The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text

Nikhil Kandpal[*1,2] Brian Lester[*1,2] Colin Raffel[*1,2,3]
Sebastian Majstorovic[4] Stella Biderman[4] Baber Abbasi[4] Luca Soldaini[5] Enrico Shippole[6]
A. Feder Cooper[†7] Aviya Skowron[4] John Kirchenbauer[8] Shayne Longpre[9] Lintang
Sutawika[4,10] Alon Albalak[‡11] Zhenlin Xu[12] Guilherme Penedo[3] Loubna Ben allal[3] Elie
Bakouch[3] John David Pressman[4] Honglu Fan[4,13] Dashiell Stander[4] Guangyu Song[4] Aaron
Gokaslan[7] Tom Goldstein[8] Brian R. Bartoldson[14] Bhavya Kailkhura[14] Tyler Murray[5]

[1]University of Toronto [2]Vector Institute [3]Hugging Face [4]EleutherAI [5]The Allen Institute for Artificial Intelligence [6]Teraflop AI [7]Cornell University [8]University of Maryland, College Park [9]MIT [10]CMU [11]Lila Sciences [12]Independent [13]poolside [14]Lawrence Livermore National Laboratory

Abstract

Large language models (LLMs) are typically trained on enormous quantities of unlicensed text, a practice that has led to scrutiny due to possible intellectual property infringement and ethical concerns. Training LLMs on openly licensed text presents a first step towards addressing these issues, but prior data collection efforts have yielded datasets too small or low-quality to produce performant LLMs. To address this gap, we collect, curate, and release the Common Pile v0.1, an eight terabyte collection of openly licensed text designed for LLM pretraining. The Common Pile comprises content from 30 sources that span diverse domains including research papers, code, books, encyclopedias, educational materials, audio transcripts, and more. Crucially, we validate our efforts by training Comma v0.1, a 7 billion parameter LLM trained on 1 trillion tokens of text from the Common Pile. Comma attains competitive performance to LLMs trained on unlicensed text with similar computational budgets, such as LLaMA 7B. In addition to releasing the Common Pile v0.1 itself, we also release the code used in its creation as well as Comma v0.1's checkpoints and training mixture.

1 Introduction

A critical stage of large language model (LLM) development is pretraining [73, 138, 144], where an LLM is trained to predict the next token (i.e., word or subword unit) in a corpus of unstructured text. Pretraining is widely regarded as the foundation for strong downstream performance, as it enables LLMs to learn the structure of natural language [32, 111, 156] and accumulate a broad base of world knowledge [135, 154]. In an effort to push the capabilities of LLMs, pre-training datasets have grown steadily over time [145], with modern datasets containing trillions of tokens [134, 169, 195]. To meet this increasing demand for pre-training data, the *de facto* approach has been to leverage the public Internet as a source of text [57, 96, 109, 134, 144].

While the web provides a diverse and continuously growing supply of text, much of this content—under most legal frameworks—is protected by copyright. Yet, this text is routinely used to pretrain

[*]Equal contribution. For a list of author contributions, see Appendix A. [†]Work done while a graduate student at Cornell University. [‡]Work done while at SynthLabs.

Preprint.

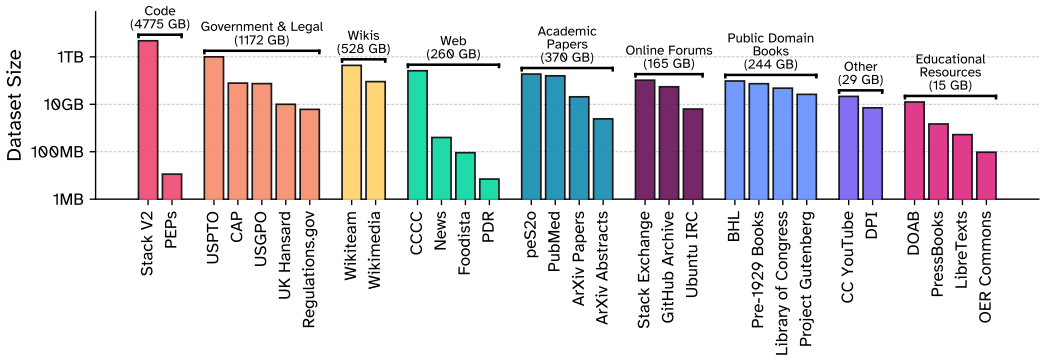

Figure 1: **The Common Pile is an 8TB dataset of openly licensed text curated from 30 diverse sources.** The sources comprising the Common Pile are shown above, categorized by textual domain.

LLMs, often without compensation to the creators of this content. Recent estimates suggest that compensating the authors of pre-training data, even at conservatively low wage rates, would cost billions of US dollars [83]. While copyright exemptions for text and data mining exist in some jurisdictions [70, 80, 93, 132, 158], many rights holders have objected to the uncompensated use of their work, resulting in numerous lawsuits against LLM developers [24, 193] that could carry financial damages in the billions [40, 97, 161]. Beyond questions of intellectual property (IP) law, the use of web-scraped data also raises ethical concerns [9], as content creators rarely explicitly consent to the downstream use of their work for LLM training. In fact, recent evidence suggests that many content owners may *not* consent to its use as LLM training data, as shown by a sharp mid-2023 increase in websites blocking AI crawlers [108], following growing awareness of web data being used to train models. Finally, while open models trained on publicly released pre-training datasets [18, 64, 104] support research into the study of learning dynamics [50, 77, 85], memorization [17, 22], data auditing [47, 130, 147], and more, the use of unlicensed training data heavily limits the ability of model trainers to share their datasets, and has previously resulted in DMCA takedowns of datasets such as the Pile [57].

The current landscape reflects a growing divide between LLM developers and content creators. We submit that a natural first step toward resolving this tension is to ask: *Is it possible to train performant language models using only public domain and openly licensed text?* We define "openly licensed" text as content that follows the Open Knowledge Foundation's Open Definition 2.1 (further detailed in Section 2 and Appendix C), which refers to content where the copyright holder has granted explicit permission for the content to be freely accessed, used, modified, and shared for any purpose. Our primary contribution in this paper is to demonstrate that this is indeed possible by collecting, curating, and releasing *the Common Pile v0.1*, an 8TB dataset that—to our knowledge—constitutes the largest collection of openly licensed text to date. The Common Pile comprises 30 text sources (detailed in Section 3), covering diverse domains including research publications, open-source code, government documents, historical books, educational resources, audio transcripts, and more. Crucially, we demonstrate that after appropriate filtering, deduplication, and reweighting, the Common Pile v0.1 can be used as the foundation for a competitive LLM: *Comma v0.1*, a 7-billion-parameter model with comparable performance budget-matched models trained on unlicensed datasets such as LLaMA-1 7B, MPT 7B, and RedPajama-INCITE 7B. In the spirit of openness and transparency, we release the Common Pile v0.1, Comma v0.1 and its filtered and deduplicated pre-training dataset, and all data collection and processing code.

2 What do we mean by "openly licensed"?

Copyright law grants content creators certain rights, such as exclusive rights (with certain exceptions) to reproduce, distribute, and create derivatives of their original works. Although copyright laws vary across jurisdictions, original, creative works (that are "fixed" in a tangible medium, such as physically or digitally [see, e.g., 1]) typically fall within the scope of copyright. Works in the *public domain* [38] have had their copyrights expire (after a legally dictated time period), were never eligible for copyright protection due to specific carve-outs (e.g., government documents in the U.S. [2]), or were otherwise dedicated to the public domain by their copyright owners (e.g., with a CC0 license [35]). Copyright

owners can *license* their protected works, allowing others to adapt and reuse them under specified terms. For example, Creative Commons (CC) Licenses (except CC0) grant the right to "reproduce and Share the Licensed Material, in whole or in part; and produce, reproduce, and Share Adapted Material" [36]. For a more in-depth and accessible discussion about licenses and generative AI, see Lee et al. [97, Parts II.I–II.J].

For the Common Pile, we collect and curate public domain and openly licensed text, where we consider "openly licensed" to mean any license that meets the Open Knowledge Foundation's Open Definition 2.1. Some prominent examples of licenses that are considered to be "open" under this definition include CC BY [37], CC BY-SA [39], and software licenses certified by the Blue Oak Council (e.g., the MIT license) [20]. We note that CC NC (non-commercial) and ND (no derivatives) licenses are not considered open under this definition and we therefore do not include content distributed under these licenses. While the use of an open license does not necessarily imply that the rights holder has specifically contemplated use of their content to train LLMs, most open licenses include text like "the above rights may be exercised in all media and formats whether now known or hereafter devised" [37]. Overall, we consider our use of openly licensed data to be a substantial first step towards ethical pre-training dataset curation.

2.1 License due diligence

License laundering There is a large quantity of data on the internet with incorrect, ambiguous, or missing licensing metadata [97, 107]. A common pitfall is "license laundering," where a copyrighted work is redistributed (typically by a non-rights holder) with an incorrect license. License laundering can undermine our ability to confidently source openly licensed content since it implies that we cannot always trust the license distributed with a piece of content. To address this issue, we set strict standards for data sourcing, only including data from sources where we were confident that the licensing information was provided by the copyright holder, which ultimately led us to exclude certain sources such as OpenAlex [81, 128], YouTube Commons [75], and the Hacker News dataset on Kaggle.

Use of collection licenses A related issue is the licensing status of compilations of existing works. Many training corpora are released under open licenses, but these licenses do not necessarily align with the licensing status of the underlying documents [97, Part II.A]. As an example, the ODC-By license has been commonly used for large-scale web corpora such as Dolma [169], FineWeb [134], and TxT360 [175]. ODC-By, by definition, does not extend to *individual documents* within the corpus; therefore, the copyright of documents in these collections is still controlled by the document authors, and does *not* imply that the text itself is openly licensed.

LLM-generated synthetic datasets Datasets containing text generated by LLMs trained on unlicensed data have been released under open licenses [e.g. 197]. It has not yet been established whether it is permissible to apply arbitrary licenses to the generations of an LLM that was trained on unlicensed data [97]. We therefore take a conservative stance and avoid synthetic content that was generated by an LLM.

Caveats Despite our best efforts at due diligence, data that falls outside of our curatorial principles and choices may have still ended up in our dataset. License laundering is a notoriously hard problem to identify exhaustively in practice [97]. Copyright owners may also change the license they associate with their content. Since we collected and curated the Common Pile v0.1 in late 2024, the licensing information we include and rely on may not be completely aligned with more recent updates. Further, some *documents* that we collect that are in the public domain or are openly licensed may contain material with unclear status (e.g., quoted snippets of in-copyright books in public domain U.S. government publications). Finally, we note that while it is relatively straightforward to obey attribution requirements when redistributing data, attributing model predictions back to the training data points that influenced them remains an active area of research [131, 28].

2.2 Comparisons with related work

Our work is not the first that aims to construct a dataset of openly licensed and/or public domain data for the purposes of training machine learning models. Past efforts include CommonCanvas [61], a collection of approximately 70 million Creative Commons-licensed images designed for training image generation models, the PG19 dataset [142] of public domain novels sourced from Project

Gutenberg used for benchmarking language models, the C4Corpus tools for sourcing Creative Commons text from Common Crawl snapshots [68], and many datasets comprising CC BY-SA-licensed text from Wikipedia [66, 116].

More relevant to our work are the recent Open License Corpus (OLC) [121], Common Corpus [75, 92], and KL3M [79] datasets, which were constructed for use as LLM pre-training data. On the whole, OLC uses similar selection criteria to ours, including text that is in the public domain or is openly licensed. However, OLC also includes conversations scraped from Hacker News, which does not have an open license. Additionally, OLC is considerably smaller than the Common Pile v0.1, comprising data from 12 sources (vs. 30 for Common Pile) totaling 0.85 TB of text (vs. 7.6 TB for Common Pile). Common Corpus also uses a similar set of allowable licenses/copyright statuses (e.g. CC BY, CC BY-SA, public domain, MIT-style, etc.) although the specific licenses/statuses are not clear because Common Corpus does not retain full per-document licensing information across all sources. Additionally, Common Corpus incorporates data from OpenAlex [128] which is known to provide inaccurate licensing information [e.g., 81]. Furthermore, while the Common Pile and Common Corpus are similar in size (7.6 TB vs. 7.4 TB), Common Corpus targets a broader set of languages and therefore contains significantly less English text. Conversely, KL3M does not consider CC BY-SA to be acceptable and, as a result, almost exclusively consists of government documents. Accordingly, the Common Pile is much larger than KL3M (3 TB), and is built from significantly more diverse data sources (Figure 1 & Section 3). In Section 4.3, we compare the Common Pile v0.1 to these datasets in a controlled setting, ultimately showing that it produces substantially more performant LLMs.

3 Composition of the Common Pile

The Common Pile comprises content drawn from a wide range of domains, including scholarly publications, government documents, online discussions, books, open educational resources, and more. In this section, we provide an overview of each of the domains contained in the Common Pile and briefly discuss their constituent data sources. In-depth discussion of each source is provided in Appendix B.

Scientific and scholarly texts, which are often distributed under open licenses due to open access mandates, appear in many LLM pre-training datasets [e.g. 57, 169, 187] since they expose models to technical terminology, formal reasoning, and long-range document structure. To attain broad coverage of scholarly text, we filter peS2o [168] (a collection text extracted from open-access scientific PDFs based on S2ORC [105]) to only retain openly licensed research papers. For medical-domain text, we collect text from openly licensed articles in the U.S. National Institutes of Health's National Library of Medicine's PubMed Central archive. Additionally, we collect data from ArXiv, which contains over 2.4 million articles in the quantitative sciences, most of which are uploaded as LaTeX source and may be distributed under various licenses chosen by a given article's author. We include openly licensed articles sourced from ArXiv's bulk-access S3 bucket and parsed using LaTeXML and Trafilatura [10]. Furthermore, according to ArXiv's licensing policy, all metadata (including abstracts) of articles posted to ArXiv are distributed under the CC0 license; we therefore include the abstracts for *all* ArXiv papers in the Common Pile, regardless of the paper's full-text license.

Online discussion forums comprise multi-turn question-answer pairs and discussions and therefore can be useful for training language models to follow conversational structure as well as for improving performance on question answering and dialogue-centric tasks. StackExchange is a collection of websites that host user-provided questions and answers and allow their redistribution under a CC BY-SA license. We leverage the user-provided StackExchange dumps from the Internet Archive and format questions/answers in the same order they appear on StackExchange, using PyMarkdown to convert each comment into plain text. Additionally, we collect text from issues, pull requests, and comments on GitHub, which, according to GitHub's terms of service, inherit the license of their associated repository. We extract this content from repositories with Blue Oak Council-approved licenses from the GitHub Archive. Finally, we include logs of all discussions on the Ubuntu-hosted Internet Relay Chat (IRC) since 2004, which are released into public domain.

Government and legal texts are often published directly into the public domain or under open licenses. For example, in the US, text written by federal government employees is considered to be in the public domain. We therefore include all plain-text documents made available through the United

States Government Publishing Office (USGPO)'s GovInfo.gov developer API. Additionally, we include all plain text regulatory documents published by U.S. federal agencies from Regulations.gov, an online platform that hosts newly proposed rules and regulations from federal agencies. The Common Pile also incorporates US Patents and Trademark Office (USPTO) patent documents sourced from the Google Patents Public Data dataset [78], containing millions of public domain patents and published patent applications dating back to 1782. Similarly, the Hansard (the official record of parliamentary proceedings) of the United Kingdom is distributed under the Open Parliament License, which stipulates similar terms to the CC BY license. We source UK Hansard data from ParlParse [133], covering Commons debates from 1918 forward and Lords proceedings from the 1999 reform. For legal text, we leverage the Caselaw Access Project (comprising 40 million pages of U.S. federal and state court decisions and judges' opinions from the last 365 years) and Court Listener (including 900 thousand cases scraped from 479 courts). Only legal texts in the public domain were selected for the Common Pile.

Curated task datasets are typically designed for fine-tuning on specific downstream tasks such as question answering, summarization, or text classification. To source datasets that are distributed under an open license and only contain content owned by the dataset's rights holder (to avoid license laundering), we use metadata and redistributed datasets from the Data Provenance Initiative [107, 110]. Full details on the datasets we include are available in Appendix D.

Books, particularly historic text, can fall into the public domain due to copyright expiration—for example, in the United States, books published prior to 1929 are currently in the public domain. We source public domain books from various sources, including the Biodiversity Heritage Library (BHL), an open-access digital library for biodiversity literature and archives; pre-1929 books digitized by the Internet Archive on behalf of HathiTrust member libraries; the collection of public domain books called "Selected Digitized Books" released by the Library of Congress; and select books from Project Gutenberg, an online collection of over 75,000 digitized books, most of which are in the public domain.

