removed copyright statements in code files from document preamble[39];
- **Quality filter**[34]: Removed documents matching any of the RedPajama v1 (Together Computer, 2023a) code filters (*41.49% of data tagged for removal*):
  - Maximum line length > 1000 characters.
  - Average line length > 100 characters.
  - Proportion of alpha-numeric characters < 0.25.
  - Ratio of alphabetical characters to number of tokens < 1.5[40].
- **Quality filter**[34]: Removed documents matching any of the following Starcoder filters (Li et al., 2023b):
  - Contains XML template code.
  - HTML code-to-text ratio <= 0.2.
  - Java, Javascript, Python code-to-comment ratio <= 0.01 or > 0.8.
- **Content filter**: Mask Personal Identifiable Information (PII) using regular expressions that identify emails, phone numbers, and IP addresses; pages containing 6 or more PIIs are completely removed from the corpus.

The **Common Crawl**, **C4**, **Reddit**, and **Code** subsets used the same regular expressions for identifying PII:

- **Email addresses:**
  ```
  [.\s@,?!;:)(]*([\^\s@]+@[\^\s@,?!;:)(]+?)[.\s@,?!;:)(]?[\s\n\r]
  ```
- **IP addresses:**
  ```
  \s+\(?(\d{3})\)?[-\. ]*(\d{3})[-. ]?(\d{4})
  ```
- **Phone numbers:**
  ```
  (?:(?:25[0-5]|2[0-4][0-9]|[01]?[0-9]{1,2})\.){3}(?:25[0-5]|2[0-4][0-9]|[01]?[0-9]{1,2})
  ```

For the **Wikipedia and Wikibooks** subsets, we remove pages that contain fewer than 25 UTF-8 words.

For the **Gutenberg** subset:

- **Language identification**: for each paragraph (defined as newline-separated spans of text), we use FastText to perform language identification. Then, we compute the average language score by averaging the score for all passages. If a document has a language score lower than 0.5, it is discarded;
- **Quality filter**[34]: we remove pages that contain fewer than 25 UTF-8 words;
- **Quality filter**[34]: Remove any document that contains a token or sequence of tokens repeating over 100 times[38].

For the **Semantic Scholar** subset, we remove any document that contains a token or sequence of tokens repeating over 100 times[38] .

For Dolma versions `1.0` and `1.5`, we perform decontamination for all subsets of Dolma. In particular, we remove paragraphs that are shared with documents in the Paloma evaluation suite (Magnusson et al., 2023). Overall, only 0.003% of our dataset is removed due to contamination with this evaluation set. Dolma version `1.6` is not decontaminated.

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 constrains, we only keep linearized version of Common Crawl snapshots, filtered by Language ID as described above.

Raw data is not available for download outside the Allen Institute for AI. Interested individuals may contact authors of this manuscript if they require access to raw data.

Is the preprocessing software available?

Yes, all preprocessing software is available on GitHub at `github.com/allenai/dolma` and on PyPI[41].

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

Yes, it does.

[39] Code license and provenance is still tracked in metadata.

[40] Tokens counted using whitespace tokenizer

[41] `pypi.org/project/dolma`

N.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 Hugging Face Hub, which offers access via the `datasets` (Lhoest et al., 2021) 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.

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?

Information about the license associated with Dolma are available on its release page on the Hugging Face Hub: `huggingface.co/datasets/allenai/dolma`.

Are there any fees or access/export restrictions?

The dataset is distributed for free. Users should verify any restrictions on its release page on the Hugging Face Hub: `huggingface.co/datasets/allenai/dolma`.

N.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[42] or on the community tab of dataset page[43] (the former being preferred over the latter). Any other inquiry should be sent to `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. Information about latest version of Dolma are available on its release page on the Hugging Face Hub: `huggingface.co/datasets/allenai/dolma`. Dolma users should cite this manuscript when using this data.

[42] `github.com/allenai/dolma/issues`

[43] `hf.co/datasets/allenai/dolma/discussions`

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?

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.

N.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 (Soldaini and Lo, 2023) is derived from manuscripts that are licensed for permissive distribution by their authors. Reddit content was acquired through a public API adherent to terms of service; individual authors of Reddit posts were not contacted directly. 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 Dolma 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[44] for individuals who wish to be removed from the dataset.

[44] `forms.gle/q4BNUUxUxKwKkfdT6`

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[44] 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.

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[44].

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.

O All Raw Ablation Results

O.1 Comparing Dolma With Other Corpora

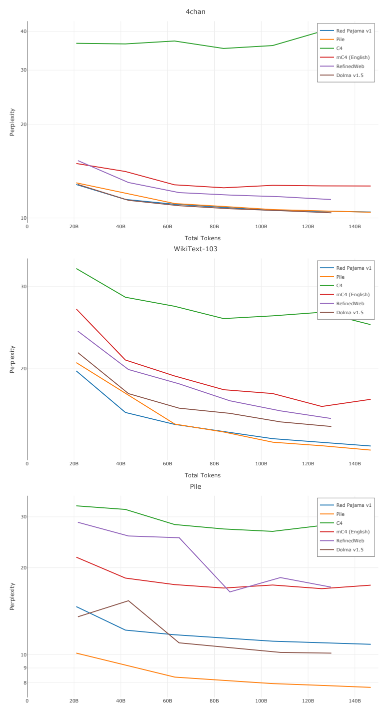


Figure 13: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), WikiText 103 (Merity et al., 2016), and Pile (Gao et al., 2020) (Val)

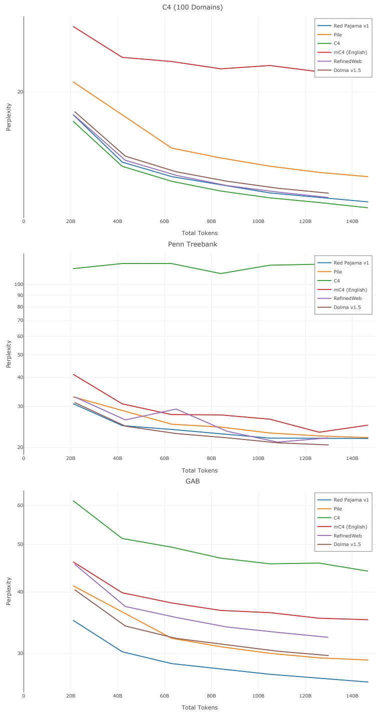


Figure 14: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 100 dom (Chronopoulou et al., 2022), Penn Tree Bank (Marcus et al., 1994), and Gab (Zannettou et al., 2018)

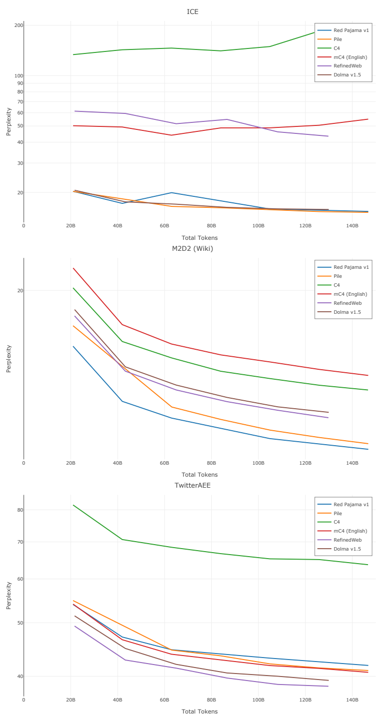


Figure 15: Perplexity results on Paloma (Magnusson et al., 2023); subsets ICE (Greenbaum, 1991), M2D2 (Reid et al., 2022) (Wiki), and Twitter AAE (Blodgett et al., 2016)

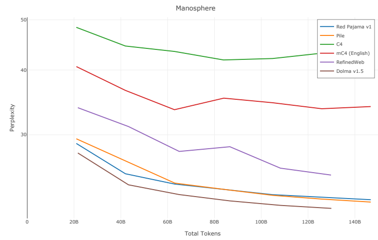


Figure 16: Perplexity results on Paloma (Magnusson et al., 2023); subsets Manosphere (Ribeiro et al., 2021)

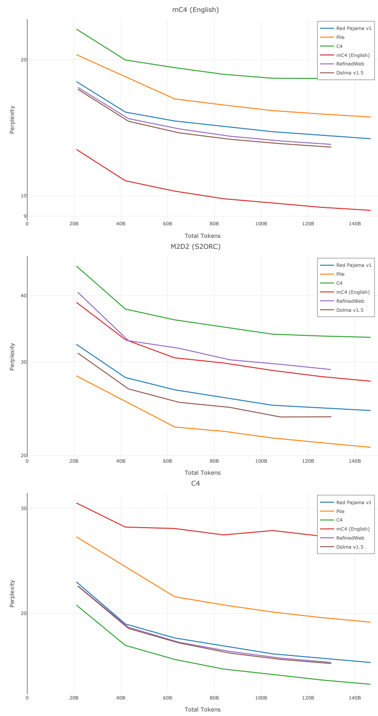


Figure 17: Perplexity results on Paloma (Magnusson et al., 2023); subsets mC4 (Xue et al., 2020) (English), M2D2 (Reid et al., 2022) (S2ORC), and C4 (Raffel et al., 2020; Dodge et al., 2021)