Open Educational Resources (OERs) are educational materials (e.g. textbooks, lecture notes, lesson plans, etc.), typically published under Creative Commons licenses that support free and equitable access to education. We collect data from multiple OER repositories, including the Directory of Open Access Books (DOAB), an online index of over 94,000 peer-reviewed books curated from trusted open-access publishers; PressBooks, a searchable catalog of over 8,000 open access books; OERCommons, an online platform where educators share open-access instructional materials; and LibreTexts, a catalog of over 3,000 open-access textbooks.

Wikis are topic- or domain-specific encyclopedic websites that are collaboratively written, maintained, and moderated. Historical and cultural precedent has led many wikis to have an open license. We downloaded the official database dumps of wikitext (Mediawiki's custom markup language) of the English-language wikis that are directly managed by the Wikimedia foundation and converted wikitext to plain text using `wtf_wikipedia`. For wikis not managed by Wikimedia, we make use of wikiteam's unofficial database dumps and apply the same conversion process.

Source code has proven to be a useful part of LLM pre-training corpora, not only to support coding abilities but also to improve reasoning [7, 114, 122]. Due to the Free and Open Source Software (FOSS) movement, a great deal of source code is distributed with an open license. We leverage prior work done by the Software Heritage Foundation and BigCode to compile the openly licensed subset of the Stack V2 [114], based on the license detection performed by the creators of Stack V2. Additionally we collected all Python Enhancement Proposals (PEPs)—design documents that generally provide a technical specification and rationale for new features of the Python programming language—that were released into the public domain.

YouTube allows users to upload content under a CC BY license. We therefore sourced and transcribed speech-heavy CC BY videos from YouTube. To avoid license laundering and focus on high-quality speech-based textual content, we manually curated a set of over 2,000 YouTube channels that release original openly licensed content containing speech. From these channels, we retrieved and transcribed (using Whisper [140]) over 1.1 million openly licensed videos comprising more than 470,000 hours of content.

Web text is a common source of LLM pre-training data. A small fraction of content on the web is distributed under open licenses. To recover a portion of this content, we process 52 Common

Crawl snapshots using a regular expression (regex) adapted from the C4Corpus project [68] to retain pages that include a CC BY, CC BY-SA, or CC0 marker. This regex naturally results in many false positives (e.g., it would retain a page that included and provided attribution for a CC BY image but otherwise contained unlicensed content), so we manually verified the top 1000 domains by content volume, retaining only those for which all content was assigned a Creative Commons license. Text was extracted using a pipeline similar to the one used for Dolma [169]. We provide more details on the composition of our web-sourced text, called CCCC, in Appendix G. Apart from CCCC, we additionally manually collected data from a few select sites, including Foodista, a community-maintained site with recipes and food-related news as well as nutrition information; news sites that publish content under CC BY or CC BY-SA according to Open Newswire; and the Public Domain Review, an online journal dedicated to exploration of works of art and literature that have aged into the public domain.

4 Assessing the Common Pile v0.1's quality

The utility of an LLM pre-training dataset is mostly assessed in terms of whether or not it can be used to train performant LLMs. To validate our efforts in curating the Common Pile, we use it as the basis of an LLM pre-training dataset created through additional filtering (Section 4.1) and rebalancing (Section 4.2). Then, we perform a controlled data ablation study (Section 4.3) where we train otherwise-identical LLMs on different pre-training datasets, including prior datasets comprised of openly licensed text mentioned in Section 2.2 as well as a selection of representative pre-training datasets of unlicensed text. Finally, we train Comma v0.1, a 7 billion parameter LLM trained on 1 trillion tokens of Common Pile-sourced content, and compare it to models with a similar parameter count and training budget that were trained on unlicensed text (Section 4.4).

4.1 Dataset preprocessing and filtering

Before training a language model, it is considered important to "clean" data in hopes of retaining only high-quality text under some notion of quality [4, 106]. Consequently, before training on data from the Common Pile (which is distributed in a relatively "raw" format), we independently preprocessed each of the Common Pile's non-code datasets using pipelines implemented with the Dolma data processing toolkit [169].

Since the Common Pile v0.1 focuses primarily on English content, we apply **language identification** using a FastText classifier [82] to filter out non-English text. When processing web text from CCCC, we employ the **text quality classifier** adapted from DataComp-LM [100] with an extremely low threshold to remove noisy text. We remove documents with pervasive OCR errors using the **likelihood-based filtering** approach from [168], which removes documents that are assigned an excessively low log-likelihood under a unigram language model constructed from the Trillion Word Corpus [119]. To reduce the prevalence of toxic or inappropriate content, we apply a pair of FastText **toxicity classifiers** implemented in Dolma [169] that were trained on the Jigsaw Toxic Comment Classification Challenge dataset [30]. We apply regex-based **personally identifiable information (PII) redaction** to remove email addresses, phone numbers, and IP addresses, and replace them with `<EMAIL_ADDRESS>`, `<PHONE_NUMBER>`, and `<IP_ADDRESS>` respectively. Finally, we perform source-specific **regex filtering** to remove repetitive or boilerplate text (e.g., page numbers, document preambles, license statements, etc.). For a detailed breakdown of the pre-processing applied to each dataset, see Table 5 (appendix).

After filtering, we perform global document-level fuzzy deduplication across all sources, as excessive data duplication is known to harm language modeling performance [95] and increase memorization [84]. We use the bloom filter-based deduplication functionality from Dolma [169] and deem two documents duplicates if they share more than 90% of their 20-grams.

For code data from the Stack v2, we apply the Red Pajama V1 [187] code filtering heuristics. These include filters based on the mean and maximum line length in a document, the proportion of alphanumeric characters, and the ratio of alphabetical characters to tokens. After this initial filter, we adopt the process used by SmolLM2 [5] where we keep only code in Python, C, C++, SQL, Java, PHP, Rust, Javascript, Typescript, Go, Ruby, Markdown, C#, Swift, or shell and filter this set using language-specific quality classifiers to retain only educational and well-documented code. We use a lower threshold to filter out low-quality code than was used for SmolLM2, resulting in a larger

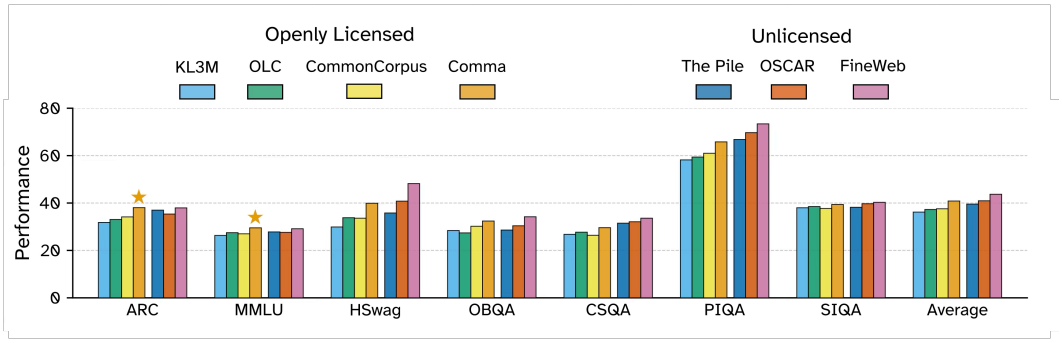

Figure 2: **The Common Pile consistently outperforms other openly licensed corpora as a pre-training dataset.** Following the setup from Penedo et al. [134], we train and evaluate 1.7B parameter models on 28B tokens of data from each dataset. Stars denote benchmarks on which the model trained using the Common Pile outperforms all other models.

set of post-filtered text. Finally, we extract plaintext from HTML documents in the Stack V2 using Trafilatura [10] and apply our standard plaintext filtering pipeline including language, length, toxicity, and PII filtering.

4.2 Data mixing

Recent work [3, 180, 191] has shown that up- or down-weighting pre-training data sources in accordance with some notion of data quality can produce more performant models. Indeed, the sources in the Common Pile vary drastically in their characteristics, and we don't necessarily expect that our largest sources contain the highest quality text. For example, patent text sourced from the USPTO (our second-largest source) exhibits substantially different wording, terminology, and repetition than typical natural language. Consequently, we anticipate that appropriately mixing the sources in the Common Pile (rather than simply combining all sources, i.e., mixing in proportion to source size) is of particular importance. Additionally, while LLM pre-training datasets have been continuously scaled to avoid the diminishing returns that result from repeating data [95], recent work has highlighted that repeating high-quality data can be preferable to avoiding repetition by training on low-quality data [122, 52].

To determine mixing weights, we first trained per-source language models using the procedure outlined in Section 4.3 below for 28 billion tokens on all sources that were sufficiently large to be repeated less than four times at this data budget. Based on the performance of these per-source models, we heuristically set mixing weights to up- and down-weight high- and low-performance sources respectively while targeting a maximum of six repetitions over the course of a 1 trillion token training run. Additionally, we assumed that our smaller sources were high quality and set their mixing rates such that they were also repeated six times over the course of 1 trillion tokens. The resulting mixture and per-source repetition rates are given in Table 7 (appendix). We also experimented with using MixMin [180] to automatically determine mixing weights but found that it did not improve over our heuristically determined mixture.

Because we use this dataset mixture to train Comma v0.1 7B (Section 4.4) and because it comprises a heavily filtered and remixed version of the Common Pile v0.1, we refer to it as the "Comma dataset" to distinguish it from the Common Pile itself.

4.3 Controlled dataset quality experiments

As a preliminary measure of the Common Pile's quality, we adopt the experimental setting of Penedo et al. [134] and identically train models on the Comma dataset and various preexisting datasets. By using a controlled setting across datasets, we can assert that differences in model performance stem primarily from the quality of each dataset. Specifically, we train 1.7 billion parameter decoder-only Transformer models [184] that follow the Llama architecture [181] on 28 billion tokens of data from each dataset, tokenized using the GPT-2 tokenizer [139]. We follow the hyperparameters and setup of Penedo et al. [134] exactly, except that we used a weight decay of 0.2 instead of 0.1 due to slightly improved performance (possibly due to the large amount of repetition in the Comma dataset).

Each model was then evaluated using the set of "early signal" tasks identified by Penedo et al. [134] which cover commonsense reasoning and knowledge capabilities; specifically, we evaluate zero-shot performance on ARC [33], MMLU [71], HellaSwag (HSwag) [194], OpenBookQA (OBQA) [120], CommonSenseQA (CSQA) [173], PIQA [19], and SocialIQA (SIQA) [162]. We omit Winogrande because it is included in the set of datasets we sourced from the Data Provenance Initiative. We highlight that a significant portion of the Comma dataset is code, but none of the tasks we evaluate on measure code capabilities. While it is possible that we could improve performance by omitting code data in this setting, we retained code for reliable reporting of the Comma dataset's performance.

As baselines, we compare to the prior datasets that aim to provide open licensed text discussed in Section 2.2: OLC [121], Common Corpus [92], and KL3M [79]. We additionally compare to the Pile, as it one of the only LLM pre-training datasets that contains a comparable number of diverse sources to the Common Pile (22 vs. 30). Finally, we report the performance of two web text-based unlicensed pre-training datasets: OSCAR [171], which incorporates relatively little filtering; and FineWeb [134], an recent dataset that reflects current best practice for LLM pre-training dataset curation.

The resulting performance of each model is shown in Fig. 2, with detailed results in Table 9 (appendix). Notably, the Comma dataset-based model outperforms the models trained OLC, Common Corpus, and KL3M across all benchmarks and outperforms the Pile-based model on all but two benchmarks. While the performance of the FineWeb-based model is the best on most benchmarks, the Comma dataset-based model performs best on the scientific and scholarly knowledge-based benchmarks MMLU and ARC, possibly due to the Common Pile's large proportion of domain-relevant text. On the other hand, on the commonsense reasoning datasets HellaSwag, PIQA, and CommonSenseQA, the model trained on the Comma dataset has significantly worse performance than models trained on the Pile, OSCAR, and FineWeb, possibly indicating a lack of relevant data in the Common Pile. We additionally note that recent work [190] highlights that performance on HellaSwag is most heavily influenced by coverage of certain domains and topics such as personal blogs, tutorials, hobbies, and sports, which are poorly represented in the Common Pile. Overall, these findings confirm that the Comma dataset performs best among datasets that aim to contain only openly licensed data and is also a strong candidate in general, particularly when targeting scientific and scholarly applications.

We note that the Comma dataset is the only dataset we evaluate that explicitly includes task-like data due to inclusion of data from the Data Provenance Initiative (DPI). To verify that this does not confer an unfair advantage, we trained an additional model on the Comma dataset with all sources retained except for the DPI-sourced data. Removing this source had a minimal impact on model performance (full results in Table 9), with a notable decrease only on HellaSwag, possibly suggesting that the DPI data contains domain-relevant data for this benchmark that other sources lack.

4.4 Comma v0.1

Having established that Comma's dataset produces models with competitive performance when compared to other datasets, we now validate our efforts at a larger, more realistic scale. Specifically, we train Comma v0.1, a 7 billion-parameter model trained on 1 trillion tokens of text, and compare with other models trained using a similar computational budget.

Tokenization While training a tokenizer on unlicensed text is less likely to raise ethical or IP-related issues than training an LLM, we nevertheless trained a custom tokenizer on the Comma dataset to ensure that our entire modeling pipeline was based on openly licensed data. In addition, the different characteristics of our dataset likely makes existing tokenizers (which are often trained on web text) suboptimal. We therefore trained a BPE-based [55] tokenizer using the Hugging Face `tokenizers` library using a vocabulary size of 64,000. We follow the same splitting regex as Llama 3.2 [62] and the Hugging Face `ByteLevel` preprocessor; no Unicode normalization was used. The tokenizer was trained on a 600GB sample [152] of text from the Comma dataset.

Training setup We trained Comma v0.1 using the `lingua` framework [185]. We base our model architecture and training hyperparameters on `lingua`'s Llama-7B configuration, which closely follows the conventions set by the Llama series of models [181, 62]. We trained with an effective batch size of 512 length-4096 sequences using the AdamW [112] optimizer and a weight decay of 0.2. We performed two stage training, with a first stage following a cosine learning rate schedule for 460,000 steps with 2,000 steps of warmup, an initial learning rate of $1e{-}3$, a minimum learning rate of $1e{-}9$, and a period of 500,000 steps. In the second stage, we performed a 18,000-step "cool-down" [74], where we train only on a subset of high-quality sources using the mixing weights provided

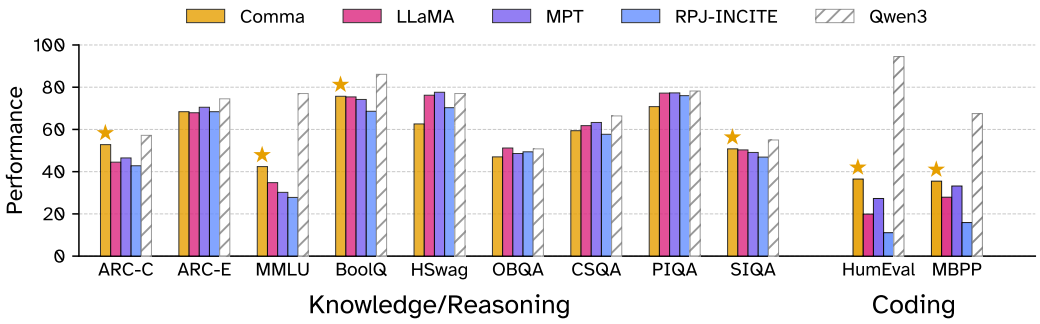

Figure 3: **Compared to models trained with similar resources (7 billion parameters, 1 trillion tokens), Comma is the strongest model on several standard benchmarks.** To contextualize these results, we include Qwen3 8B (trained on 36 trillion tokens) as a "current best-practices" upper bound. Stars denote benchmarks on which Comma outperforms all other compute-matched models (i.e., all models other than Qwen3)

in Table 8 (appendix) while decaying the learning rate linearly to 0. Finally, we average together ten checkpoints from the cool-down phase to produce a final model as suggested by Grattafiori et al. [62]. The full training run was completed in 18 days using 64 H100 GPUs. Apart from our main Comma v0.1 training run, we completed several additional runs to better understand how hyper-parameters impact the model, including using a larger batch size, following a three-stage (rather than two-stage) curriculum, and training for a longer budget. Overall, the results of these runs were consistent with the findings from our main training run. Additional details can be found in Appendix O.

Evaluation We evaluate Comma 7B on the suite of benchmarks used by Groeneveld et al. [64] in addition to two additional code benchmarks. Specifically, we evaluate models on ARC [33], MMLU [71], BoolQ [31], HellaSwag [194], OpenBookQA [120], CommonsenseQA [173], PIQA [19], and SIQA [162] to probe world knowledge and reasoning and HumanEval [25] and MBPP [8] to evaluate coding capabilities. Following Groeneveld et al. [64], we evaluate using OLMES [65], using a zero-shot format for all tasks except MMLU, which uses a 5-shot format.

Baseline models For fairness, we primarily compare to prior models with the same parameter count and token budget. Since we are not aware of any such models trained on openly licensed data, we compare only to models trained on unlicensed data: specifically, LLaMA 1 7B [181], MPT-7B [177], RPJ-INCITE-7B [187], StableLM-7B [12], and OpenLLaMA-7B [58]. Over time, the token budgets of open pre-trained LLMs have continually grown [145], and current standard practice is to pretrain on significantly more than 1 trillion tokens. Consequently, recent models tend to outperform our baselines, which were released in 2023 and 2024. To provide a state-of-the-art point of reference, we additionally include results for the recently released Qwen3 8B [178], which was trained for 36 trillion tokens. We emphasize that we cannot reliably compare to a model with a 36× larger training budget and we primarily include it as a point of reference.

Results As shown in Fig. 3, Comma 7B outperforms budget-matched baseline models on over half of the benchmarks tested. In line with our results from Section 4.3, we observe that Comma 7B excels on knowledge-based benchmarks like ARC-C and MMLU, but lags behind on HellaSwag and PIQA. Comma 7B is also particularly strong at code-related tasks where it outperforms baseline models by a wide margin. Comparisons to StableLM and OpenLLama can be found in Table 10 (appendix), but show similar trends. Qwen3 8B's superior performance across all benchmarks confirms the benefit of larger training budgets and motivates future efforts on scaling up the Common Pile.

5 Conclusion

We release *Common Pile v0.1*, an 8TB corpus that—to our knowledge—constitutes the largest dataset built exclusively from openly licensed text. Alongside our dataset, we release *Comma v0.1*, a highly performant 7-billion-parameter LLM trained on text from the Common Pile, as well as the filtered and rebalanced data mixture we used for training. Our results demonstrate that not only is the Common Pile the strongest dataset for pretraining under an open-license constraint, but also that it produces models comparable to those trained on an equivalent amount of unlicensed data. This positive result

holds promise for future of open-license pretraining, especially if the research community invests in collecting larger quantities of openly licensed text data in the future. Ultimately, we believe that the Common Pile v0.1 represents the first step on the path towards a more ethical language model ecosystem, where performance need not come at the cost of creator rights and legal transparency.

Acknowledgments

We thank Chris Maddison, Anvith Thudi, Pierre-Carl Langlais, Alec Radford, Adam Roberts, Sewon Min, and Weijia Shi for fruitful discussions and constructive feedback. An early draft of this work was shared at the Dataset Convening hosted by the Mozilla Foundation and EleutherAI. We thank the participants for their discussion and feedback.

This work was supported by funding from the Mozilla Foundation and Sutter Hill Ventures. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC). Researchers funded through the NSERC-CSE Research Communities Grants do not represent the Communications Security Establishment Canada or the Government of Canada. Any research, opinions or positions they produce as part of this initiative do not represent the official views of the Government of Canada.

Parts of this work were performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 and was supported by the LLNL-LDRD Program under Project No. 24-ERD-010 and Project No. 24-SI-008 (LLNL-CONF-2006420).

References

[1] 17 U.S. Code § 102. Subject matter of copyright: In general, December 1990. URL https://www.law.cornell.edu/uscode/text/17/102.

[2] 17 U.S. Code § 105. Subject matter of copyright: United States Government works, December 2024. URL https://www.law.cornell.edu/uscode/text/17/105.

[3] Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training. *arXiv preprint arXiv:2312.02406*, 2023.

[4] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. A survey on data selection for language models. *Transactions on Machine Learning Research*, 2024.

[5] Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

[6] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In *Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, 2019.

[7] Viraat Aryabumi, Yixuan Su, Raymond Ma, Adrien Morisot, Ivan Zhang, Acyr Locatelli, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. To code, or not to code? exploring impact of code in pre-training. *arXiv preprint arXiv:2408.10914*, 2024.

[8] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

[9] Stefan Baack, Stella Biderman, Kasia Odrozek, Aviya Skowron, Ayah Bdeir, Jillian Bommarito, Jennifer Ding, Maximilian Gahntz, Paul Keller, Pierre-Carl Langlais, Greg Lindahl, Sebastian Majstorovic, Nik Marda, Guilherme Penedo, Maarten Van Segbroeck, Jennifer Wang, Leandro von Werra, Mitchell Baker, Julie Belião, Kasia Chmielinski, Marzieh Fadaee, Lisa

Gutermuth, Hynek Kydlíček, Greg Leppert, EM Lewis-Jong, Solana Larsen, Shayne Longpre, Angela Oduor Lungati, Cullen Miller, Victor Miller, Max Ryabinin, Kathleen Siminyu, Andrew Strait, Mark Surman, Anna Tumadóttir, Maurice Weber, Rebecca Weiss, Lee White, and Thomas Wolf. Towards best practices for open datasets for llm training, 2025. URL https://arxiv.org/abs/2501.08365.

[10] Adrien Barbaresi. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.acl-demo.15.

[11] Max Bartolo, A. Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678, 2020.

[12] Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. Stable lm 2 1.6 b technical report. *arXiv preprint arXiv:2402.17834*, 2024.

[13] Jonathan Berant, A. Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, 2013.

[14] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In Leif Azzopardi, Allan Hanbury, Gabriella Pasi, and Benjamin Piwowarski, editors, *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, March 2018. Springer.

[15] Janek Bevendorff, Martin Potthast, and Benno Stein. FastWARC: Optimizing Large-Scale Web Archive Analytics. In Andreas Wagner, Christian Guetl, Michael Granitzer, and Stefan Voigt, editors, *3rd International Symposium on Open Search Technology (OSSYM 2021)*. International Open Search Symposium, October 2021.

[16] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

[17] Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090, 2023.

[18] Stella Biderman, Hailey Schoelkopf, Quentin 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. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL https://arxiv.org/abs/2304.01373.

[19] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

[20] Blue Oak Council. License List (version 15), 2025. URL https://blueoakcouncil.org/list.

[21] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. In *Neural Information Processing Systems*, pages 9560–9572, 2018.

[22] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[23] Ilias Chalkidis, Abhik Jana, D. Hartung, M. Bommarito, Ion Androutsopoulos, D. Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Annual Meeting of the Association for Computational Linguistics*, pages 4310–4330, 2021.

[24] Chat GPT Is Eating the World, 2024. URL https://chatgptiseatingtheworld.com.

[25] 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. Evaluating large language models trained on code, 2021.

[26] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.91. URL https://aclanthology.org/2020.findings-emnlp.91/.

[27] Zhiyu Chen, Wenhu Chen, Hanwen Zha, Xiyou Zhou, Yunkai Zhang, Sairam Sundaresan, and William Yang Wang. Logic2text: High-fidelity natural language generation from logical forms. *ArXiv*, abs/2004.14579, 2020.

[28] Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, Jeff Schneider, Eduard Hovy, Roger Grosse, and Eric Xing. What is your data worth to gpt? llm-scale data valuation with influence functions, 2024. URL https://arxiv.org/abs/2405.13954.

[29] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. In *Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, 2018.

[30] cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. Toxic comment classification challenge. https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge, 2017. Kaggle.

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

[32] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019.

[33] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

[35] Creative Commons. CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, 2025. URL https://creativecommons.org/publicdomain/zero/1.0/.

[36] Creative Commons. Creative Commons Attribution 4.0 International License § 2(a)(1)(A), 2025. URL https://creativecommons.org/licenses/by/4.0/legalcode.

[37] Creative Commons. Creative Commons Attribution 4.0 International License § 3(a)(1)(A)(i), 2025. URL https://creativecommons.org/licenses/by/4.0/legalcode.

[38] Creative Commons. Public Domain Mark 1.0, 2025. URL https://creativecommons.org/publicdomain/mark/1.

[39] Creative Commons. Creative Commons Attribution-ShareAlike 4.0 International License, 2025. URL https://creativecommons.org/licenses/by-sa/4.0/.

[40] A. Feder Cooper and James Grimmelmann. The Files are in the Computer: Copyright, Memorization, and Generative AI. *arXiv preprint arXiv:2404.12590*, 2024.

[41] A. Feder Cooper, Aaron Gokaslan, Amy B. Cyphert, Christopher De Sa, Mark A. Lemley, Daniel E. Ho, and Percy Liang. Extracting memorized pieces of (copyrighted) books from open-weight language models. *arXiv preprint arXiv:2505.12546*, 2025.

[42] Yiming Cui, Ting Liu, Li Xiao, Zhipeng Chen, Wentao Ma, Wanxiang Che, Shijin Wang, and Guoping Hu. A span-extraction dataset for chinese machine reading comprehension. In *EMNLP-IJCNLP*, pages 5882–5888, 2019.

[43] Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen tau Yih, and Peter Clark. Tracking state changes in procedural text: a challenge dataset and models for process paragraph comprehension. In *North American Chapter of the Association for Computational Linguistics*, pages 1595–1604, 2018.

[44] Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Conference on Empirical Methods in Natural Language Processing*, volume abs/1908.05803, 2019.

[45] Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. Few-nerd: A few-shot named entity recognition dataset. *ArXiv*, abs/2105.07464, 2021.

[46] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *North American Chapter of the Association for Computational Linguistics*, pages 2368–2378, 2019.

[47] Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

[48] S. Dumitrescu, Petru Rebeja, Beáta Lőrincz, Mihaela Găman, M. Ilie, Andrei Pruteanu, Adriana Stan, Luciana Morogan, Traian Rebedea, and Sebastian Ruder. Liro: Benchmark and leaderboard for romanian language tasks. In *NeurIPS Datasets and Benchmarks*, 2021.

[49] Yanai Elazar and Yoav Goldberg. Where's my head? definition, data set, and models for numeric fused-head identification and resolution. *Transactions of the Association for Computational Linguistics*, 7:519–535, 2019.

[50] Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. Measuring causal effects of data statistics on language model'sfactual'predictions. *arXiv preprint arXiv:2207.14251*, 2022.

[51] Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. *ArXiv*, abs/2012.15738, 2020.

[52] Alex Fang, Hadi Pouransari, Matt Jordan, Alexander Toshev, Vaishaal Shankar, Ludwig Schmidt, and Tom Gunter. Datasets, documents, and repetitions: The practicalities of unequal data quality. *arXiv preprint arXiv:2503.07879*, 2025.

[53] James Ferguson, Matt Gardner, Tushar Khot, and Pradeep Dasigi. Iirc: A dataset of incomplete information reading comprehension questions. In *Conference on Empirical Methods in Natural Language Processing*, pages 1137–1147, 2020.

[54] Nancy Fulda, Nathan Tibbetts, Zachary Brown, and D. Wingate. Harvesting common-sense navigational knowledge for robotics from uncurated text corpora. In *Conference on Robot Learning*, pages 525–534, 2017.

[55] Philip Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994. URL https://api.semanticscholar.org/CorpusID:59804030.

[56] N. Gale, G. Heath, E. Cameron, S. Rashid, and S. Redwood. Using the framework method for the analysis of qualitative data in multi-disciplinary health research. *BMC Medical Research Methodology*, 13:117 – 117, 2013.

[57] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. URL https://arxiv.org/abs/2101.00027.

[58] Xinyang Geng and Hao Liu. Openllama: An open reproduction of llama, May 2023. URL https://github.com/openlm-research/open_llama.

[59] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, D. Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

[60] Max Glockner, Vered Shwartz, and Yoav Goldberg. Breaking nli systems with sentences that require simple lexical inferences. *ArXiv*, abs/1805.02266, 2018.

[61] Aaron Gokaslan, A. Feder Cooper, Jasmine Collins, Landan Seguin, Austin Jacobson, Mihir Patel, Jonathan Frankle, Cory Stephenson, and Volodymyr Kuleshov. CommonCanvas: Open Diffusion Models Trained on Creative-Commons Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8250–8260, June 2024.

[62] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar

Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala,

Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 Herd of Models, November 2024. URL http://arxiv.org/abs/2407.21783. arXiv:2407.21783 [cs].

[63] Grobid. Grobid. https://github.com/kermitt2/grobid, 2008–2025.

[64] 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. Olmo: Accelerating the science of language models, 2024. URL https://arxiv.org/abs/2402.00838.

[65] Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. Olmes: A standard for language model evaluations, 2025. URL https://arxiv.org/abs/2406.08446.

[66] Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. Wiki-40b: Multilingual language model dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2440–2452, 2020.

[67] Aditya Gupta, Jiacheng Xu, Shyam Upadhyay, Diyi Yang, and Manaal Faruqui. Disfl-qa: A benchmark dataset for understanding disfluencies in question answering. *ArXiv*, abs/2106.04016, 2021.

[68] Ivan Habernal, Omnia Zayed, and Iryna Gurevych. C4Corpus: Multilingual web-size corpus with free license. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 914–922, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1146/.

[69] Sungjun Han, Juyoung Suk, Suyeong An, Hyungguk Kim, Kyuseok Kim, Wonsuk Yang, Seungtaek Choi, and Jamin Shin. Trillion 7b technical report. *arXiv preprint arXiv:2504.15431*, 2025.

[70] Seth Hays. AI Training and Copyright Infringement: Solutions from Asia, October 2024. URL https://www.techpolicy.press/ai-training-and-copyright-infringement-solutions-from-asia/.

[71] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[72] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *ArXiv*, abs/2103.06268, 2021.

[73] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[74] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[75] HuggingFace: Common Corpus, 2025. URL https://huggingface.co/datasets/PleIAs/common_corpus.

[76] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*, pages 6384–6392, 2020.

[77] Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*, 2024.

[78] IFI CLAIMS Patent Services and Google. Google patents public data. https://patents.google.com/, 2023. Licensed under a Creative Commons Attribution 4.0 International License.

[79] Michael J Bommarito II, Jillian Bommarito, and Daniel Martin Katz. The kl3m data project: Copyright-clean training resources for large language models, 2025. URL https://arxiv.org/abs/2504.07854.

[80] Infocomm Media Development Authority of Singapore (IMDA), Aicadium, and AI Verify Foundation. Model AI Governance Framework for Generative AI: Fostering a Trusted Ecosystem, May 2024. URL https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf.

[81] Najko Jahn, Nick Haupka, and Anne Hobert. Analysing and reclassifying open access information in OpenAlex, 2023. URL https://subugoe.github.io/scholcomm_analytics/posts/oalex_oa_status/?utm_source=chatgpt.com.

[82] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics, April 2017.

[83] Nikhil Kandpal and Colin Raffel. Position: The most expensive part of an llm should be its training data. *arXiv preprint arXiv:2504.12427*, 2025.

[84] Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models, 2022. URL https://arxiv.org/abs/2202.06539.

[85] Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR, 2023.

[86] Pride Kavumba, Naoya Inoue, Benjamin Heinzerling, Keshav Singh, Paul Reisert, and Kentaro Inui. When choosing plausible alternatives, clever hans can be clever. *ArXiv*, abs/1911.00225, 2019.

[87] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. In *AAAI Conference on Artificial Intelligence*, pages 5189–5197, 2018.

[88] Rodney Michael 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 W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline

Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. The Semantic Scholar Open Data Platform. *ArXiv*, abs/2301.10140, 2023. URL https://api.semanticscholar.org/CorpusID:256194545.

[89] Andreas Kopf, Yannic Kilcher, Dimitri von Rutte, Sotiris Anagnostidis, Zhi Rui Tam, K. Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Rich'ard Nagyfi, ES Shahul, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and A. Mattick. Openassistant conversations - democratizing large language model alignment. *ArXiv*, abs/2304.07327, 2023.

[90] T. Kwiatkowski, J. Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, D. Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

[91] Faisal Ladhak, Esin Durmus, Claire Cardie, and K. McKeown. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. *ArXiv*, abs/2010.03093, 2020.

[92] Pierre-Carl Langlais. Releasing Common Corpus: the largest public domain dataset for training LLMs, 2024. URL https://huggingface.co/blog/Pclanglais/common-corpus.

[93] LDP Headquarters for the Promotion of Digital Society and Project Team on the Evolution and Implementation of AIs. AI White Paper 2024: New Strategies in Stage II, Toward the world's most AI-friendly country, April 2024. URL https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf.

[94] R. Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, 2016.

[95] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better, 2022. URL https://arxiv.org/abs/2107.06499.

[96] Katherine Lee, A. Feder Cooper, James Grimmelmann, and Daphne Ippolito. AI and Law: The Next Generation. *SSRN*, 2023. http://dx.doi.org/10.2139/ssrn.4580739.

[97] Katherine Lee, A. Feder Cooper, and James Grimmelmann. Talkin' 'Bout AI Generation: Copyright and the Generative-AI Supply Chain. *arXiv preprint arXiv:2309.08133*, 2023.

[98] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, D. Kontokostas, Pablo N. Mendes, Sebastian Hellmann, M. Morsey, Patrick van Kleef, S. Auer, and Christian Bizer. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6:167–195, 2015.