Figure 18: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

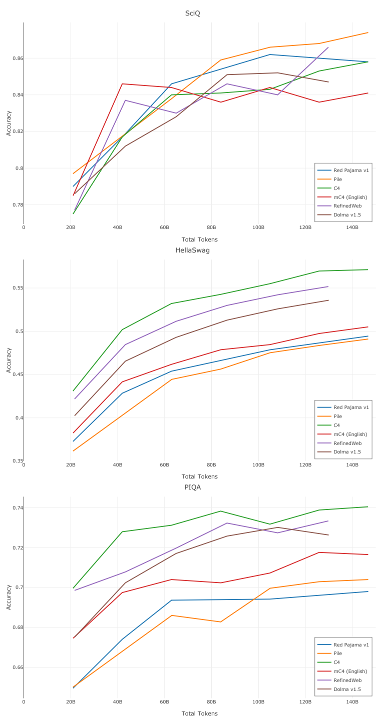


Figure 19: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

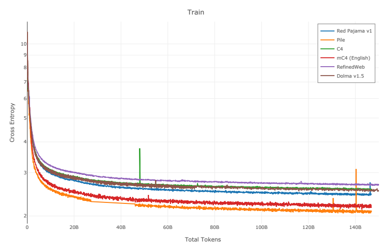


Figure 20: Training Cross Entropy

O.2 Deduping Strategy

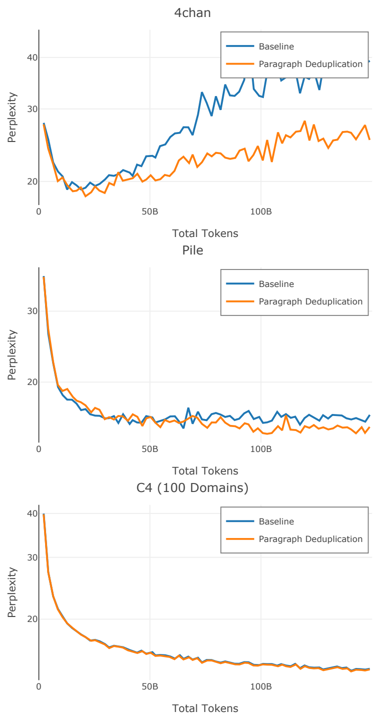


Figure 21: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Pile (Gao et al., 2020) (Val), and C4 100 dom (Chronopoulou et al., 2022)

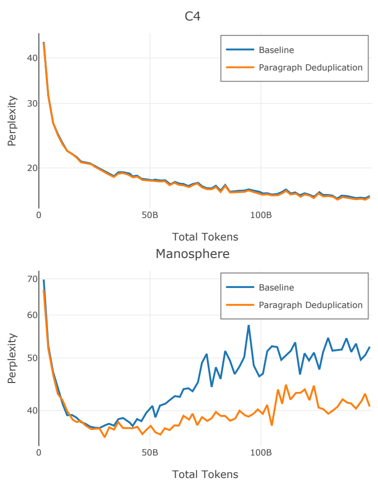


Figure 22: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 (Raffel et al., 2020; Dodge et al., 2021) and Manosphere (Ribeiro et al., 2021)

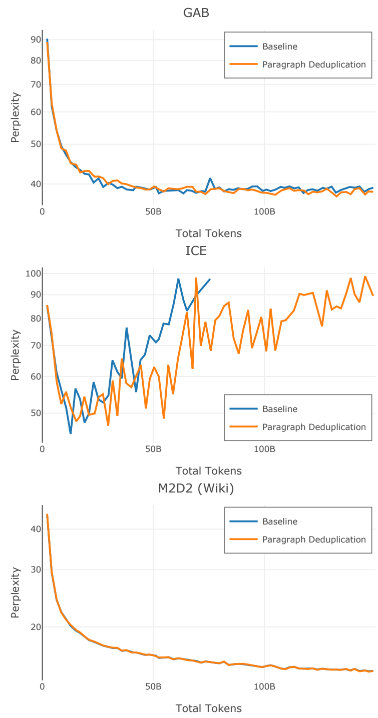


Figure 23: Perplexity results on Paloma (Magnusson et al., 2023); subsets Gab (Zannettou et al., 2018), ICE (Greenbaum, 1991), and M2D2 (Reid et al., 2022) (Wiki)

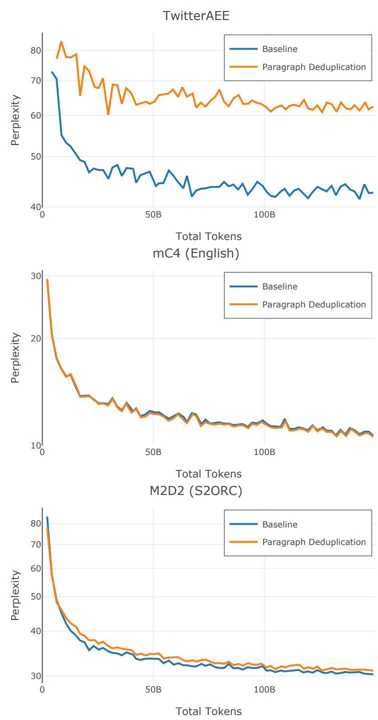


Figure 24: Perplexity results on Paloma (Magnusson et al., 2023); subsets Twitter AAE (Blodgett et al., 2016), mC4 (Xue et al., 2020) (English), and M2D2 (Reid et al., 2022) (S2ORC)

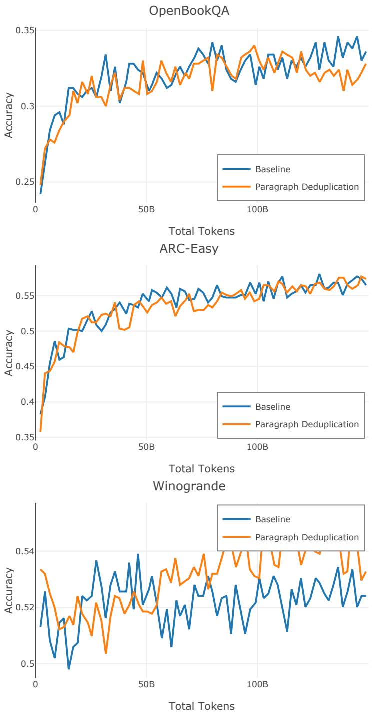


Figure 25: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

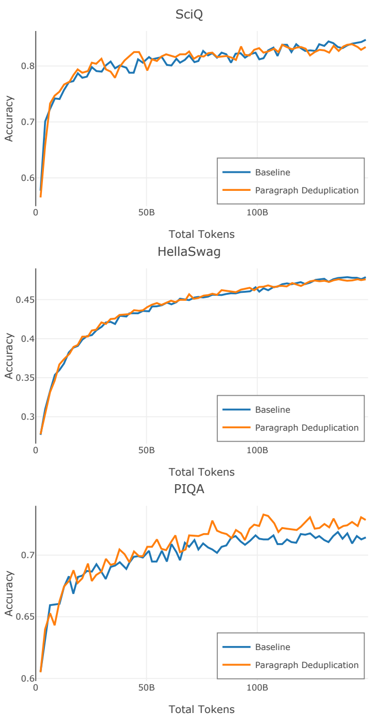


Figure 26: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

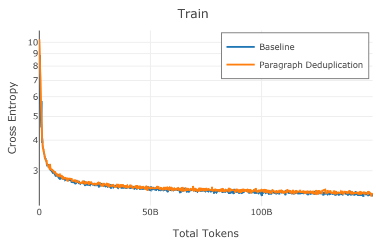


Figure 27: Training Cross Entropy

O.3 Filtering of Personal Identifiable Information

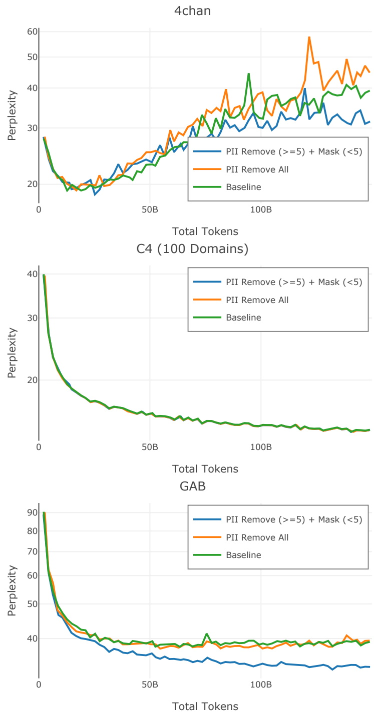


Figure 28: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), C4 100 dom (Chronopoulou et al., 2022), and Gab (Zannettou et al., 2018)

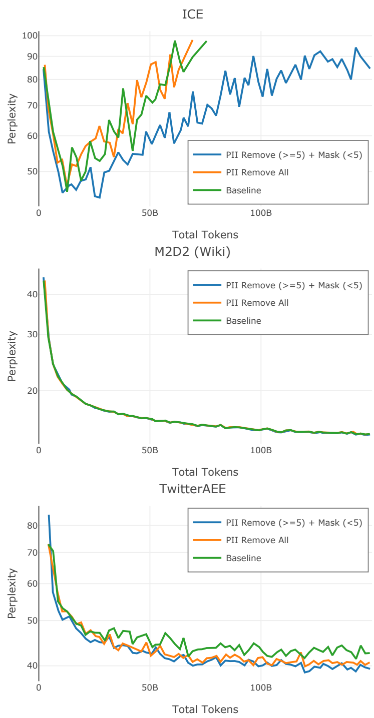


Figure 29: Perplexity results on Paloma (Magnusson et al., 2023); subsets ICE (Greenbaum, 1991), M2D2 (Reid et al., 2022) (Wiki), and Twitter AAE (Blodgett et al., 2016)