[99] H. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, 2011.

[100] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models, 2025. URL https://arxiv.org/abs/2406.11794.

[101] Xin Li and D. Roth. Learning question classifiers. In *International Conference on Computational Linguistics*, pages 1–7, 2002.

[102] Stephanie C. Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Annual Meeting of the Association for Computational Linguistics*, pages 3214–3252, 2021.

[103] Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language. *ArXiv*, abs/2204.12632, 2022.

[104] 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. Llm360: Towards fully transparent open-source llms, 2023. URL https://arxiv.org/abs/2312.06550.

[105] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447. URL https://www.aclweb.org/anthology/2020.acl-main.447.

[106] Shayne Longpre, Stella Biderman, Alon Albalak, Hailey Schoelkopf, Daniel McDuff, Sayash Kapoor, Kevin Klyman, Kyle Lo, Gabriel Ilharco, Nay San, et al. The responsible foundation model development cheatsheet: A review of tools & resources. *Transactions on Machine Learning Research*, 2024.

[107] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, Xinyi (Alexis) Wu, Enrico Shippole, Kurt Bollacker, Tongshuang Wu, Luis Villa, Sandy Pentland, and Sara Hooker. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence*, 6(8):975–987, August 2024. doi: 10/gt8f5p.

[108] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, and Sandy Pentland. Consent in crisis: The rapid decline of the AI data commons. *Advances in Neural Information Processing Systems*, 37, 2024.

[109] Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Katy Gero, Sandy Pentland, and Jad Kabbara. Data authenticity, consent, & provenance for ai are all broken: what will it take to fix them? *arXiv preprint arXiv:2404.12691*, 2024.

[110] Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, et al. Bridging the data provenance gap across text, speech and video. *arXiv preprint arXiv:2412.17847*, 2024.

[111] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, 2024.

[112] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

[113] Annie Louis, D. Roth, and Filip Radlinski. "i'd rather just go to bed"': Understanding indirect answers. In *Conference on Empirical Methods in Natural Language Processing*, volume abs/2010.03450, 2020.

[114] Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, Zhuang Li, Wen-Ding Li, Megan Risdal, Jia Li, Jian Zhu, Terry Yue Zhuo, Evgenii Zheltonozhskii, Nii Osae Osae Dade, Wenhao Yu, Lucas Krauß, Naman Jain, Yixuan Su, Xuanli He, Manan Dey, Edoardo Abati, Yekun Chai, Niklas Muennighoff, Xiangru Tang, Muhtasham Oblokulov, Christopher Akiki, Marc Marone, Chenghao Mou, Mayank Mishra, Alex Gu, Binyuan Hui, Tri Dao, Armel Zebaze, Olivier Dehaene, Nicolas Patry, Canwen Xu, Julian McAuley, Han Hu, Torsten Scholak, Sebastien Paquet, Jennifer Robinson, Carolyn Jane Anderson, Nicolas Chapados, Mostofa Patwary, Nima Tajbakhsh, Yacine Jernite, Carlos Muñoz Ferrandis, Lingming Zhang, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder 2 and the stack v2: The next generation, 2024.

[115] Robert Mahari and Shayne Longpre. Discit ergo est: Training data provenance and fair use. *Robert Mahari and Shayne Longpre, Discit ergo est: Training Data Provenance And Fair Use, Dynamics of Generative AI (ed. Thibault Schrepel & Volker Stocker), Network Law Review, Winter*, 2023.

[116] Matt Mahoney. Large text compression benchmark, 2011.

[117] Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the sdes and scaling rules for adaptive gradient algorithms. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 7697–7711. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/32ac710102f0620d0f28d5d05a44fe08-Paper-Conference.pdf.

[118] Stephen Merity, Caiming Xiong, James Bradbury, and R. Socher. Pointer sentinel mixture models. *ArXiv*, abs/1609.07843, 2016.

[119] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182, 2011. doi: 10.1126/science.1199644. URL https://www.science.org/doi/abs/10.1126/science.1199644.

[120] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

[121] Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah A. Smith, and Luke Zettlemoyer. SILO language models: Isolating legal risk in a nonparametric datastore. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ruk0nyQPec.

[122] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2023.

[123] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 50358–50376. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9d89448b63ce1e2e8dc7af72c984c196-Paper-Conference.pdf.

[124] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Conference on Empirical Methods in Natural Language Processing*, pages 1953–1967, 2020.

[125] Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. The e2e dataset: New challenges for end-to-end generation. *ArXiv*, abs/1706.09254, 2017.

[126] Tomoko Ohta, Sampo Pyysalo, Junichi Tsujii, and S. Ananiadou. Open-domain anatomical entity mention detection. In *Annual Meeting of the Association for Computational Linguistics*, pages 27–36, 2012.

[127] Yasumasa Onoe, Michael J.Q. Zhang, Eunsol Choi, and Greg Durrett. Creak: A dataset for commonsense reasoning over entity knowledge. *ArXiv*, abs/2109.01653, 2021.

[128] OpenAlex, 2025. URL https://openalex.org.

[129] Vassil Panayotov, Guoguo Chen, Daniel Povey, and S. Khudanpur. Librispeech: An asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[130] Ashwinee Panda, Xinyu Tang, Milad Nasr, Christopher A Choquette-Choo, and Prateek Mittal. Privacy auditing of large language models. *arXiv preprint arXiv:2503.06808*, 2025.

[131] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. Trak: Attributing model behavior at scale, 2023. URL https://arxiv.org/abs/2303.14186.

[132] European Parliament and Council of the European Union. Directive (eu) 2019/790, 2019. URL https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32019L0790#art_3.

[133] ParlParse. Parser for uk parliament proceedings. https://parser.theyworkforyou.com/, 2025. Accessed: 2025-05-09.

[134] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37, 2024.

[135] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[136] E. Ponti, Goran Glavavs, Olga Majewska, Qianchu Liu, Ivan Vulic, and A. Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. In *Conference on Empirical Methods in Natural Language Processing*, pages 2362–2376, 2020.

[137] Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. Dynasent: A dynamic benchmark for sentiment analysis. *ArXiv*, abs/2012.15349, 2020.

[138] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

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

[140] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

[141] Filip Radlinski, K. Balog, B. Byrne, and K. Krishnamoorthi. Coached conversational preference elicitation: A case study in understanding movie preferences. In *SIGDIAL Conferences*, pages 353–360, 2019.

[142] Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. Compressive transformers for long-range sequence modelling. *arXiv preprint*, 2019. URL https://arxiv.org/abs/1911.05507.

[143] 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, Nat 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. Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446, 2021. URL https://api.semanticscholar.org/CorpusID:245353475.

[144] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140), 2020.

[145] Robi Rahman and David Owen. The size of datasets used to train language models doubles approximately every seven months, 2024. URL https://epoch.ai/data-insights/dataset-size-trend. Accessed: 2025-05-08.

[146] Nazneen Rajani, Bryan McCann, Caiming Xiong, and R. Socher. Explain yourself! leveraging language models for commonsense reasoning. *ArXiv*, abs/1906.02361, 2019.

[147] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 557–571, 2022.

[148] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, 2016.

[149] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Annual Meeting of the Association for Computational Linguistics*, volume abs/1806.03822, 2018.

[150] Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Schema-guided dialogue state tracking task at dstc8. *ArXiv*, abs/2002.01359, 2020.

[151] Abhilasha Ravichander, Matt Gardner, and Ana Marasović. Condaqa: A contrastive reading comprehension dataset for reasoning about negation. *ArXiv*, abs/2211.00295, 2022.

[152] Varshini Reddy, Craig W. Schmidt, Yuval Pinter, and Chris Tanner. How much is enough? the diminishing returns of tokenization training data, 2025. URL https://arxiv.org/abs/2502.20273.

[153] Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, V. Thang, N. Thai, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, K. Soe, K. Nwet, M. Utiyama, and Chenchen Ding. Introduction of the asian language treebank. In *Oriental COCOSDA International Conference on Speech Database and Assessments*, pages 1–6, 2016.

[154] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[155] Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting closer to ai complete question answering: A set of prerequisite real tasks. In *AAAI Conference on Artificial Intelligence*, pages 8722–8731, 2020.

[156] Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how BERT works. *Transactions of the association for computational linguistics*, 8, 2021.

[157] Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Thinking like a skeptic: Defeasible inference in natural language. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.418. URL https://aclanthology.org/2020.findings-emnlp.418/.

[158] Matthew Sag and Peter K. Yu. The globalization of copyright exceptions for ai training. *Emory Law Journal*, 74, 2025. doi: http://dx.doi.org/10.2139/ssrn.4976393. URL https://ssrn.com/abstract=4976393.

[159] Swarnadeep Saha, Yixin Nie, and Mohit Bansal. Conjnli: Natural language inference over conjunctive sentences. In *Conference on Empirical Methods in Natural Language Processing*, pages 8240–8252, 2020.

[160] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande. *Communications of the ACM*, 64:99 – 106, 2019.

[161] Pamela Samuelson. How to Think About Remedies in the Generative AI Copyright Cases. *Lawfare*, February 2024. URL https://www.lawfaremedia.org/article/how-to-think-about-remedies-in-the-generative-ai-copyright-cases.

[162] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019. URL https://arxiv.org/abs/1904.09728.

[163] A. Sboev, A. Naumov, and R. Rybka. Data-driven model for emotion detection in russian texts. In *BICAAI*, pages 637–642, 2020.

[164] Tal Schuster, Adam Fisch, and R. Barzilay. Get your vitamin c! robust fact verification with contrastive evidence. In *North American Chapter of the Association for Computational Linguistics*, pages 624–643, 2021.

[165] Emily Sheng and David C. Uthus. Investigating societal biases in a poetry composition system. *ArXiv*, abs/2011.02686, 2020.

[166] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew J. Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. *ArXiv*, abs/2010.03768, 2020.

[167] Shivalika Singh, Freddie Vargus, Daniel Dsouza, B"orje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemi'nski, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Minh Chien Vu, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, A. Ustun, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning. *ArXiv*, abs/2402.06619, 2024. URL https://api.semanticscholar.org/CorpusID:267617144.

[168] Luca Soldaini and Kyle Lo. peS2o (Pretraining Efficiently on S2ORC) Dataset. Technical report, Allen Institute for AI, 2023. ODC-By, https://github.com/allenai/pes2o.

[169] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob

Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[170] Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. Evaluating gender bias in machine translation. *ArXiv*, abs/1906.00591, 2019.

[171] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*. Leibniz-Institut für Deutsche Sprache, 2019.

[172] Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. Quartz: An open-domain dataset of qualitative relationship questions. In *Conference on Empirical Methods in Natural Language Processing*, volume abs/1909.03553, 2019.

[173] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[174] Niket Tandon, Keisuke Sakaguchi, Bhavana Dalvi, Dheeraj Rajagopal, Peter Clark, Michal Guerquin, Kyle Richardson, and E. Hovy. A dataset for tracking entities in open domain procedural text. *ArXiv*, abs/2011.08092, 2020.

[175] Liping Tang, Nikhil Ranjan, Omkar Pangarkar, Xuezhi Liang, Zhen Wang, Li An, Bhaskar Rao, Linghao Jin, Huijuan Wang, Zhoujun Cheng, Suqi Sun, Cun Mu, Victor Miller, Xuezhe Ma, Yue Peng, Zhengzhong Liu, and Eric P. Xing. Txt360: A top-quality llm pre-training dataset requires the perfect blend, 2024.

[176] Ishan Tarunesh, Somak Aditya, and M. Choudhury. Trusting roberta over bert: Insights from checklisting the natural language inference task. *ArXiv*, abs/2107.07229, 2021.

[177] MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL www.mosaicml.com/blog/mpt-7b. Accessed: 2023-05-05.

[178] Qwen Team. Qwen3, April 2025. URL https://qwenlm.github.io/blog/qwen3/.

[179] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. *ArXiv*, abs/1803.05355, 2018.

[180] Anvith Thudi, Evianne Rovers, Yangjun Ruan, Tristan Thrush, and Chris J. Maddison. Mixmin: Finding data mixtures via convex minimization, 2025. URL https://arxiv.org/abs/2502.10510.

[181] 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. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

[182] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[183] UK Parliament. Open parliament license. https://www.parliament.uk/site-information/copyright-parliament/open-parliament-licence/, Unknown. Accessed: 2025-05-09.

[184] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[185] Mathurin Videau, Badr Youbi Idrissi, Daniel Haziza, Luca Wehrstedt, Jade Copet, Olivier Teytaud, and David Lopez-Paz. Meta Lingua: A minimal PyTorch LLM training library, 2024. URL https://github.com/facebookresearch/lingua.

[186] Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. Helpsteer: Multi-attribute helpfulness dataset for steerlm. *ArXiv*, abs/2311.09528, 2023.

[187] Maurice Weber, Daniel Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. Redpajama: an open dataset for training large language models, 2024. URL https://arxiv.org/abs/2411.12372.

[188] Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. Resolving gendered ambiguous pronouns with bert. *ArXiv*, abs/1906.01161, 2019.

[189] Wei Wei, Quoc V. Le, Andrew M. Dai, and Jia Li. Airdialogue: An environment for goal-oriented dialogue research. In *Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854, 2018.

[190] Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*, 2025.

[191] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36, 2023.

[192] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, R. Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, 2018.

[193] Cat Zakrzewski, Nitasha Tiku, and Elizabeth Dwoskin. OpenAI prepares to fight for its life as legal troubles mount. *The Washington Post*, 2024.

[194] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[195] 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. MAP-Neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv:2405.19327*, 2024.

[196] Hongming Zhang, Xinran Zhao, and Yangqiu Song. Winowhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge. In *Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, 2020.

[197] Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

[198] Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and D. Roth. Temporal reasoning on implicit events from distant supervision. *ArXiv*, abs/2010.12753, 2020.

Appendix

Table of Contents

A Contributions

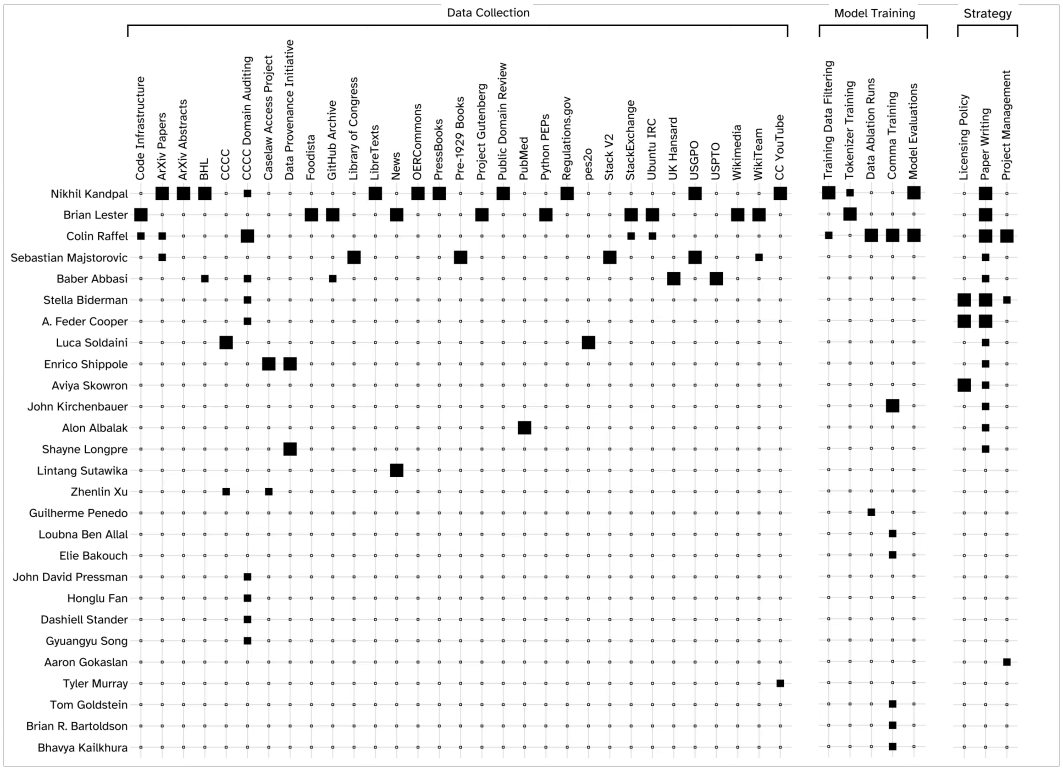

Figure 4: **Author contributions to the Common Pile and Comma.** Large squares indicate a major contribution and small squares indicate a supporting contribution.

B Detailed Description of Sources

Below, we give a more in-depth overview of the sources that make up the Common Pile, including specific license decisions and tools used during collection.

B.1 Scientific and Scholarly Text

Scientific and scholarly texts are a staple of modern LLM pretraining corpora, appearing in nearly all large-scale datasets [e.g. 57, 187, 169] since they expose models to technical terminology, formal reasoning, and long-range document structure—skills that are essential for downstream tasks in science, education, and question answering. Thanks to open access mandates and academic cultural norms, many scholarly texts are either in the public domain or are distributed under open licenses.

peS2o To ensure broad coverage across many scientific disciplines, we include a version of peS2o [168] restricted to openly licensed articles. pes2o is derived from S2ORC [105], a corpus of openly licensed abstract and full-text papers that have been converted to a structured format using Grobid [63]. Starting from Grobid's XML output, peS2o filters papers that are too short, have incorrect metadata, are in languages other than English, and contain OCR errors using a combination of heuristic- and model-based filtering steps. We refer the reader to the datasheet and code for more details on this processing pipeline. The subset of peS2o included in the Common Pile starts from v3 of the corpus, which contains documents from January 1, 1970 to October 6, 2024. We retain full-text papers with CC BY, CC BY-SA, or CC0 licenses, or that have been labeled as public domain; metadata is provided by the Semantic Scholar APIs [88]. After filtering, this set contains 6.3 million

papers, or 35.7 billion whitespace-separated segments. We provide more details on the composition of this subset in Appendix H.

PubMed PubMed Central (PMC) is an open-access archive of biomedical and life sciences research papers maintained by the U.S. National Institutes of Health's National Library of Medicine. We collected papers from PMC whose metadata indicated that the publishing journal had designated a CC BY, CC BY-SA, or CC0 license. PMC stores the text content of each article as a single XML file, which we convert to markdown using pandoc.

ArXiv Papers ArXiv is an online open-access repository of over 2.4 million scholarly papers covering fields such as computer science, mathematics, physics, quantitative biology, economics, and more. When uploading papers, authors can choose from a variety of licenses. We included text from all papers uploaded under CC BY, CC BY-SA, and CC0 licenses in the Common Pile through a three-step pipeline: first, the latex source files for openly licensed papers were downloaded from ArXiv's bulk-access S3 bucket; next, the LaTeXML conversion tool was used to convert these source files into a single HTML document; finally, the HTML was converted to plaintext using the Trafilatura [10] HTML-processing library.

ArXiv Abstracts Each paper uploaded to ArXiv includes structured metadata fields, including an abstract summarizing the paper's findings and contributions. According to ArXiv's licensing policy, the metadata for any paper submitted to ArXiv is distributed under the CC0 license, regardless of the license of the paper itself. Thus, we include as an additional source the abstracts for every paper submitted to ArXiv. We source the abstracts from ArXiv's API via the Open Archives Initiative Protocol for Metadata Harvesting endpoint and reproduce them as-is.

B.2 Online Discussions and Forums

Online forums are a rich source of multi-turn, user-generated dialogue covering a wide range of topics. These platforms often feature question–answer pairs, problem-solving discussions, and informal explanations of technical and non-technical concepts. The Common Pile incorporates online discussions from sources that distribute content under an open license.

StackExchange While StackExchange formerly provided structured XML dumps of all of their content, since July of 2024, StackExchange has stopped publishing dumps to the Internet Archive. Instead, each site can provide a logged-in user with a custom URL to download the dump for that site. This means that dumps for defunct sites like windowsphone.stackexchange.com are inaccessible. Additionally, in dumps produced by the new export tool, many questions that are available in past dumps (and accessible on the site) are not present. We therefore extract all questions and answers from community uploaded dumps from December of 2024 from the Internet Archive and additionally extract missing questions and answers from the last official dumps in July of 2024 to account for the deficiencies listed above. We use a question, its comments, its answers and the comments on each answer as a single document. Following the display order on StackExchange, answers are ordered by the number of votes they received, with the exception that the "accepted answer" always appears first. PyMarkdown was used to convert each comment into plain text.

GitHub Archive According to GitHub's terms of service, issues and pull request descriptions—along with their comments—inherit the license of their associated repository. To collect this data, we used the GitHub Archive's public BigQuery table of events to extracted all issue, pull request, and comment events since 2011 and aggregated them into threads. The table does not include "edit" events so the text from each comment is the original from when it was first posted. We filtered out comments from bots. This resulted in approximately 177 million threads across 19 million repositories. We then removed threads whose repositories did not have a Blue Oak Council-approved license. License information for each repository comes from either 1) the "public-data:github_repos" BigQuery Table, 2) metadata from the StackV2, or 3) the GitHub API. License filtering left 10 million repositories. PyMarkdown was used to convert from GitHub-flavored markdown to plain text. When parsing failed, the raw markdown was kept.

Ubuntu IRC Logs of all discussions on the Ubuntu-hosted Internet Relay Chat (IRC) since 2004 have been archived and released into the Public Domain. We downloaded all chats from all channels up until March of 2025. We consider all messages for given channel on a given day as a single document. We removed system messages as well as those from known bots.

B.3 Government and Legal Texts

Governments produce a vast amount of informational text, ranging from legislation and legal opinions to scientific reports, public communications, and regulatory notices. This content is explicitly intended to inform the public, and as such, in many jurisdictions it is published directly into the public domain or under open licenses. In the United States, for example, works authored by federal employees as part of their official duties are not subject to copyright. Government and legal texts offer language models exposure to formal argumentation, legal reasoning, and procedural language.

US Government Publishing Office The United States Government Publishing Office (USGPO) is a federal agency responsible for disseminating official documents authored by the U.S. government. The Common Pile v0.1 includes all plain-text documents made available through the USGPO's GovInfo.gov developer API. This collection comprises over 2.7 million documents, spanning issues of the Federal Register, congressional hearing transcripts, budget reports, economic indicators, and other federal publications.

US Patents and Trademark Office In the US, patent documents are released into the public domain as government works. Patents follow a highly standardized format with distinct required sections for background, detailed description, and claims. We include parents from the US Patents and Trademark Office (USPTO) as provided by the Google Patents Public Data dataset [78], which includes millions of granted patents and published patent applications dating back to 1782. We processed these documents to extract clean text while preserving this structured format. Mathematical expressions and equations were converted into LaTeX.

Caselaw Access Project and Court Listener The Common Pile contains 6.7 million cases from the Caselaw Access Project and Court Listener. The Caselaw Access Project consists of nearly 40 million pages of U.S. federal and state court decisions and judges' opinions from the last 365 years. In addition, Court Listener adds over 900 thousand cases scraped from 479 courts. The Caselaw Access Project and Court Listener source legal data from a wide variety of resources such as the Harvard Law Library, the Law Library of Congress, and the Supreme Court Database. From these sources, we only included documents that were in the public domain. Erroneous OCR errors were further corrected after digitization, and additional post-processing was done to fix formatting and parsing.

UK Hansard Hansard represents the official record of parliamentary proceedings across the United Kingdom's legislative bodies. The Common Pile incorporates records from multiple sources, including debates and written answers from the UK Commons and Lords, devolved legislatures (Scottish Parliament, Senedd in both English and Welsh, Northern Ireland Assembly), London Mayor's Questions, and ministerial statements. Data was sourced from ParlParse [133], covering Commons debates from 1918 forward and Lords proceedings from the 1999 reform. Each document was processed to preserve complete parliamentary sessions as cohesive units, maintaining the natural flow of debate. All content is published under the Open Parliament License [183].

Regulations.gov Regulations.gov is an online platform operated by the U.S. General Services Administration that collates newly proposed rules and regulations from federal agencies along with comments and feedback from the general public. The Common Pile includes all plain-text regulatory documents published by U.S. federal agencies on this platform, acquired via the bulk download interface provided by Regulations.gov.

B.4 Curated Task Data

Curated datasets that cover specific tasks such as question answering, summarization, or text classification are often released via open licenses to the research community. While not traditionally part of pretraining corpora, including a small amount of task-oriented data during pretraining can help models acquire early familiarity with task formats and prompt–completion structures.

Data Provenance Initiative The Data Provenance Initiative is a digital library of supervised datasets that have been manually annotated with their source and license information [107, 110]. We leverage their tooling to filter HuggingFace datasets, based on a range of criteria, including their licenses, which may be particularly relevant for supervised datasets [115]. Specifically, we filter the data according to these criteria: contains English language or code data, the text is not model-generated, the dataset's audit yielded a open license and the original sources of the data are only from recognized public domain sources.

B.5 Books in the Public Domain

Books represent a time-tested resource for language model pretraining, offering carefully edited, long-form prose that supports learning of narrative coherence and long-range dependency modeling. For these reasons, many large-scale pretraining corpora—including the Pile [57], Dolma [169], and RedPajama [187]—include content from books [41]. In the United States, as of 2024, books published prior to 1929 are in the public domain. Thus, the Common Pile includes public domain books drawn from curated collections, covering topics such as literature, science, and history.

Biodiversity Heritage Library The Biodiversity Heritage Library (BHL) is an open-access digital library for biodiversity literature and archives. The Common Pile contains over 42 million public domain books and documents from the BHL collection. These works were collected using the bulk data download interface provided by the BHL and were filtered based on their associated license metadata. We use the optical character recognition (OCR)-generated text distributed by BHL.

Pre-1929 Books Books published in the US before 1929 passed into the public domain on January 1, 2024. We used the bibliographic catalog Hathifiles produced by HathiTrust to identify digitized books which were published in the US before 1929. The collection contains over 130,000 books digitized and processed by the Internet Archive on behalf of HathiTrust member libraries. The OCR plain text files were downloaded directly from the Internet Archive website.

Library of Congress The Library of Congress (LoC) curates a collection of public domain books called "Selected Digitized Books". We downloaded over 130,000 English-language books from this public domain collection as OCR plain text files using the LoC APIs.

Project Gutenberg Project Gutenberg is an online collection of over 75,000 digitized books available as plain text. We use all books that are 1) English and 2) marked as in the Public Domain according to the provided metadata. Additionally, we include any books that are part of the pg19 [142] dataset, which only includes books that are over 100 years old. Minimal preprocessing is applied to remove the Project Gutenberg header and footers, and many scanned books include preamble information about who digitized them.

B.6 Open Educational Resources

Open Educational Resources (OERs) are educational materials, typically published under open licenses, to support free and equitable access to education. These resources include educational artifacts such as textbooks, lecture notes, lesson plans, syllabi, and problem sets. For language models, OERs offer exposure to instructional formatting and domain-specific information, making them valuable for improving performance on knowledge-based downstream tasks. The Common Pile includes a range of such materials sourced from major OER repositories, including collections of open-access books and structured teaching resources.

Directory of Open Access Books The Directory of Open Access Books (DOAB) is an online index of over 94,000 peer-reviewed books curated from trusted open-access publishers. To collect the openly licensed content from DOAB, we retrieve metadata using their official metadata feed. We then filter the collection to include only English-language books released under CC BY and CC BY-SA licenses. The filtered books are downloaded in PDF format and converted to plaintext using the Marker PDF-to-text converter. As an additional validation step, we manually create a whitelist of open license statements and retain only texts explicitly containing one of these statements in their front- or back-matter.

PressBooks PressBooks is a searchable catalog of over 8,000 open access books. To collect openly licensed content from PressBooks we construct a search query to retrieve URLs for all books written in English and listed as public domain or under CC BY or CC BY-SA licenses. For each matched book, we collect its contents directly from the publicly available web version provided by PressBooks.

OERCommons OERCommons is an online platform where educators share open-access instructional materials—such as textbooks, lesson plans, problem sets, course syllabi, and worksheets—with the goal of expanding access to affordable education. To collect the openly licensed content available on OERCommons, we construct a search query to retrieve English-language content released into the public domain or under CC BY or CC BY-SA licenses. The resulting documents are converted to plain text directly from the HTML pages hosted on the OERCommons website.

LibreTexts LibreTexts is an online platform that provides a catalog of over 3,000 open-access textbooks. To collect openly licensed content from LibreTexts we gather links to all textbooks in the catalog and check each textbook section for a license statement indicating that it is in the public domain or under a CC BY, CC BY-SA, or the GNU Free Documentation License. We extract plain text from these textbook sections directly from the HTML pages hosted on the LibreTexts website.

B.7 Wikis

Wikis are collaboratively maintained websites that organize information around specific topics or domains. Their crowd-sourced nature, coupled with community moderation and citation requirements, often results in text that is both informative and well-structured. Prominent examples such as Wikipedia have become staples in large-scale language model pretraining corpora due to their breadth of coverage and high quality. In addition, most major wikis are distributed under open licenses such as CC BY and CC BY-SA. The Common Pile includes content from a range of openly licensed wikis to provide models with structured and well-researched informational text.

Wikimedia We downloaded the official database dumps from March 2025 of the English-language wikis that are directly managed by the Wikimedia foundation (see Appendix F for a complete list). These database dumps include the wikitext—Mediawiki's custom markup language—for each page as well as talk pages, where editors discuss changes made to a page. We only use the most recent version of each page. We converted wikitext to plain text using `wtf_wikipedia` after light adjustments in formatting to avoid errors in section ordering caused by a bug. Before parsing, we converted wikitext math into LaTeX math using our custom code. Finally, any remaining HTML tags were removed via regexes.

Wikiteam There are many wikis on the internet that are not managed by the Wikimedia foundation, but do use their MediaWiki software to power their wiki. Many of these wikis have been archived by wikiteam, a collection of volunteers that create unofficial database dumps of wikis and upload them to the Internet Archive. We download all dumps made by wikiteam when the metadata indicates the wiki was licensed under CC BY, CC BY-SA, or released into the public domain on the Internet Archive in September of 2024. This results in downloading approximately 330,000 wikis. When multiple dumps of the same wiki exists, we use the most recent dump. The wikitext was converted to plain text following the same steps as with Wikimedia wikis. After preprocessing, we removed documents from wikis that appeared to contain large amounts of license laundering, e.g. those that were collections of song lyrics or transcripts.

B.8 Source Code

Source code has become an increasingly important component of large-scale language model pretraining corpora, as it enables models to learn syntax, program structure, and problem solving strategies useful for both code generation and reasoning tasks. Thanks to the Free and Open Source Software (FOSS) movement, code also happens to be one of the most openly licensed forms of text, with many software repositories distributed under open licenses such as MIT, BSD, Apache 2.0, and the GNU Free Documentation License (GFDL). The Common Pile includes high-quality, openly licensed source code from large-scale public code datasets and documentation standards, enabling models trained on it to perform better on coding and technical writing tasks.

The Stack V2 The Stack V2 [114] consists of a mixture of openly licensed and unlicensed work. We use the tooling that the Software Heritage Foundation and BigCode created to build our dataset. In particular, we relied on the license detection performed by the creators of Stack V2. When multiple licenses are detected in a single repository, we make sure that *all* of them meet our definition of "openly licensed".

Python Enhancement Proposals Python Enhancement Proposals, or PEPs, are design documents that generally provide a technical specification and rationale for new features of the Python programming language. There are been 661 PEPs published. The majority of PEPs are published in the Public Domain, but 5 were published under the "Open Publication License" and omitted. PEPs are long, highly-polished, and technical in nature and often include code examples paired with their prose. PEPs are authored in ReStructured Text; we used pandoc, version 3.5, to convert them to plain text.

B.9 Transcribed Audio Content

A historically underutilized source of text data is speech transcribed from audio and video content. Spoken language in educational videos, speeches, and interviews provide an opportunity for models to learn conversational speech patterns.

Creative Commons YouTube YouTube is large-scale video-sharing platform where users have the option of uploading content under a CC BY license. To collect high-quality speech-based textual content and combat the rampant license laundering on YouTube, we manually curated a set of over 2,000 YouTube channels that consistently release original openly licensed content containing speech. The resulting collection spans a wide range of genres, including lectures, tutorials, reviews, video essays, speeches, and vlogs. From these channels, we retrieved over 1.1 million openly licensed videos comprising more than 470,000 hours of content. Finally, each video was transcribed to text using the Whisper speech recognition model [140].

B.10 Web Text

The success of modern LLM pre-training relies on text scraped indiscriminately from the web, as web text covers an extremely diverse range of textual domains. In the Common Pile, we restrict this approach to only include web content with clear public domain status or open license statements.

Creative Commons Common Crawl We sourced text from 52 Common Crawl snapshots, covering about half of Common Crawl snapshots available to date and covering all years of operations of Common Crawl up to 2024. We found a higher level of duplication across this collection, suggesting that including more snapshots would lead to a modest increase in total token yield. From these snapshots, we extract HTML content using FastWarc [15]. Then, using a regular expression adapted from the C4Corpus project [68], we retain only those pages where a CC BY, CC BY-SA, or CC0 license appears. To ensure license accuracy, we manually verified the top 1000 domains by content volume, retaining only the 537 domains with confirmed licenses where the Creative Commons designation is applied to all text content rather than only embedded media or a subset of the text on the domain. We extract the main content of these documents and remove boilerplate using Resiliparse [14]. We perform URL-level exact deduplication and use Bloom filters to remove near-duplicates with 80% ngram overlap. We also employ rule-based filters matching Dolma [169]; namely, we use C4-derived heuristics [144] to filter pages containing Javascript, Lorem Ipsum, and curly braces `{}`. We also apply all Gopher rules [143] to remove low-quality pages. We provide more details on the composition of this subset in Appendix G.