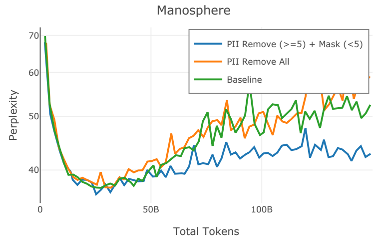


Figure 30: Perplexity results on Paloma (Magnusson et al., 2023); subsets Manosphere (Ribeiro et al., 2021)

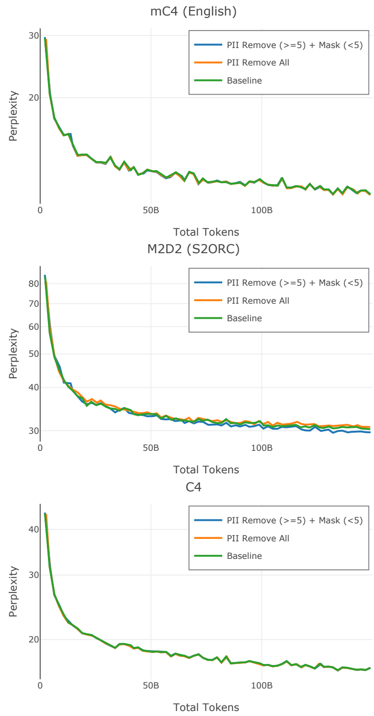


Figure 31: Perplexity results on Paloma (Magnusson et al., 2023); subsets mC4 (Xue et al., 2020) (English), M2D2 (Reid et al., 2022) (S2ORC), and C4 (Raffel et al., 2020; Dodge et al., 2021)

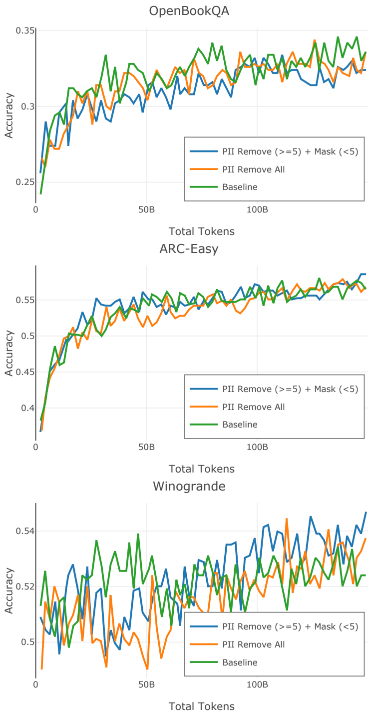


Figure 32: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

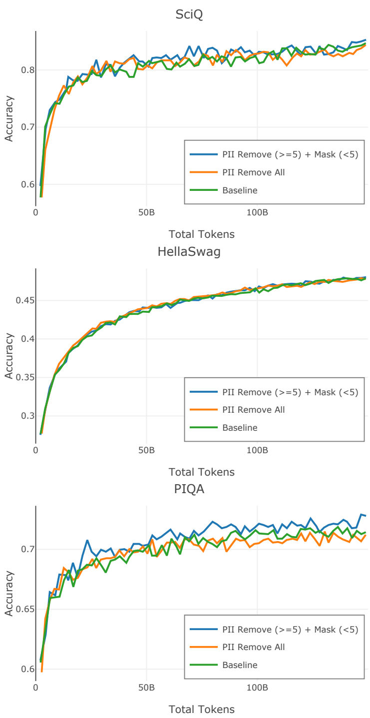


Figure 33: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

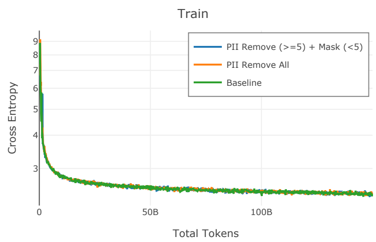


Figure 34: Training Cross Entropy

O.4 Comparing Quality Filters for Web Pipeline

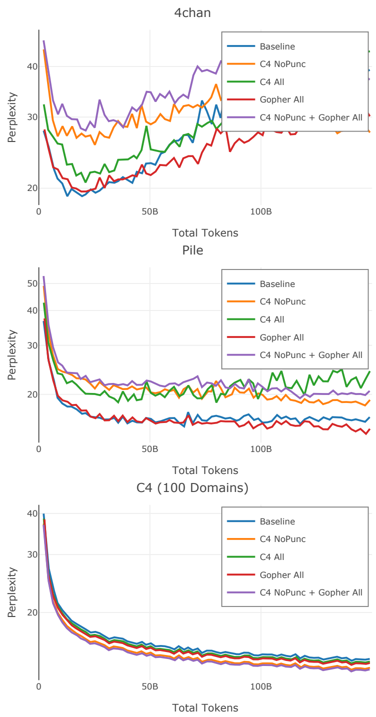


Figure 35: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Pile (Gao et al., 2020) (Val), and C4 100 dom (Chronopoulou et al., 2022)

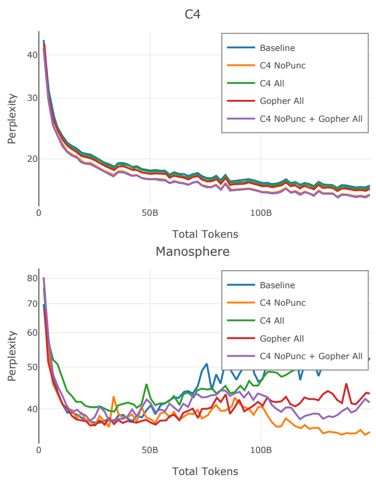


Figure 36: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 (Raffel et al., 2020; Dodge et al., 2021) and Manosphere (Ribeiro et al., 2021)

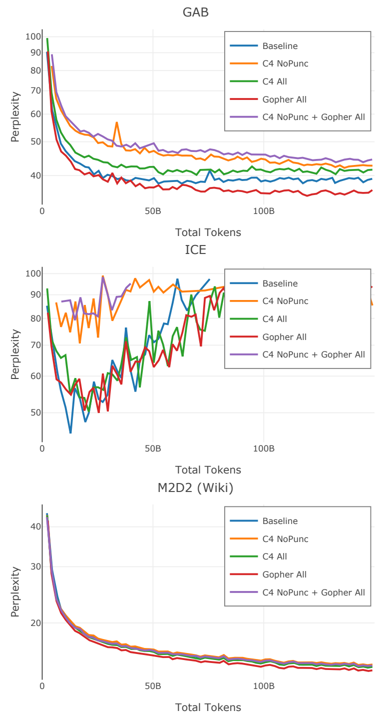


Figure 37: Perplexity results on Paloma (Magnusson et al., 2023); subsets Gab (Zannettou et al., 2018), ICE (Greenbaum, 1991), and M2D2 (Reid et al., 2022) (Wiki)

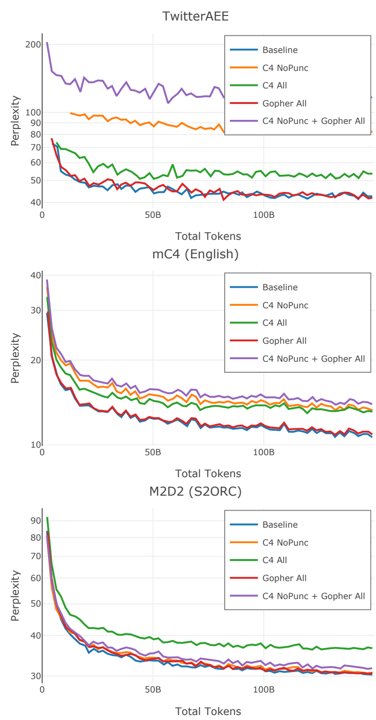


Figure 38: Perplexity results on Paloma (Magnusson et al., 2023); subsets Twitter AAE (Blodgett et al., 2016), mC4 (Xue et al., 2020) (English), and M2D2 (Reid et al., 2022) (S2ORC)