Foodista Foodista is a community-maintained site with recipes, food-related news, and nutrition information. All content is licensed under CC BY. Plain text is extracted from the HTML using a custom pipeline that includes extracting title and author information to include at the beginning of the text. Additionally, comments on the page are appended to the article after we filter automatically generated comments.

News We scrape the news sites that publish content under CC BY or CC BY-SA according to opennewswire. A full list of sites can be found in Appendix E. Plain text was extracted from the HTML using our custom pipeline, including extraction of the title and byline to include at the beginning of each article.

Public Domain Review The Public Domain Review is an online journal dedicated to exploration of works of art and literature that have aged into the public domain. We collect all articles published in the Public Domain Review under a CC BY-SA license.

C Additional insights on licensing

There are many standards we could have chosen for what licenses to include in our dataset. The open source, knowledge, and culture movements have harmonized on the high level principles described in Section 1: "open" means that permission is granted for content to be freely used, studied, modified, and shared for any purpose. This language is found in the Open Knowledge Definition we follow as well as the Open Source Institute's Open Definition, Creative Commons's statement on Open Culture, Wikimedia's Acceptable licenses policy and more. Our work was also developed to be consistent with the Open movement's work in the specific context of AI technologies such as the

Open Source Initiative's Open Source AI Definition and in consultations with leading members of the community [9].

C.1 Why we can't always trust automatic license detection

There are many reasons why identifying the licensing status of internet text with automatic tooling can be challenging. In this section, we briefly discuss some major themes from our experience.

There are many ways to say the same thing. While there exist standards for how to express a license, people don't always follow those standards and failure to follow the standards doesn't mean that the license is invalid. For example, simple string matching on "CC BY" misses a huge amount of CC BY licensed text because a very common way to denote Creative Commons licenses is using an image badge. Current web-processing tools are substantially stronger at identifying text than images, and the failure rate on sites using image badges is quite high.

Lack of understanding of licenses. Most people are not lawyers and do not understand the full legal scope and meaning of the licenses that they attempt to put on their text. Developers routinely tweak boilerplate to produce ambiguous language like ("Licensed under MIT-ish terms") or write contradictory statements ("All rights reserved / CC-BY"). In general, it is common for people to write quasi-legal language along side a more traditional license. Non-standard licenses require substantial amounts of work to interpret and are not always valid or meaningful.

Licensing signals can be noisy. Even when a developer intends to clearly communicate a specific license, contradictions and errors can occur in practice. For example, Longpre et al. [108] found that there were substantial disagreements between the terms of service of a website and the restrictions found in a robots.txt file. We have not yet found a reliable way to have an automatic system identify licensed text and therefore frequently resort to manual review by humans.

D List of Data Provenance Initiative sources

The openly licensed supervised datasets included in the Common Pile are listed in Table 1. These datasets were identified and collected using metadata from the Data Provenance Initiative. For more information on these datasets, consult the Data Provenance Initiative Dataset Explorer.

Table 1: Supervised datasets included in the Common Pile from the Data Provenance Initiative collection.

Collection	Dataset Identifier	Licenses
AgentInstruct	AgentInstruct-alfworld[166]	MIT License
HelpSteer	HelpSteer[186]	CC BY 4.0
Aya Dataset	aya-english[167]	Apache License 2.0
CommitPackFT	commitpackft-abap[167]	MIT License
CommitPackFT	commitpackft-agda[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-apl[167]	MIT License, ISC License
CommitPackFT	commitpackft-arc[167]	MIT License
CommitPackFT	commitpackft-aspectj[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-ats[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-blitzmax[167]	MIT License
CommitPackFT	commitpackft-bluespec[167]	MIT License
CommitPackFT	commitpackft-boo[167]	MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
CommitPackFT	commitpackft-brainfuck[167]	Apache License 2.0, BSD 2-Clause License, MIT License
CommitPackFT	commitpackft-bro[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-cartocss[167]	MIT License
CommitPackFT	commitpackft-chapel[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-clean[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-coldfusion[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-creole[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-crystal[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-dns-zone[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-dylan[167]	MIT License
CommitPackFT	commitpackft-eiffel[167]	MIT License
CommitPackFT	commitpackft-emberscript[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-fancy[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-flux[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-forth[167]	MIT License
CommitPackFT	commitpackft-g-code[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-gdscript[167]	Apache License 2.0, CC0 1.0, MIT License
CommitPackFT	commitpackft-genshi[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-graphql[167]	Apache License 2.0, BSD 3-Clause License, CC0 1.0, MIT License
CommitPackFT	commitpackft-harbour[167]	MIT License
CommitPackFT	commitpackft-hlsl[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-http[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-idris[167]	MIT License, BSD 3-Clause License, BSD 2-Clause License
CommitPackFT	commitpackft-igor-pro[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-inform-7[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-ioke[167]	MIT License
CommitPackFT	commitpackft-isabelle[167]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-jflex[167]	MIT License
CommitPackFT	commitpackft-json5[167]	MIT License, BSD 3-Clause License, BSD 2-Clause License
CommitPackFT	commitpackft-jsonld[167]	Apache License 2.0, BSD 3-Clause License, CC0 1.0, MIT License
CommitPackFT	commitpackft-krl[167]	MIT License
CommitPackFT	commitpackft-latte[167]	MIT License
CommitPackFT	commitpackft-lean[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-lfe[167]	Apache License 2.0, MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
CommitPackFT	commitpackft-lilypond[167]	MIT License
CommitPackFT	commitpackft-liquid[167]	Apache License 2.0, CC0 1.0, MIT License
CommitPackFT	commitpackft-literate-agda[167]	MIT License
CommitPackFT	commitpackft-literate-coffeescript[167]	MIT License
CommitPackFT	commitpackft-literate-haskell[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-llvm[167]	Apache License 2.0, BSD 3-Clause License, BSD 2-Clause License, MIT License
CommitPackFT	commitpackft-logos[167]	Apache License 2.0, BSD 3-Clause License, BSD 2-Clause License, MIT License, ISC License
CommitPackFT	commitpackft-lsl[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-maple[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-mathematica[167]	MIT License, CC0 1.0
CommitPackFT	commitpackft-metal[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-mirah[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-monkey[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-moonscript[167]	MIT License
CommitPackFT	commitpackft-mtml[167]	MIT License
CommitPackFT	commitpackft-mupad[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-nesc[167]	MIT License
CommitPackFT	commitpackft-netlinx[167]	MIT License
CommitPackFT	commitpackft-ninja[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-nit[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-nu[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-ooc[167]	MIT License
CommitPackFT	commitpackft-openscad[167]	MIT License, CC0 1.0, BSD 2-Clause License
CommitPackFT	commitpackft-oz[167]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-pan[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-piglatin[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-pony[167]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-propeller-spin[167]	MIT License
CommitPackFT	commitpackft-pure-data[167]	MIT License
CommitPackFT	commitpackft-purebasic[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-purescript[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-ragel-in-ruby-host[167]	MIT License
CommitPackFT	commitpackft-rebol[167]	Apache License 2.0, MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
CommitPackFT	commitpackft-red[167]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-rouge[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-sage[167]	MIT License
CommitPackFT	commitpackft-sas[167]	MIT License
CommitPackFT	commitpackft-scaml[167]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-scilab[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-slash[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-smt[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-solidity[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-sourcepawn[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-squirrel[167]	MIT License
CommitPackFT	commitpackft-ston[167]	MIT License
CommitPackFT	commitpackft-systemverilog[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-unity3d-asset[167]	Apache License 2.0, BSD 3-Clause License, BSD 2-Clause License, MIT License, ISC License, CC0 1.0
CommitPackFT	commitpackft-uno[167]	MIT License
CommitPackFT	commitpackft-unrealscript[167]	MIT License
CommitPackFT	commitpackft-urweb[167]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-vcl[167]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-xbase[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-xpages[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-xproc[167]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-yacc[167]	MIT License, ISC License, BSD 2-Clause License
CommitPackFT	commitpackft-zephir[167]	MIT License
CommitPackFT	commitpackft-zig[167]	MIT License
Dolly 15k	dolly-brainstorming[167]	CC BY-SA 3.0
Dolly 15k	dolly-classification[167]	CC BY-SA 3.0
Dolly 15k	dolly-closedqa[167]	CC BY-SA 3.0
Dolly 15k	dolly-creative_writing[167]	CC BY-SA 3.0
Dolly 15k	dolly-infoextract[167]	CC BY-SA 3.0
Dolly 15k	dolly-openqa[167]	CC BY-SA 3.0
Dolly 15k	dolly-summarization[167]	CC BY-SA 3.0
DialogStudio	ds-ABCD[167]	Apache License 2.0, MIT License
DialogStudio	ds-ATIS[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-ATIS-NER[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-AirDialogue[167]	Apache License 2.0
DialogStudio	ds-AntiScam[167]	Apache License 2.0, CC0 1.0
DialogStudio	ds-BANKING77[167]	Apache License 2.0, CC BY 4.0

Continued on next page

Collection	Dataset Identifier	Licenses
DialogStudio	ds-BANKING77-OOS[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-BiTOD[167]	Apache License 2.0
DialogStudio	ds-CLINC-Single-Domain-OOS-banking[167]	Apache License 2.0, CC BY 3.0
DialogStudio	ds-CLINC-Single-Domain-OOS-credit_cards[167]	Apache License 2.0, CC BY 3.0
DialogStudio	ds-CLINC150[167]	Apache License 2.0, CC BY-SA 3.0
DialogStudio	ds-CaSiNo[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-CoQA[167]	Apache License 2.0, MIT License
DialogStudio	ds-CoSQL[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-ConvAI2[167]	Apache License 2.0
DialogStudio	ds-CraigslistBargains[167]	Apache License 2.0, MIT License
DialogStudio	ds-DART[167]	Apache License 2.0, MIT License
DialogStudio	ds-DSTC8-SGD[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-DialogSum[167]	Apache License 2.0, MIT License
DialogStudio	ds-Disambiguation[167]	Apache License 2.0, MIT License
DialogStudio	ds-FeTaQA[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-GECOR[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-GrailQA[167]	Apache License 2.0
DialogStudio	ds-HDSA-Dialog[167]	Apache License 2.0, MIT License
DialogStudio	ds-HH-RLHF[167]	Apache License 2.0, MIT License
DialogStudio	ds-HWU64[167]	Apache License 2.0, CC BY-SA 3.0
DialogStudio	ds-HybridQA[167]	Apache License 2.0, MIT License
DialogStudio	ds-KETOD[167]	Apache License 2.0, MIT License
DialogStudio	ds-MTOP[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-MULTIWOZ2_2[167]	Apache License 2.0, MIT License
DialogStudio	ds-MulDoGO[167]	Apache License 2.0, CDLA Permissive 1.0
DialogStudio	ds-MultiWOZ_2.1[167]	Apache License 2.0, MIT License
DialogStudio	ds-Prosocial[167]	Apache License 2.0, MIT License
DialogStudio	ds-RESTAURANTS8K[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-SGD[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-SNIPS[167]	Apache License 2.0
DialogStudio	ds-SNIPS-NER[167]	Apache License 2.0
DialogStudio	ds-SParC[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-SQA[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-STAR[167]	Apache License 2.0, MIT License
DialogStudio	ds-Spider[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-TOP[167]	Apache License 2.0, CC BY-SA
DialogStudio	ds-TOP-NER[167]	Apache License 2.0, CC BY-SA
DialogStudio	ds-Taskmaster1[167]	Apache License 2.0, CC BY 4.0

Continued on next page

Collection	Dataset Identifier	Licenses
DialogStudio	ds-Taskmaster2[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-Taskmaster3[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-ToTTo[167]	Apache License 2.0, CC BY-SA 3.0
DialogStudio	ds-TweetSumm[167]	Apache License 2.0, CC0 1.0
DialogStudio	ds-WOZ2_0[167]	Apache License 2.0
DialogStudio	ds-WebQSP[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-WikiSQL[167]	Apache License 2.0, BSD 3-Clause License
DialogStudio	ds-WikiTQ[167]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-chitchat-dataset[167]	Apache License 2.0, MIT License
DialogStudio	ds-wizard_of_internet[167]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-wizard_of_wikipedia[167]	Apache License 2.0, CC BY 4.0
Flan Collection (Chain-of-Thought)	fc-cot-cot_gsm8k[34]	MIT License
Flan Collection (Chain-of-Thought)	fc-cot-cot_strategyqa[59]	CC BY-SA 3.0
Flan Collection (Chain-of-Thought)	fc-cot-stream_creak[127]	MIT License, CC BY-SA 4.0
Flan Collection (Chain-of-Thought)	fc-cot-stream_esnli[21]	MIT License, CC BY-SA 4.0
Flan Collection (Flan 2021)	fc-flan-drop[46]	CC BY 4.0
Flan Collection (Flan 2021)	fc-flan-e2e_nlg[125]	CC BY-SA 4.0
Flan Collection (Flan 2021)	fc-flan-natural_questions[90]	Apache License 2.0, CC BY-SA 3.0
Flan Collection (Flan 2021)	fc-flan-quac[29]	CC BY-SA 4.0
Flan Collection (Flan 2021)	fc-flan-squad_v1[149]	CC BY-SA 4.0
Flan Collection (Flan 2021)	fc-flan-squad_v2[149]	CC BY-SA 4.0
Flan Collection (Flan 2021)	fc-flan-trec[101]	CC0 1.0
Flan Collection (Flan 2021)	fc-flan-true_case[101]	CC0 1.0
Flan Collection (Flan 2021)	fc-flan-wiki_lingua_english_en[91]	CC BY 3.0
Flan Collection (Flan 2021)	fc-flan-winogrande[160]	Apache License 2.0, CC BY 4.0
Flan Collection (Flan 2021)	fc-flan-wnli[99]	CC BY 4.0
Flan Collection (Flan 2021)	fc-flan-word_segment[99]	CC0 1.0
Flan Collection (Flan 2021)	fc-flan-wsc[99]	CC BY 4.0
Flan Collection (P3)	fc-p3-adversarial_qa[11]	CC BY-SA 3.0
Flan Collection (P3)	fc-p3-cos_e[146]	BSD 3-Clause License
Flan Collection (P3)	fc-p3-dbpedia_14[98]	CC BY-SA 3.0
Flan Collection (P3)	fc-p3-hotpotqa[192]	Apache License 2.0, CC BY-SA 4.0
Flan Collection (P3)	fc-p3-quarel[155]	CC BY 4.0
Flan Collection (P3)	fc-p3-quartz[172]	CC BY 4.0
Flan Collection (P3)	fc-p3-quoref[44]	CC BY 4.0
Flan Collection (P3)	fc-p3-web_questions[13]	CC BY 4.0
Flan Collection (P3)	fc-p3-wiki_bio[94]	CC BY-SA 3.0
Flan Collection (P3)	fc-p3-wiki_hop[94]	CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-adversarial_qa[11]	CC BY-SA 3.0

Continued on next page

Collection	Dataset Identifier	Licenses
Flan Collection (Super-NaturalInstructions)	fc-sni-adverserial_qa[11]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-air_dialogue[189]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-ancora_ca_ner[189]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-anem[126]	MIT License, CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-argkp	Apache License 2.0, CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-asian_language_-treebank[153]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-atomic[76]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-bard[54]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-cedr[163]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-circa[113]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-clue_cmrc2018[42]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-coached_conv_pref[141]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-copa_hr	BSD 2-Clause License
Flan Collection (Super-NaturalInstructions)	fc-sni-crows_pairs[124]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-cuad[72]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-defeasible_nli_atomic[157]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-disfl_qa[67]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-e_snli[21]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-gap[188]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-hotpotqa[192]	Apache License 2.0, CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-human_ratings_of_natural_-language_generation_outputs[192]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-hybridqa[26]	CC BY 4.0, MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-iirc[53]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-jigsaw[53]	CC0 1.0

Continued on next page

Collection	Dataset Identifier	Licenses
Flan Collection (Super-NaturalInstructions)	fc-sni-librispeech_asr[129]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-logic2text[27]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-numeric_fused_head[49]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-offenseval_dravidian[49]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-open_pi[174]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-paper_reviews_data_set[174]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-poem_sentiment[165]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-propara[43]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-quarel[155]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-quartz[172]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-quoref[44]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-ro_sts_parallel[48]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-schema_guided_dstc8[150]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-scitail[87]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-scitailv1.1[87]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-semeval_2020_task4[87]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-sms_spam_collection_v.1[87]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-splash[87]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-squad2.0[149]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-squad_1.1[148]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-strategyqa[59]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-universal_dependencies___-english_dependency_treebank[59]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-web_questions[13]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-wiki_hop[94]	CC BY-SA 3.0