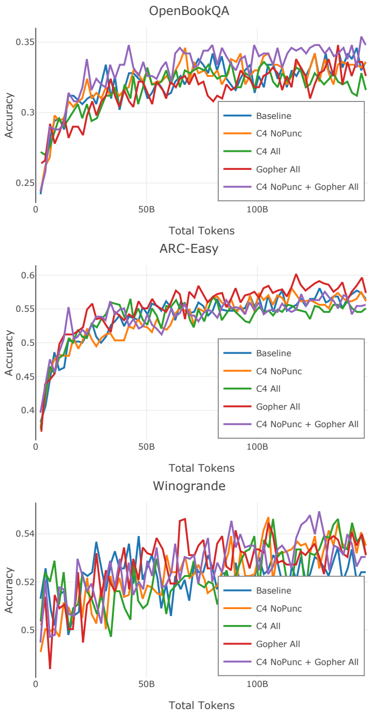


Figure 39: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

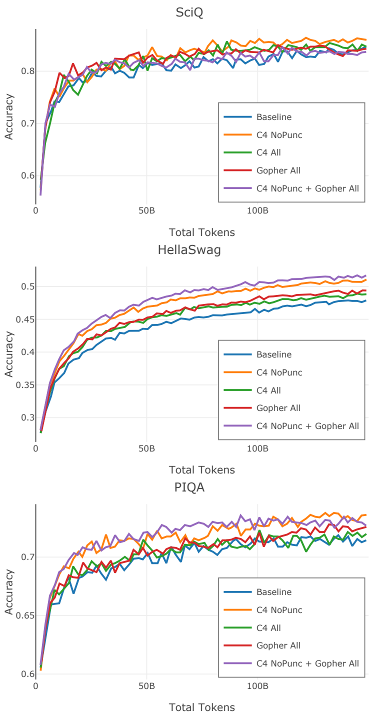


Figure 40: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

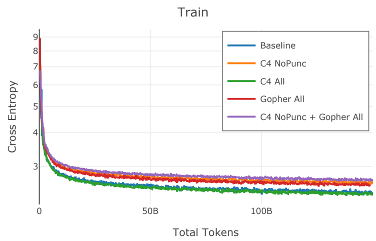


Figure 41: Training Cross Entropy

O.5 Full Comparison of Web Pipeline

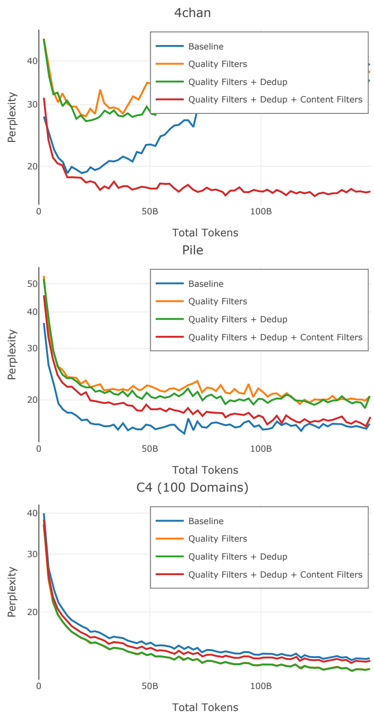


Figure 42: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Pile (Gao et al., 2020) (Val), and C4 100 dom (Chronopoulou et al., 2022)

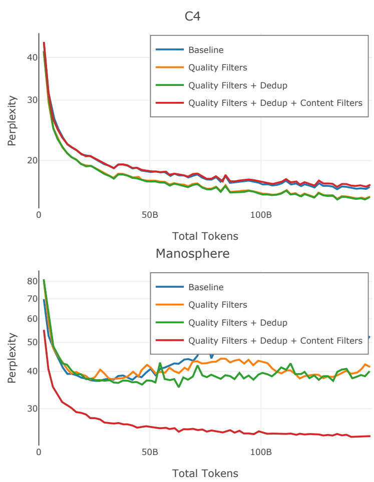


Figure 43: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 (Raffel et al., 2020; Dodge et al., 2021) and Manosphere (Ribeiro et al., 2021)

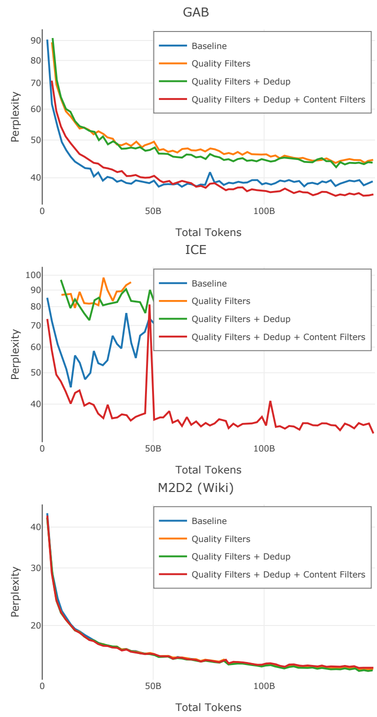


Figure 44: Perplexity results on Paloma (Magnusson et al., 2023); subsets Gab (Zannettou et al., 2018), ICE (Greenbaum, 1991), and M2D2 (Reid et al., 2022) (Wiki)

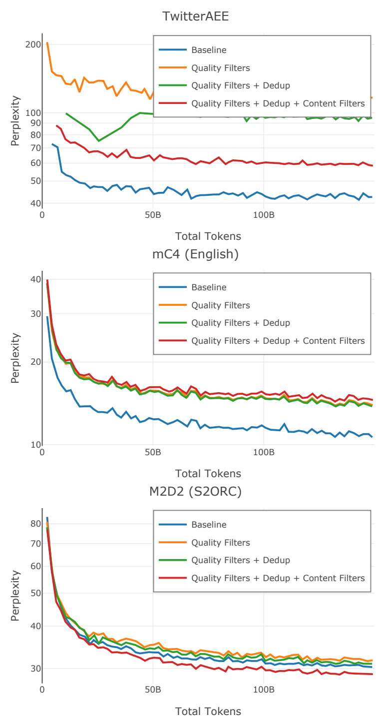


Figure 45: Perplexity results on Paloma (Magnusson et al., 2023); subsets Twitter AAE (Blodgett et al., 2016), mC4 (Xue et al., 2020) (English), and M2D2 (Reid et al., 2022) (S2ORC)

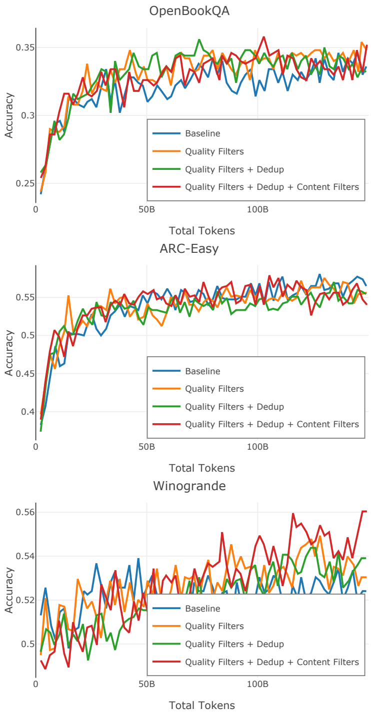


Figure 46: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

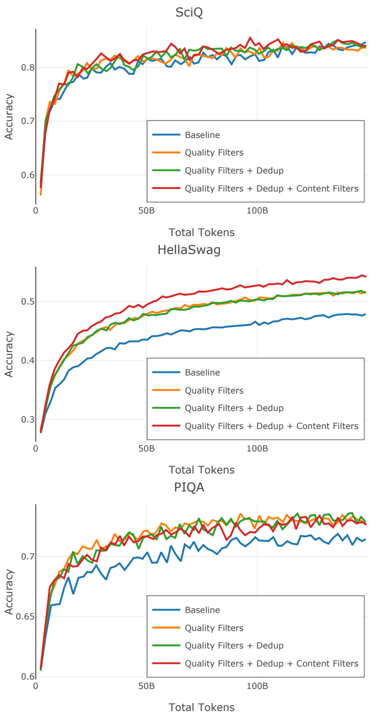


Figure 47: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

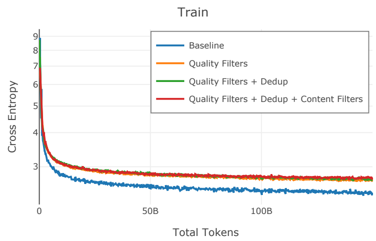


Figure 48: Training Cross Entropy

O.6 Toxicity Filtering in Web Pipeline

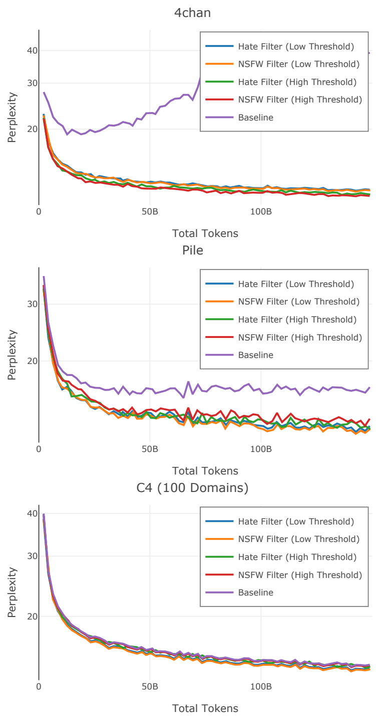


Figure 49: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Pile (Gao et al., 2020) (Val), and C4 100 dom (Chronopoulou et al., 2022)

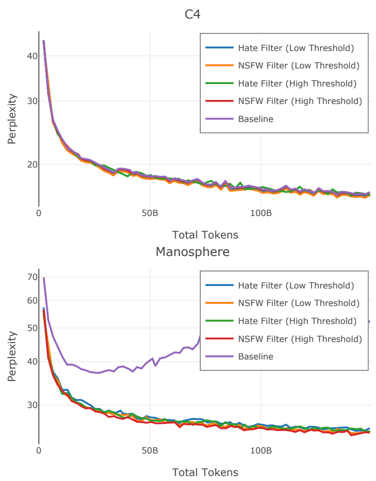


Figure 50: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 (Raffel et al., 2020; Dodge et al., 2021) and Manosphere (Ribeiro et al., 2021)

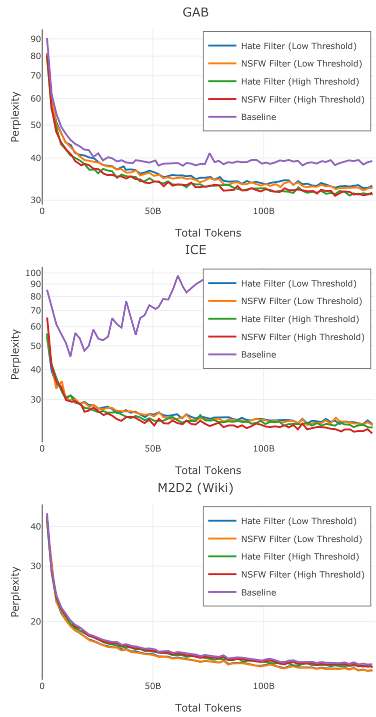


Figure 51: Perplexity results on Paloma (Magnusson et al., 2023); subsets Gab (Zannettou et al., 2018), ICE (Greenbaum, 1991), and M2D2 (Reid et al., 2022) (Wiki)

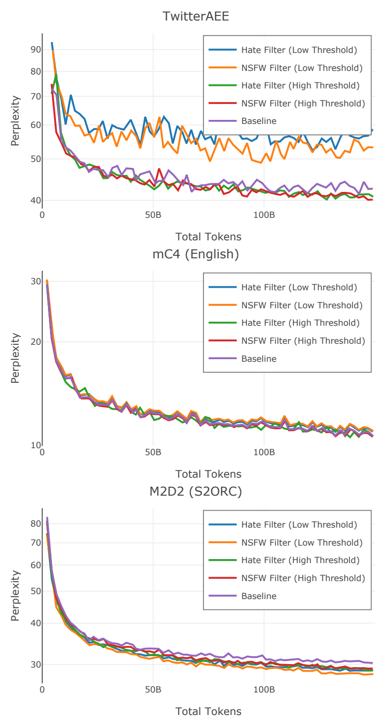


Figure 52: Perplexity results on Paloma (Magnusson et al., 2023); subsets Twitter AAE (Blodgett et al., 2016), mC4 (Xue et al., 2020) (English), and M2D2 (Reid et al., 2022) (S2ORC)

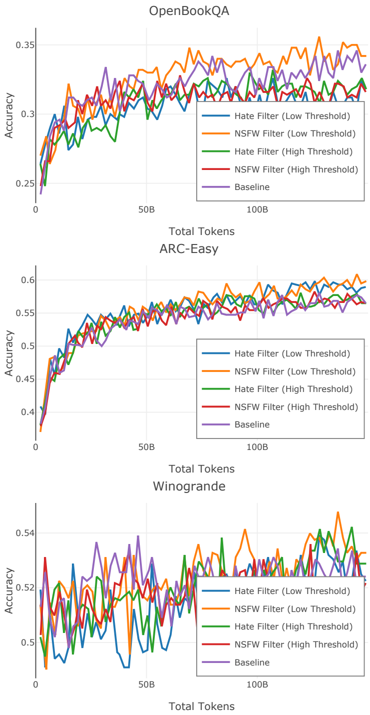


Figure 53: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

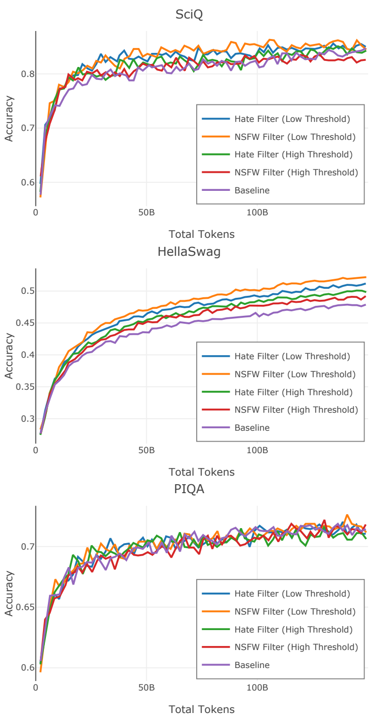


Figure 54: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

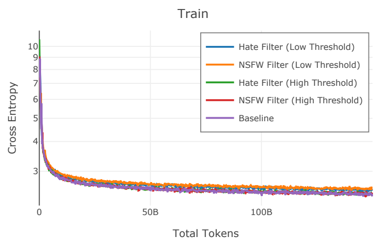


Figure 55: Training Cross Entropy

O.7 Comparing Code Processing Pipeline

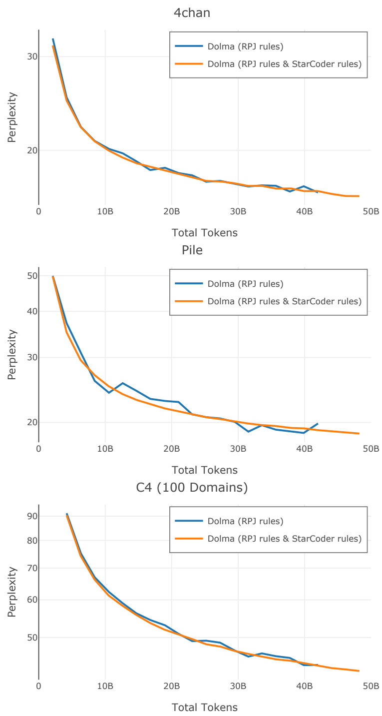


Figure 56: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Pile (Gao et al., 2020) (Val), and C4 100 dom (Chronopoulou et al., 2022)

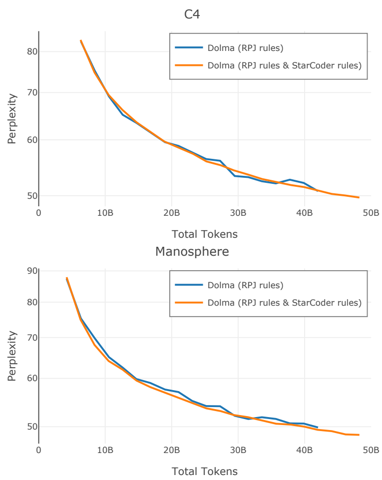


Figure 57: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 (Raffel et al., 2020; Dodge et al., 2021) and Manosphere (Ribeiro et al., 2021)

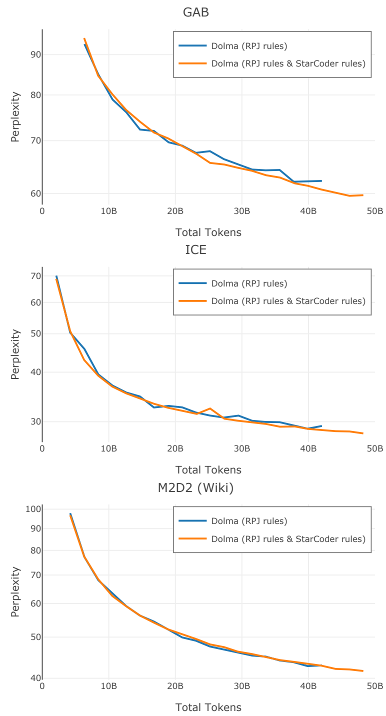


Figure 58: Perplexity results on Paloma (Magnusson et al., 2023); subsets Gab (Zannettou et al., 2018), ICE (Greenbaum, 1991), and M2D2 (Reid et al., 2022) (Wiki)

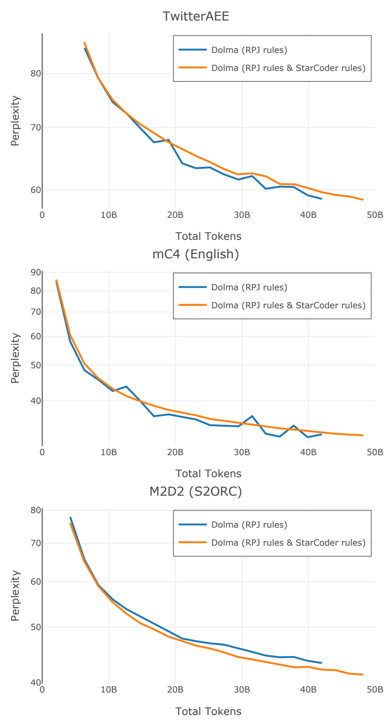


Figure 59: Perplexity results on Paloma (Magnusson et al., 2023); subsets Twitter AAE (Blodgett et al., 2016), mC4 (Xue et al., 2020) (English), and M2D2 (Reid et al., 2022) (S2ORC)