Continued on next page

Collection	Dataset Identifier	Licenses
Flan Collection (Super-NaturalInstructions)	fc-sni-wikitext[118]	CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-winograd_wsc[99]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-winomt[170]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-winowhy[196]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-wsc; enhanced_wsc[196]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-wsc_fiexed[196]	CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-xcopa[136]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-xquad[6]	CC BY-SA 4.0
Open Assistant	oasst-en[89]	Apache License 2.0, CC BY 4.0
Open Assistant OctoPack	oasst-en-octopack[89]	Apache License 2.0, CC BY 4.0
Open Assistant v2	oasst2-en[89]	Apache License 2.0
OIG	oig-unified_canadian_-parliament[89]	Apache License 2.0
OIG	oig-unified_cuad[89]	Apache License 2.0, CC BY 4.0
OIG	oig-unified_grade_school_math_-instructions[89]	Apache License 2.0, MIT License
OIG	oig-unified_nq[89]	Apache License 2.0, CC BY-SA 3.0
OIG	oig-unified_sqlv1[89]	Apache License 2.0, CC BY-SA 4.0
OIG	oig-unified_sqlv2[89]	Apache License 2.0, CC BY-SA 4.0
OIG	oig-unified_squad_v2_more_-neg[89]	Apache License 2.0, CC BY-SA 4.0
Tasksource Instruct	tsi-balanced_copa[86]	BSD 2-Clause License
Tasksource Instruct	tsi-breaking_nli[60]	CC BY-SA 4.0
Tasksource Instruct	tsi-cladder[60]	MIT License
Tasksource Instruct	tsi-condaqa[151]	Apache License 2.0
Tasksource Instruct	tsi-conj_nli[159]	MIT License
Tasksource Instruct	tsi-defeasible_nli-atomic[157]	MIT License
Tasksource Instruct	tsi-defeasible_nli-snli[157]	MIT License
Tasksource Instruct	tsi-dynasent-dynabench.dynasent.r1.all-r1[137]	CC BY 4.0
Tasksource Instruct	tsi-dynasent-dynabench.dynasent.r2.all-r2[137]	CC BY 4.0
Tasksource Instruct	tsi-fever_evidence_related-mwong_-_fever_related[179]	CC BY-SA 4.0
Tasksource Instruct	tsi-few_nerd-supervised[45]	CC BY-SA 4.0
Tasksource Instruct	tsi-fig_qa[103]	MIT License
Tasksource Instruct	tsi-fracas[56]	MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
Tasksource Instruct	tsi-hyperpartisan_news[56]	CC BY 4.0
Tasksource Instruct	tsi-lex_glue-case_hold[23]	Apache License 2.0
Tasksource Instruct	tsi-lonli[176]	MIT License
Tasksource Instruct	tsi-moral_stories-full[51]	MIT License
Tasksource Instruct	tsi-neqa[51]	CC BY 4.0
Tasksource Instruct	tsi-prost	Apache License 2.0
Tasksource Instruct	tsi-quote_repetition	CC BY 4.0
Tasksource Instruct	tsi-recast-recast_factuality	CC BY-SA 4.0
Tasksource Instruct	tsi-recast-recast_megaveridicality	CC BY-SA 4.0
Tasksource Instruct	tsi-recast-recast_ner	CC BY-SA 4.0
Tasksource Instruct	tsi-recast-recast_puns	CC BY-SA 4.0
Tasksource Instruct	tsi-recast-recast_sentiment	CC BY-SA 4.0
Tasksource Instruct	tsi-recast-recast_verbcorner	CC BY-SA 4.0
Tasksource Instruct	tsi-recast-recast_verbnet	CC BY-SA 4.0
Tasksource Instruct	tsi-redefine_math	CC BY 4.0
Tasksource Instruct	tsi-tracie[198]	Apache License 2.0
Tasksource Instruct	tsi-truthful_qa-multiple_-choice[102]	Apache License 2.0
Tasksource Instruct	tsi-vitaminc-tals__vitaminc[164]	MIT License
Tasksource Instruct	tsi-winowhy[196]	MIT License
Tasksource Symbol-Tuning	tsy-breaking_nli[60]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-cladder[60]	MIT License
Tasksource Symbol-Tuning	tsy-condaqa[151]	Apache License 2.0
Tasksource Symbol-Tuning	tsy-conj_nli[159]	MIT License
Tasksource Symbol-Tuning	tsy-defeasible_nli-atomic[157]	MIT License
Tasksource Symbol-Tuning	tsy-defeasible_nli-snli[157]	MIT License
Tasksource Symbol-Tuning	tsy-dynasent-dynabench.dynasent.r1.all-r1[137]	CC BY 4.0
Tasksource Symbol-Tuning	tsy-dynasent-dynabench.dynasent.r2.all-r2[137]	CC BY 4.0
Tasksource Symbol-Tuning	tsy-fever_evidence_related-mwong__fever_related[179]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-fracas[56]	MIT License
Tasksource Symbol-Tuning	tsy-hyperpartisan_news[56]	CC BY 4.0
Tasksource Symbol-Tuning	tsy-lonli[176]	MIT License
Tasksource Symbol-Tuning	tsy-recast-recast_factuality[176]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_-megaveridicality[176]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_ner[176]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_puns[176]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_sentiment[176]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_verbcorner[176]	CC BY-SA 4.0

Continued on next page

Collection	Dataset Identifier	Licenses
Tasksource Symbol-Tuning	tsy-recast-recast_verbnet[176]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-tracie[198]	Apache License 2.0
Tasksource Symbol-Tuning	tsy-vitaminc-tals__vitaminc[164]	MIT License
Tasksource Symbol-Tuning	tsy-winowhy[196]	MIT License

E List of News sources

The Common Pile contains a variety of openly licensed news sources released under CC BY and CC BY-SA licenses. The sources licensed under CC BY include: 360info, Africa is a Country, Alt News, Balkan Diskurs, Factly, Freedom of the Press Foundation, Agenzia Fides, Global Voices, Meduza, Mekong Eye, Milwaukee Neighborhood News Service, Minority Africa, New Canadian Media, SciDev.Net, The Solutions Journalism Exchange, Tasnim News Agency, and ZimFact. The sources licensed under CC BY-SA include: Oxpeckers, Propastop, and The Public Record.

F List of WikiMedia wikis

Official Wikimedia wikis are released under a CC BY-SA license. The Common Pile includes the following Wikimedia wikis: Wikipedia, Wikinews, Wikibooks, Wikiquote, Wikisource, Wikiversity, Wikivoyage, and Wiktionary.

G CCCC Source Statistics

We provide additional statistics on the CCCC subset of the Common Pile, including the number of unicode words and documents sourced from each Common Crawl snapshot, in Table 2.

Table 2: Counts of words and documents extracted from 52 snapshots after filtering with our pipeline.

Snapshot	Unicode Words	Documents
CC-MAIN-2013-20	3,851,018,197	5,529,294
CC-MAIN-2013-48	4,544,197,252	6,997,831
CC-MAIN-2014-10	4,429,217,941	6,682,672
CC-MAIN-2014-15	4,059,132,873	5,912,779
CC-MAIN-2014-23	5,193,195,765	8,253,690
CC-MAIN-2014-35	4,254,690,945	6,551,673
CC-MAIN-2014-41	4,289,814,449	6,558,170
CC-MAIN-2014-42	3,986,284,741	6,144,797
CC-MAIN-2014-49	3,316,075,452	4,699,472
CC-MAIN-2014-52	4,307,765,289	6,338,983
CC-MAIN-2015-06	3,675,982,679	5,181,955
CC-MAIN-2015-11	3,932,442,900	5,438,533
CC-MAIN-2015-14	3,658,107,765	4,954,273

Continued on next page

Snapshot	Unicode Words	Documents
CC-MAIN-2015-18	4,451,734,946	6,319,757
CC-MAIN-2015-22	4,285,945,319	5,949,267
CC-MAIN-2015-27	3,639,904,128	4,975,152
CC-MAIN-2016-07	1,588,496,703	3,798,207
CC-MAIN-2016-18	3,228,754,200	4,446,815
CC-MAIN-2016-22	3,217,827,676	4,242,762
CC-MAIN-2017-04	3,852,699,213	5,239,605
CC-MAIN-2017-09	4,186,915,498	5,119,171
CC-MAIN-2017-13	4,950,110,931	5,923,670
CC-MAIN-2017-17	4,684,050,830	5,645,725
CC-MAIN-2017-22	4,683,569,278	5,514,717
CC-MAIN-2017-26	4,744,689,137	5,514,047
CC-MAIN-2017-51	1,981,004,306	2,529,289
CC-MAIN-2018-13	4,816,417,930	5,520,099
CC-MAIN-2018-22	3,921,533,251	4,401,956
CC-MAIN-2018-26	4,506,583,931	4,916,546
CC-MAIN-2018-30	4,936,722,403	5,282,886
CC-MAIN-2018-34	3,865,953,978	3,808,725
CC-MAIN-2018-47	3,933,439,841	3,637,947
CC-MAIN-2018-51	4,745,124,422	4,616,832
CC-MAIN-2019-04	4,475,679,190	4,140,277
CC-MAIN-2019-09	4,287,868,800	4,142,190
CC-MAIN-2019-13	3,966,330,348	3,849,631
CC-MAIN-2019-30	4,179,526,188	4,430,572
CC-MAIN-2019-35	5,144,426,270	5,048,106
CC-MAIN-2019-39	4,572,972,457	4,527,430
CC-MAIN-2020-29	5,200,565,501	4,984,248
CC-MAIN-2020-34	4,458,827,947	4,297,009
CC-MAIN-2021-17	1,768,757,386	1,824,942
CC-MAIN-2021-39	4,599,961,675	4,287,356
CC-MAIN-2021-43	5,337,349,331	5,304,846
CC-MAIN-2021-49	3,980,018,773	4,050,641
CC-MAIN-2022-05	4,517,850,019	4,503,863
CC-MAIN-2023-06	5,135,614,227	4,959,915
CC-MAIN-2023-14	5,117,143,765	4,675,097
CC-MAIN-2023-23	5,461,486,807	4,869,627
CC-MAIN-2023-50	5,881,860,014	4,901,306
CC-MAIN-2024-10	5,164,171,562	4,335,071
CC-MAIN-2024-18	4,745,457,054	3,949,186
Total	**221,715,271,483**	**259,728,610**

H PeS2o Source Statistics

Additional statistics on the composition of the peS2o subset of the Common Pile can be found in Tables 3 and 4.

Table 3: Distribution of licenses in the peS2o subset.

License	Train Split	Validation Split
CC BY	6,088,325	37,754
CC BY-SA	120,150	1,231
CC0	36,373	121
Public domain	10,060	6

Table 4: Distribution of papers across 23 fields of study, as identified by the Semantic Scholar API [88]. A paper may belong to one or more fields of study.

Field of Study	Train Split	Validation Split
Medicine	2,435,244	23,734
Biology	1,518,478	8,879
Environmental Science	993,499	7,601
Engineering	656,021	5,005
Computer Science	462,320	3,003
Materials Science	416,045	3,166
Physics	413,461	1,285
Chemistry	406,429	2,781
Psychology	364,441	2,126
Education	220,014	1,532
Business	193,536	946
Economics	185,716	921
Agricultural and Food Sciences	333,776	2,013
Sociology	137,257	1,535
Mathematics	135,676	199
Political Science	106,748	378
Geology	67,258	217
Geography	44,269	257
Linguistics	41,737	228
History	36,848	192
Law	30,888	251
Philosophy	27,518	148
Art	26,658	75

I Growth rates of openly licensed data

Over time, the volume of openly licensed data continues to grow as more creators release content under open licenses. In Figure 5, we quantify this growth between 2010 and 2024 by analyzing subsets of the Common Pile for which reliable creation date metadata is available. We plot the cumulative proportion of data created up to various cutoff dates and find that approximately half of the Common Pile (around 3.8TB) was created since 2020. This trend provides insight into the growing availability of openly licensed data and suggests a promising trajectory for future LLMs trained entirely on openly licensed sources.

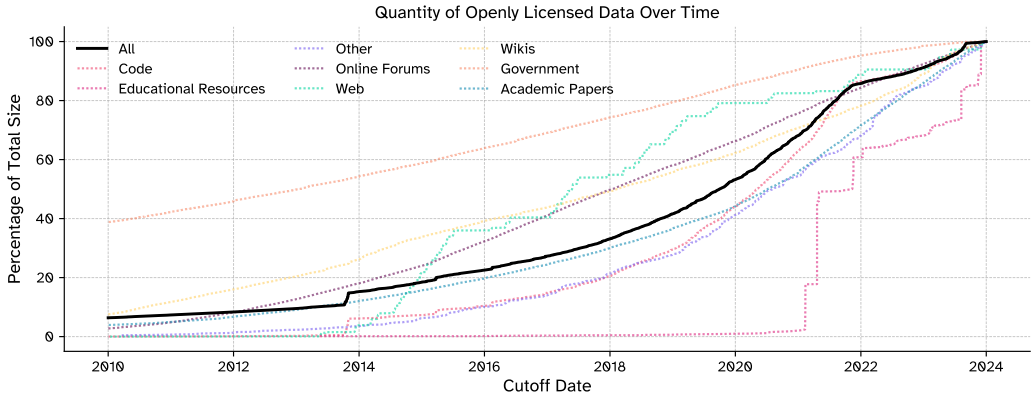

Figure 5: **The amount of openly licensed text grows steadily over time.** We visualize the cumulative proportion of data created up to various cutoff dates for sources in the Common Pile with reliable creation date metadata. This includes all sources except for the Caselaw Access Project, Data Provenance Initiative, and the sources covering early 20th century Public Domain books.

J Details on filtering pipelines

In Section 4.1, we detail the steps used to produce the Comma training dataset from the raw text in the Common Pile. These include applying filters based on language, text quality, length, likelihood, and toxicity; removing various forms of PII; and removal of source-specific boilerplate text using regular expressions. The Common Pile contains a diverse range of sources and we therefore design separate filtering thresholds for each source. The exact source-specific thresholds used to post-process the Common Pile can be found in Table 5. Additionally, statistics on the pre- and post-filtered sizes of each source can be found in Table 6.

Table 5: Pre-processing pipelines applied to each source in the Common Pile to construct the Comma 7B pre-training dataset.

Source	Language	Text Quality	Doc Length	Log-Likelihood	Toxicity	PII	Regex Filter
ArXiv Abstracts	–	–	–	–	–	Y	N
ArXiv Papers	> 0.5	–	–	–	–	Y	N
Biodiversity Heritage Library	> 0.5	–	> 100	> -20	–	N	Y
Caselaw Access Project	–	–	> 100	–	> 0.1	Y	N
CC Common Crawl	> 0.5	> 0.0001	> 100	–	> 0.1	Y	N

Continued on next page

Source	Language	Text Quality	Doc Length	Log-Likelihood	Toxicity	PII	Regex Filter
Data Provenance Initiative	–	–	–	–	–	N	N
Database of Open Access Books	> 0.5	–	> 200	–	> 0.1	Y	N
Foodista	> 0.5	–	> 100	–	–	N	N
GitHub Archive	> 0.5	–	> 100	–	> 0.1	Y	N
Library of Congress	–	–	–	> -20	> 0.1	N	Y
LibreTexts	> 0.5	–	> 700	–	> 0.1	Y	N
News	> 0.5	–	> 100	–	–	Y	N
OERCommons	> 0.5	–	> 300	–	> 0.1	Y	N
peS2o	–	–	–	–	–	Y	N
Pre-1929 Books	–	–	–	> -20	> 0.1	N	Y
PressBooks	> 0.5	–	> 600	–	> 0.1	Y	N
Project Gutenberg	> 0.5	–	–	> -20	–	N	N
Public Domain Review	–	–	> 100	–	–	Y	N
PubMed	> 0.5	–	> 100	–	–	Y	N
PEPs	–	–	–	–	–	Y	N
Regulations.gov	–	–	> 100	–	–	Y	Y
StackExchange	> 0.5	–	–	–	–	Y	N
Ubuntu IRC	> 0.5	–	> 100	–	> 0.1	Y	N
UK Hansard	> 0.5	–	–	–	–	Y	N
USGPO	–	–	–	–	–	N	Y
USPTO	–	–	> 100	> -20	–	Y	N
Wikimedia	> 0.5	–	> 100	–	–	Y	N
Wikiteam	> 0.5	–	> 700	–	> 0.1	Y	N
CC YouTube	> 0.5	–	> 100	–	> 0.1	Y	N

Table 6: Raw and filtered sizes of the Common Pile's constituent datasets.