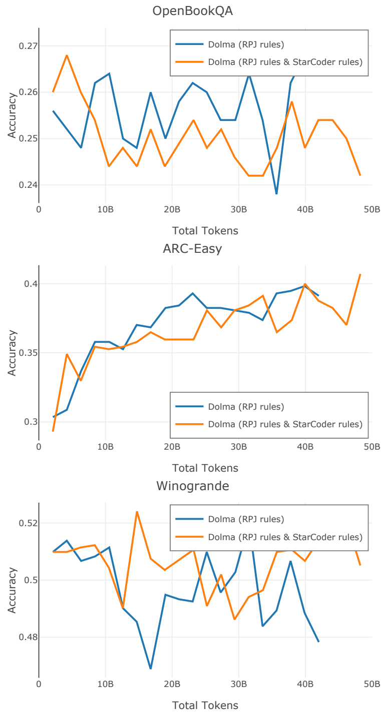


Figure 60: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

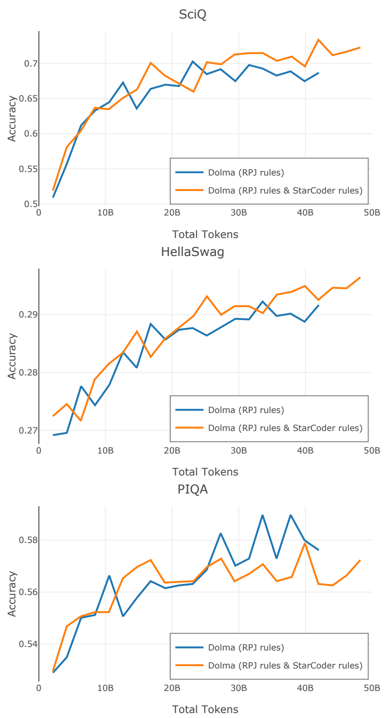


Figure 61: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

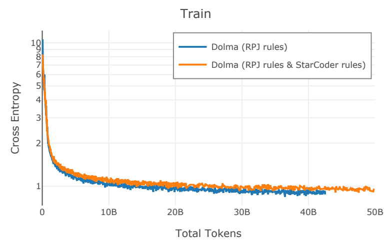


Figure 62: Training Cross Entropy

O.8 Studying Dolma Mixture

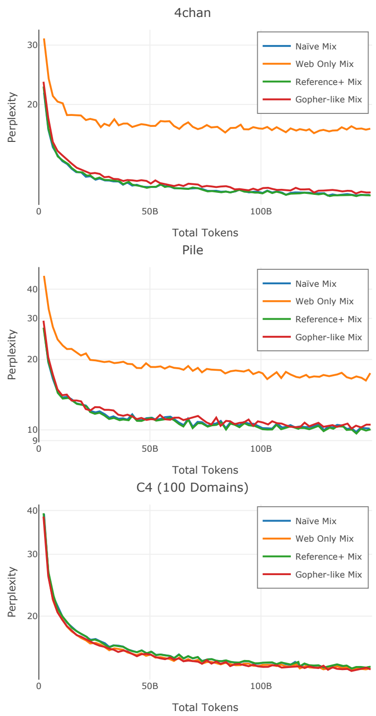


Figure 63: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Pile (Gao et al., 2020) (Val), and C4 100 dom (Chronopoulou et al., 2022)

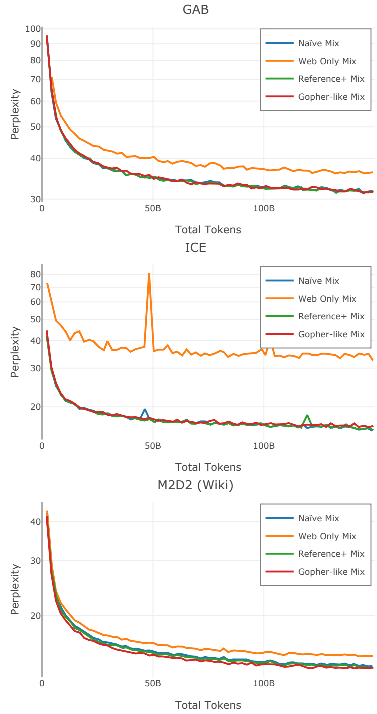


Figure 65: Perplexity results on Paloma (Magnusson et al., 2023); subsets Gab (Zannettou et al., 2018), ICE (Greenbaum, 1991), and M2D2 (Reid et al., 2022) (Wiki)

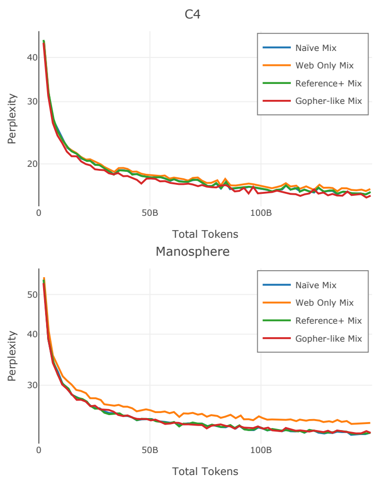


Figure 64: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 (Raffel et al., 2020; Dodge et al., 2021) and Manosphere (Ribeiro et al., 2021)

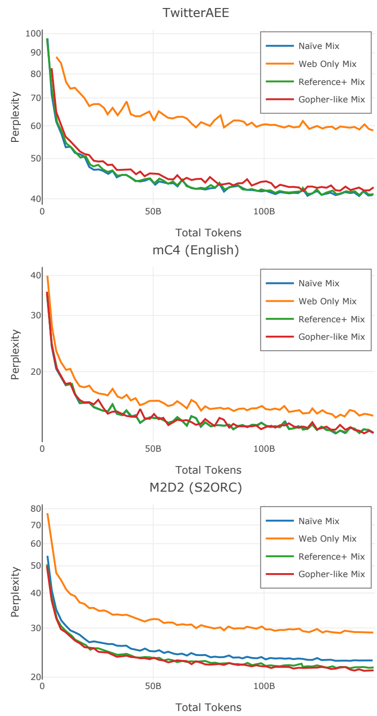


Figure 66: Perplexity results on Paloma (Magnusson et al., 2023); subsets Twitter AAE (Blodgett et al., 2016), mC4 (Xue et al., 2020) (English), and M2D2 (Reid et al., 2022) (S2ORC)

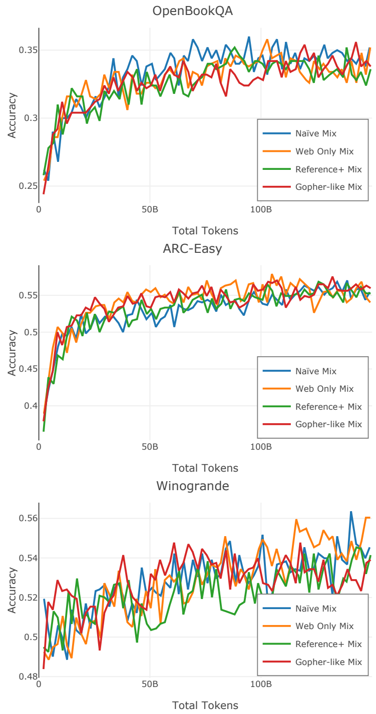


Figure 67: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

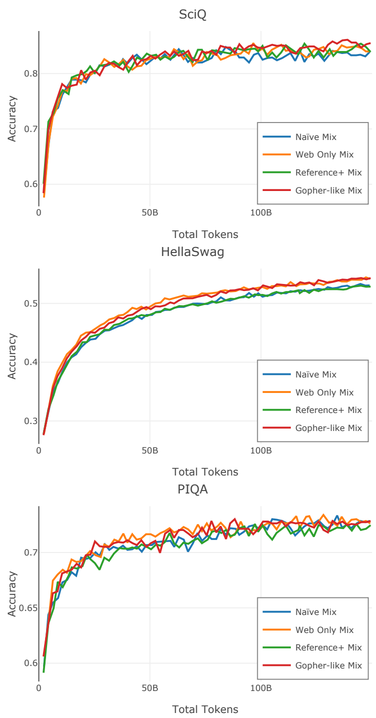


Figure 68: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

O.9 Strategies to Format Conversational Forums Pipeline

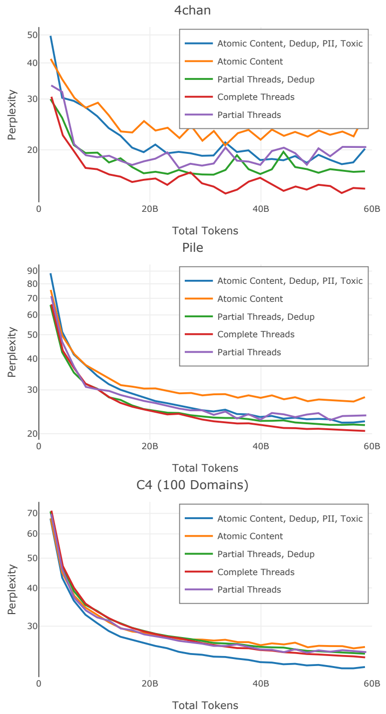


Figure 69: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Pile (Gao et al., 2020) (Val), and C4 100 dom (Chronopoulou et al., 2022)

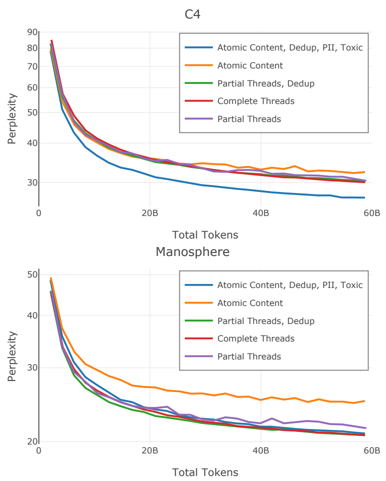


Figure 70: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 (Raffel et al., 2020; Dodge et al., 2021) and Manosphere (Ribeiro et al., 2021)

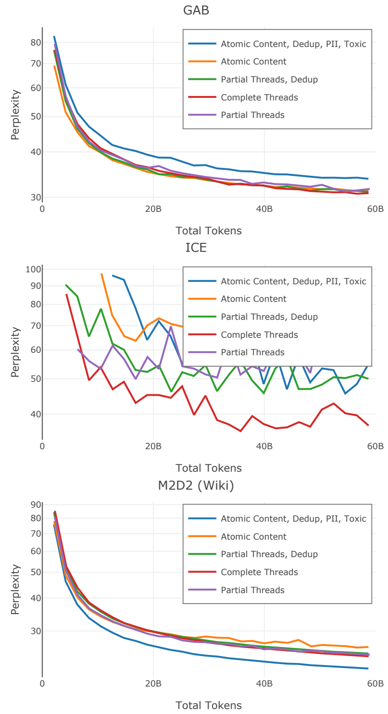


Figure 71: Perplexity results on Paloma (Magnusson et al., 2023); subsets Gab (Zannettou et al., 2018), ICE (Greenbaum, 1991), and M2D2 (Reid et al., 2022) (Wiki)

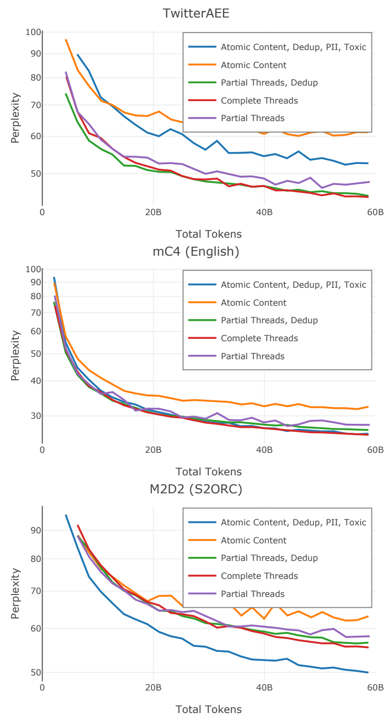


Figure 72: Perplexity results on Paloma (Magnusson et al., 2023); subsets Twitter AAE (Blodgett et al., 2016), mC4 (Xue et al., 2020) (English), and M2D2 (Reid et al., 2022) (S2ORC)

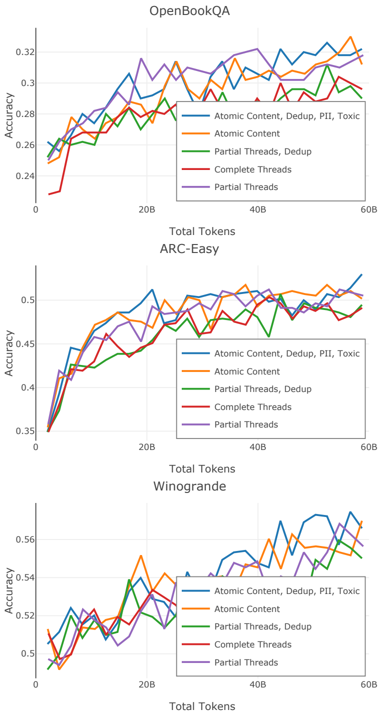


Figure 73: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

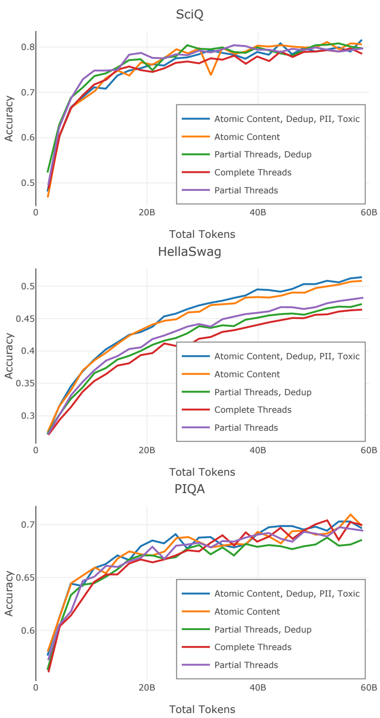


Figure 74: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

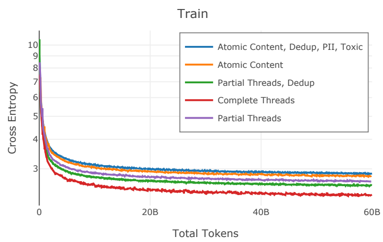


Figure 75: Training Cross Entropy

O.10 Evaluating Toxicity Filtering in Conversational Forums Pipeline

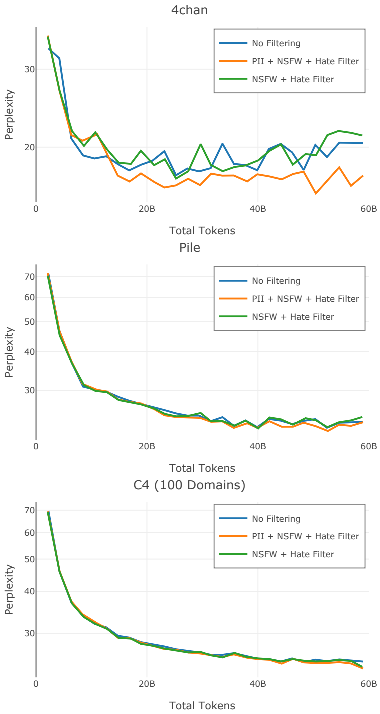


Figure 76: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Pile (Gao et al., 2020) (Val), and C4 100 dom (Chronopoulou et al., 2022)

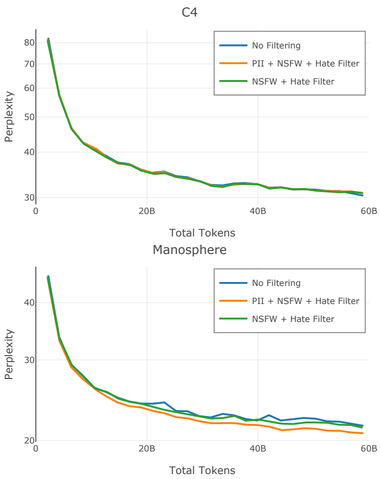


Figure 77: Perplexity results on Paloma (Magnusson et al., 2023); subsets C4 (Raffel et al., 2020; Dodge et al., 2021) and Manosphere (Ribeiro et al., 2021)

Figure 78: Perplexity results on Paloma (Magnusson et al., 2023); subsets Gab (Zannettou et al., 2018), ICE (Greenbaum, 1991), and M2D2 (Reid et al., 2022) (Wiki)

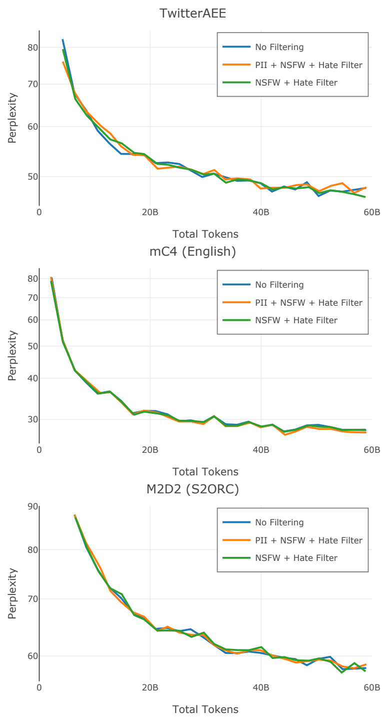


Figure 79: Perplexity results on Paloma (Magnusson et al., 2023); subsets Twitter AAE (Blodgett et al., 2016), mC4 (Xue et al., 2020) (English), and M2D2 (Reid et al., 2022) (S2ORC)

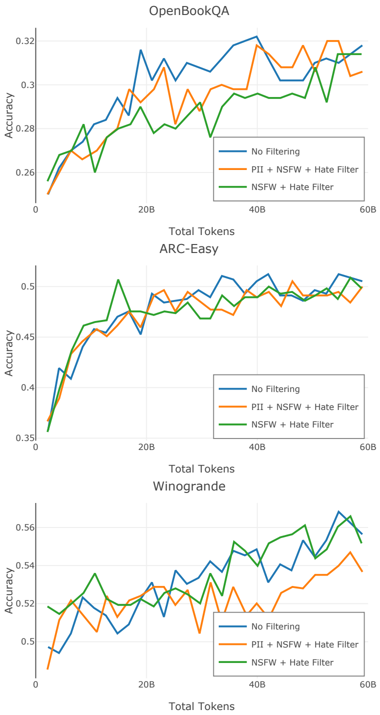


Figure 80: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

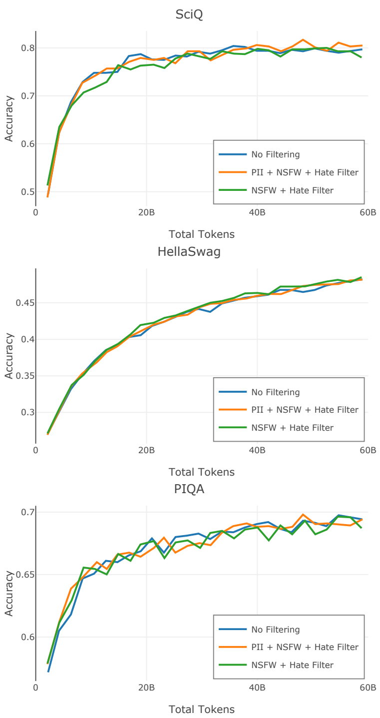


Figure 81: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

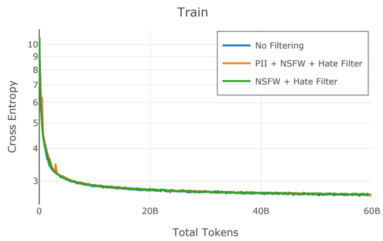


Figure 82: Training Cross Entropy

O.11 Training `OLMo-1B`

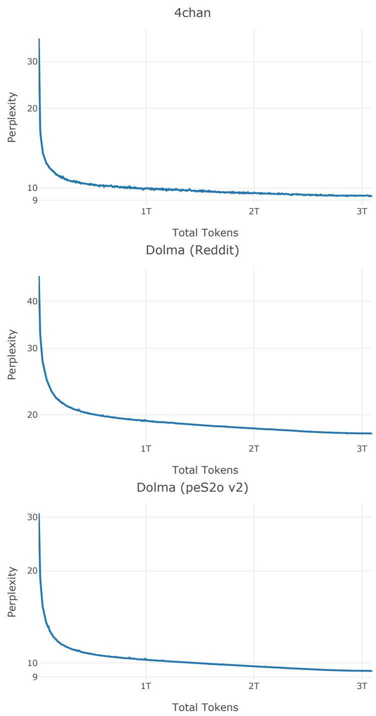


Figure 83: Perplexity results on Paloma (Magnusson et al., 2023); subsets 4chan (Papasavva et al., 2020), Dolma Reddit Subset, and Dolma Papers Subset

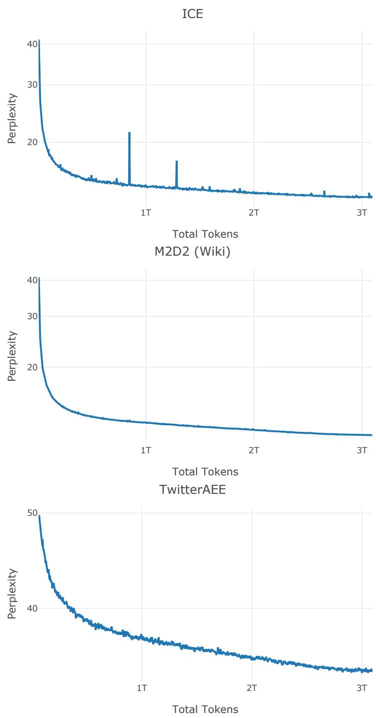


Figure 84: Perplexity results on Paloma (Magnusson et al., 2023); subsets ICE (Greenbaum, 1991), M2D2 (Reid et al., 2022) (Wiki), and Twitter AAE (Blodgett et al., 2016)

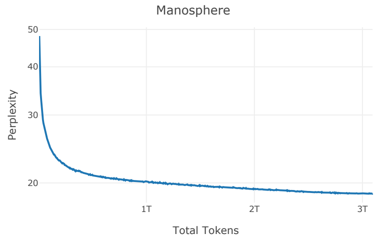


Figure 85: Perplexity results on Paloma (Magnusson et al., 2023); subsets Manosphere (Ribeiro et al., 2021)

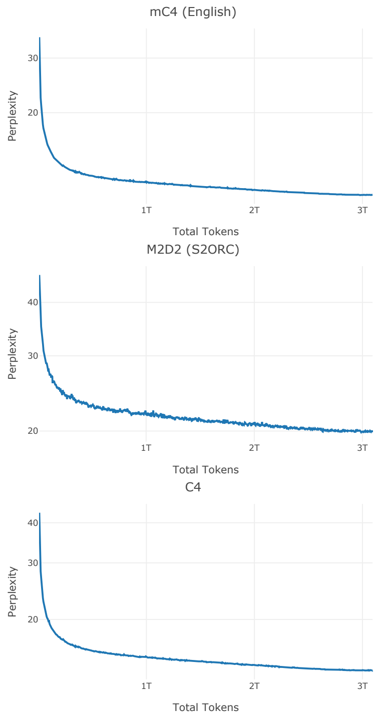


Figure 86: Perplexity results on Paloma (Magnusson et al., 2023); subsets mC4 (Xue et al., 2020) (English), M2D2 (Reid et al., 2022) (S2ORC), and C4 (Raffel et al., 2020; Dodge et al., 2021)

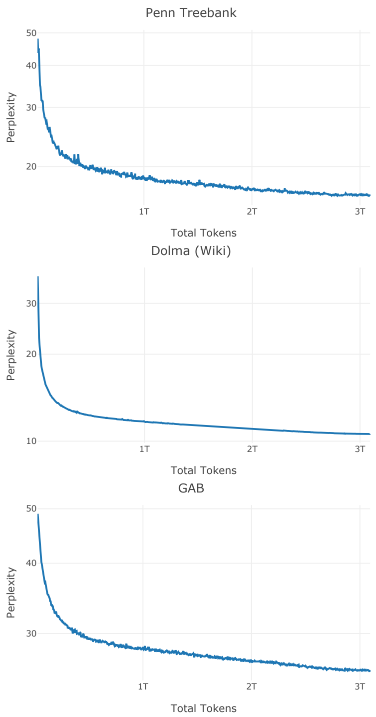


Figure 87: Perplexity results on Paloma (Magnusson et al., 2023); subsets Penn Tree Bank (Marcus et al., 1994), Dolma Wikipedia Subset, and Gab (Zannettou et al., 2018)

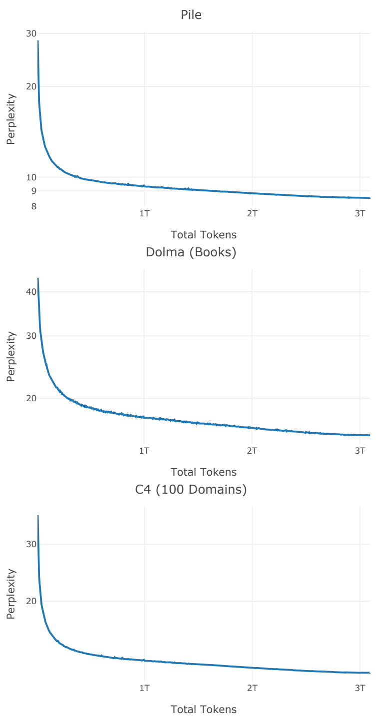


Figure 88: Perplexity results on Paloma (Magnusson et al., 2023); subsets Pile (Gao et al., 2020) (Val), Dolma Books Subset, and C4 100 dom (Chronopoulou et al., 2022)

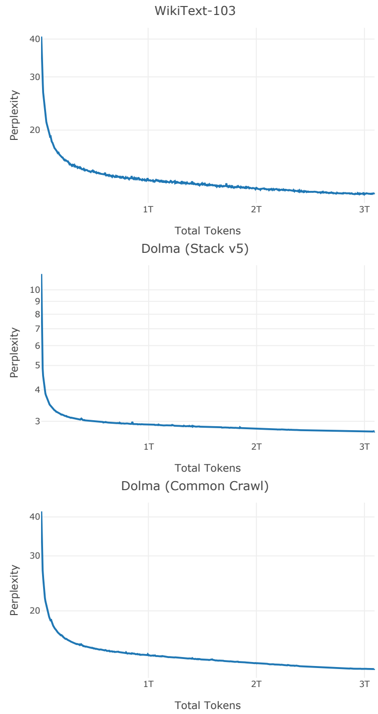


Figure 89: Perplexity results on Paloma (Magnusson et al., 2023); subsets WikiText 103 (Merity et al., 2016), Dolma Code Subset, and Dolma Web Subset

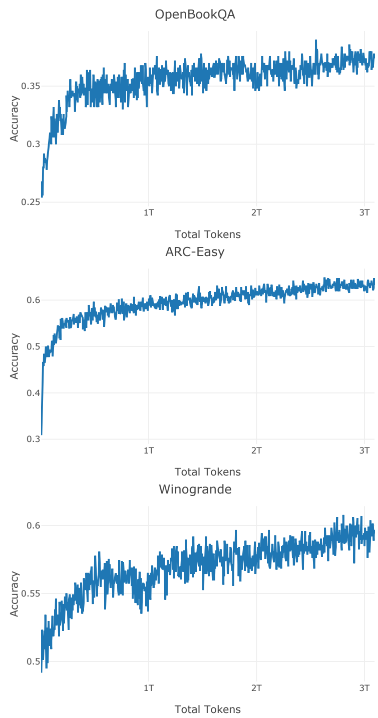


Figure 90: Results downstream tasks OpenBookQA (Mihaylov et al., 2018), ARC-E (Clark et al., 2018), and WinoGrande (Sakaguchi et al., 2019)

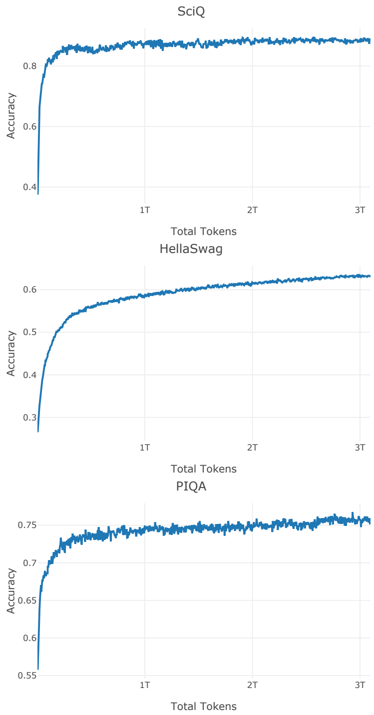


Figure 91: Results downstream tasks SciQ (Welbl et al., 2017), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2019)

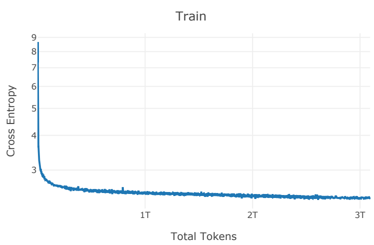


Figure 92: Training Cross Entropy