	Document Count		Size (GB)	
Source	Raw	Filtered	Raw	Filtered
ArXiv Abstracts	2,538,935	2,504,679	2.4	2.4
ArXiv Papers	321,336	304,048	21	19
Biodiversity Heritage Library	42,418,498	15,111,313	96	35
Caselaw Access Project	6,919,240	6,735,525	78	77

Continued on next page

	Document Count		Size (GB)	
Source	**Raw**	**Filtered**	**Raw**	**Filtered**
CC Common Crawl	51,054,412	6,852,137	260	58
Data Provenance Initiative	9,688,211	3,508,518	7	3
Directory of Open Access Books	474,445	403,992	12.5	12
Foodista	72,090	65,640	0.09	0.08
GitHub Archive	30,318,774	23,358,580	54.7	40.4
Library of Congress	135,500	129,052	47.8	35.6
LibreTexts	62,269	40,049	5.3	3.6
News	172,308	126,673	0.4	0.3
OERCommons	9,339	5,249	0.1	0.05
peS2o	6,294,020	6,117,280	188.2	182.6
Pre-1929 Books	137,127	124,898	73.8	46.3
PressBooks	106,881	54,455	1.5	0.6
Project Gutenberg	71,810	55,454	26.2	20.1
Public Domain Review	1,412	1,406	0.007	0.007
PubMed	4,068,867	3,829,689	158.9	147.1
PEPs	656	655	0.01	0.01
Regulations.gov	225,196	208,301	6.1	5.1
StackExchange	33,415,400	30,987,814	103.7	89.7
Stack V2	218,364,133	69,588,607	4774.7	259.9
Ubuntu IRC	329,115	234,982	6.3	5.3
UK Hansard	51,552	47,909	10	9.6
USGPO	2,732,677	2,148,548	74.5	36.1
USPTO	20,294,152	17,030,231	1003.4	661.1
Wikimedia	63,969,938	16,311,574	90.5	57.4
Wikiteam	219,139,368	26,931,807	437.5	13.7
CC YouTube	1,129,692	998,104	21.5	18.6
Total	**692,854,953**	**233,817,169**	**7557.9**	**1838.3**

K Details on Comma's pre-training data mixture

We estimated the quality of each source in the Common Pile by training a 1.7B-parameter model for 28B tokens on each source individually and evaluating the resulting models on the set of "early signal" tasks from [134]. In doing so, we found that the amount of text in each source was poorly correlated with text quality, motivating the use of heuristic mixing weights to up-/down-weight different sources in our pre-training mix. In Table 7 we list the pre-training mixture weights for each of the sources in the Common Pile.

Table 7: Overview of the data mixing used to up/down-weight individual sources in the Common Pile to construct the Comma pre-training dataset.

Source	Size (GB)	Repeats	Effective Size (GB)	Tokens (Billions)	Percentage
ArXiv Abstracts	2.4	6	14.4	3.6	0.360%
ArXiv Papers	19.5	6	117	29.3	2.932%
Biodiversity Heritage Library	35.5	0.25	8.9	2.2	0.220%
Caselaw Access Project	77.5	1	77.5	19.4	1.941%
CC Common Crawl	58.1	6	348.6	87.1	8.716%
Data Provenance Initiative	3.4	6	20.4	5.1	0.510%
Database of Open Access Books	12	6	72	18	1.801%
Foodista	0.08	6	0.48	0.12	0.012%
GitHub Archive	40.4	6	242.4	60.6	6.064%
Library of Congress	35.6	0.25	8.9	2.2	0.220%
LibreTexts	3.6	6	21.6	5.4	0.540%
News	0.25	6	1.5	0.38	0.038%
OERCommons	0.05	6	0.3	0.08	0.008%
peS2o	182.6	6	1,095.6	273.9	27.409%
Pre-1929 Books	46.3	1	46.3	11.6	1.161%
PressBooks	0.6	6	3.6	0.9	0.090%
Project Gutenberg	20.1	1	20.1	5	0.500%
Public Domain Review	0.007	6	0.04	0.01	0.001%
PubMed	147.1	1	147.1	36.8	3.683%
PEPs	0.01	6	0.06	0.02	0.002%
Regulations.gov	5.1	6	30.6	7.6	0.761%
StackExchange	89.7	6	538.2	134.6	13.469%
Stack V2	259.9	2	519.8	130	13.009%
Ubuntu IRC	5.3	6	31.8	7.9	0.791%

Continued on next page

Source	Size (GB)	Repeats	Effective Size (GB)	Tokens (Billions)	Percentage
UK Hansard	9.6	6	57.6	14.4	1.441%
USGPO	36.1	0.25	9	2.3	0.230%
USPTO	661.1	0.25	165.3	41.3	4.133%
Wikimedia	57.4	6	344.4	86.1	8.616%
Wikiteam	13.7	4	54.8	13.7	1.371%
CC YouTube	18.6	1	18.6	4.7	0.470%
Total	**1838.3**	**–**	**3997.4**	**999.3**	**100%**

L Details on Comma's cool-down data mixture

Following Hu et al. [74], we end training with a "cool-down" where we train on 37.7B tokens of high-quality data while linearly decaying the learning rate to 0. We provide the source mixture weights for this cool-down phase in Table 8.

Table 8: Overview of the data mixing used to up/down-weight individual sources in the Common Pile to construct the training distribution for Comma's cool-down phase.

Source	Size (GB)	Repeats	Effective Size (GB)	Tokens (Billions)	Percentage
ArXiv Papers	19.5	0.5	9.8	2.4	6.50%
CC Common Crawl	58.1	0.3	17.4	4.4	11.63%
Data Provenance Initiative	3.4	2	6.8	1.7	4.55%
Database of Open Access Books	12	2	24	6	16.04%
Foodista	0.08	2	0.16	0.04	0.11%
LibreTexts	3.6	2	7.2	1.8	0.48%
News	0.25	2	0.5	0.13	0.33%
OERCommons	0.05	2	0.1	0.03	0.07%
peS2o	182.6	0.1	18.3	4.6	12.18%
PressBooks	0.6	2	1.2	0.3	0.77%
Public Domain Review	0.007	2	0.014	0.004	0.01%
PEPs	0.01	2	0.02	0.005	0.02%
StackExchange	89.7	0.25	22.4	5.6	14.96%

Continued on next page

Source	Size (GB)	Repeats	Effective Size (GB)	Tokens (Billions)	Percentage
Stack V2	259.9	0.1	26.0	6.5	17.04%
Wikimedia	57.4	0.4	23	5.7	15.32%
Total	**679.4**	**–**	**149.9**	**37.5**	**100%**

M Details on small-scale data ablations

In Section 4.3 we report results from a series of small-scale data ablations where we identically trained 1.7B parameter models on various openly licensed and unlicensed datasets and evaluate their performance on the "early signal" tasks from Penedo et al. [134] to compare their data quality against the Common Pile. In Figure 6 we show how the performance of these models evolve over the course of their training run, highlighting that differences in data quality become apparent very early in training. Additionally, we provide exact numerical results for each model in Table 9, showing that the Common Pile has higher data quality than any previously released openly licensed datasets and the Pile, and nearly matches the data quality of the OSCAR dataset. To validate that this is not purely due to the presence of high-quality supervised fine-tuning data from the Data Provenance Initiative (DPI) data source, we also perform an ablation on the Common Pile excluding the DPI data and find that the final performance of this model is largely unchanged.

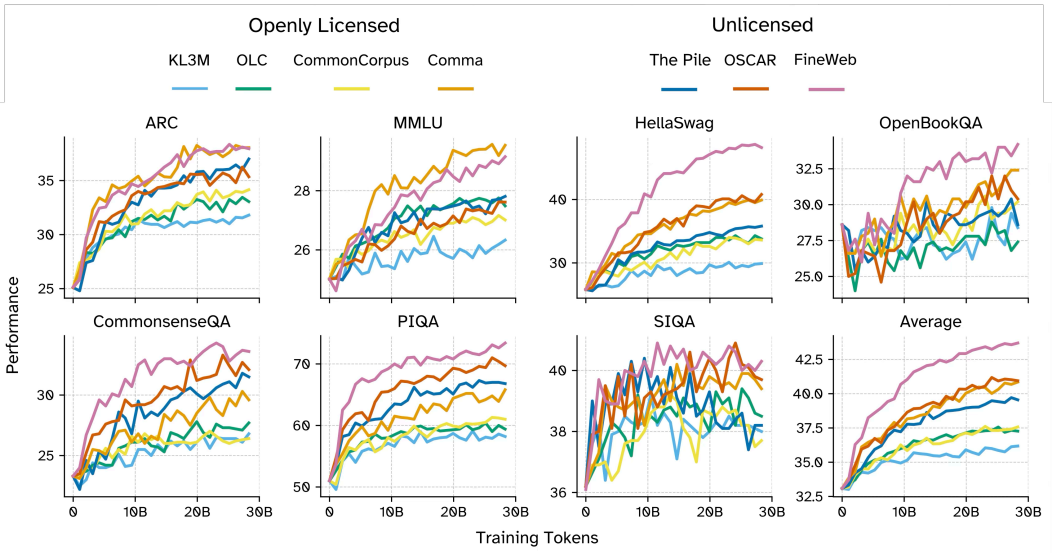

Figure 6: **Comma consistently outperforms models trained on other corpora of openly licensed text and outperforms the Pile on all but two tasks.** We train 1.7B parameter models on 28B tokens following Penedo et al. [134].

N Additional Comma results

We provide exact numerical results for Comma v0.1 and baseline models across a variety of knowledge, reasoning, and coding tasks (Table 10). We find that particularly on knowledge-based benchmarks, such as ARC-C and MMLU, and coding benchmarks, Comma outperforms baseline models trained on an equivalent amount (1T tokens) of unlicensed text.

Table 9: **Comma's training dataset has higher quality than previous openly-licensed datasets and unlicensed datasets like the Pile.** In the small-scale (1.7B parameter) data ablation setting, we find that Comma's training dataset yields better models than previous openly licensed datasets and the Pile, and nearly matches the performance of models trained on OSCAR. Additionally, we find that removing the high-quality supervised data from the Data Provenance Initiative has marginal affect on the Comma dataset's overall quality.

Dataset	ARC	MMLU	HS	OBQA	CSQA	PIQA	SIQA	Avg.
KL3M	31.8	26.3	29.9	28.4	26.8	58.2	38.0	36.2
OLC	33.1	27.5	33.8	27.4	27.7	59.4	38.5	37.3
Common Corpus	34.2	27.0	33.6	30.2	26.4	61.0	37.7	37.6
Comma (no DPI)	37.7	28.7	37.6	31.0	30.8	63.8	39.8	40.0
Comma	38.0	29.5	39.9	32.4	29.6	65.8	39.4	40.8
The Pile	37.0	27.8	35.8	28.6	31.5	66.8	38.2	39.6
OSCAR	35.4	27.6	40.8	30.4	32.1	69.7	39.7	40.9
FineWeb	38.0	29.1	48.2	34.2	33.6	73.4	40.3	43.7

Table 10: Comparison between Comma and baseline models trained with similar resources (7 billion parameters, 1 trillion tokens) across a variety of knowledge, reasoning, and coding benchmarks.

Model	ARC-C	ARC-E	MMLU	BoolQ	HS	OBQA	CSQA	PIQA	SIQA	HEval	MBPP	Avg.
RPJ-INCITE	42.8	68.4	27.8	68.6	70.3	49.4	57.7	76.0	46.9	11.1	15.9	48.6
LLaMA	44.5	67.9	34.8	75.4	76.2	51.2	61.8	77.2	50.3	19.9	27.9	53.4
StableLM	50.8	65.4	45.2	71.7	75.6	48.2	57.2	77.0	48.2	23.1	32.0	54.0
MPT	46.5	70.5	30.2	74.2	77.6	48.6	63.3	77.3	49.1	27.3	33.2	54.3
OpenLLaMA	44.5	67.2	40.3	72.6	72.6	50.8	62.8	78.0	49.7	27.6	33.9	54.5
Comma v0.1	52.8	68.4	42.4	75.7	62.6	47.0	59.4	70.8	50.8	36.5	35.5	54.7
Qwen3	57.2	74.5	77.0	86.1	77.0	50.8	66.4	78.2	55.0	94.5	67.5	71.3

O Additional training runs

To explore the sensitivity of our Comma v0.1 results to hyperparameter choices, we perform a series of additional 7B parameter/1T token training runs on AMD MI300A GPUs with slight alterations to the training recipe. Due to both a desire to reach the same 1T token target rapidly, and the lower single-GPU throughput on the system available for these ablations, for all additional runs the the training batch size is 8.3M (2^{23}) versus the 2.1M (2^{21}tokens per step of Comma v0.1. Unless otherwise specified, we did not use the two phase training process described in Section 4.4 (i.e. no separate high-quality cooldown phase is run and we do not perform checkpoint averaging at the end of training and before evaluation).

O.1 Ablations at 1T Tokens

We first performed a set of training runs for 125,000 steps, resulting in 1.048T total tokens (referred to as "1T" for brevity).

"8M Batch" We perform a run with nearly the same training hyperparameters as Comma v0.1, except with a larger 8M token batch size. We also use a single phase training setup; the base data mixture (Table 7) is run for the entire duration to 1T tokens. The learning rate schedule is 2,000 steps of warm-up from 0 to a peak of $1e-3$ with 123,000 steps of decay to a minimum of $1.8e-9$.

"Curriculum" In this experimental run, a different data mixture is used in each of three training stages of equal duration (we also use the modified hyperparameters from Batch Size Ablation above). The first stage of the curriculum comprises data from only the Common Pile's largest sources (mostly USPTO, Table 13). The second stage uses the same data mixture as Comma v0.1's main pre-training phase ("phase I"), but run for only 1/3 of the duration. Finally, the third and last stage of the curriculum up-weights Common Pile's highest quality, benchmark-relevant sources (Table 14).

We provide exact numerical results for Comma v0.1 and alternate Comma runs performed with different hyperparameters and data mixture curricula across a variety of knowledge, reasoning, and coding benchmarks (Table 11). We find that the 8M Batch and Curriculum ablations are roughly comparable on average to the main Comma v0.1 run, with the notable exception that both ablations slightly outperform Comma v0.1 on the coding benchmarks. We conclude that the benchmark results reported for Comma v0.1 in Section 4.4 seem relatively robust to minor changes in training hyperparameters, dataset mixture curriculum (assuming similar amounts of most data splits appear at some time during training), and the software environment and GPU hardware used to train the model.

Table 11: Comparison between our main training run and alternate runs performed with different hyperparameters and data mixture curricula across a variety of knowledge, reasoning, and coding benchmarks. For "Main", we report the performance of Comma without averaging the cooldownc checkpoints for ease of comparison.

Model	ARC-C	ARC-E	MMLU	BoolQ	HS	OBQA	CSQA	PIQA	SIQA	HEval	MBPP	Avg.
Curriculum	45.2	69.1	41.4	74.7	60.8	46.8	59.1	70.5	48.6	38.1	34.6	53.5
8M Batch	47.2	69.6	42.9	69.9	62.9	47.0	56.9	70.4	50.5	36.8	37.2	53.8
Main	50.8	68.4	40.2	72.9	62.3	46.2	59.5	71.0	51.2	32.1	34.6	53.6

O.2 Ablations at 2T Tokens

Compared to contemporary models [178, 62, 5], the main results are limited to a relatively small pre-training token budget to ensure that we retain sufficient data after filtering the Common Pile. To test whether the filtered dataset supports training durations beyond 1T, we complete a comparable training run to Comma v0.1 except we repeat the same data mixture approximately twice. We note that this pre-training mixture involves repeating certain sources an excessive number of times (up to 16 passes for some sources). Prior work suggests that these extreme levels of data repetition may result in diminishing returns [123]. However, these experiments still give us a preliminary picture of the performance achievable under a larger budget.

We performed a 2T token training run using the same two-phase training procedure as Comma v0.1. Due to the same hardware efficiency constraints described in Appendix O, we train at the larger 8.3M token batch size for all extended runs. In this configuration, a full duration run takes 250,000 steps. We used a base learning rate schedule with 2,000 steps of warm-up from 0 to a peak of $2e-3$ (scaling up by a factor of 2 due to the $4\times$ increase in batch size [117]) with 248,000 steps of decay to a minimum of $3.6e-9$. Similarly to the note in Appendix O.1, the batch size of 8.3M over a full run of 250,000 steps would actually correspond to 2.1T tokens. Therefore, we break the run into a 230,000 step phase 1, and a 9,000 step phase 2 choosing the step counts to train for almost exactly 2T tokens. Phase 1 uses the base data mix (Table 7) training under a cosine decay schedule and phase 2 uses the high-quality cooldown mix (Table 8) but with a learning rate that decays linearly to 0 starting from the same value the cosine schedule yields at step 230,000 (3.19e-5). Finally, we report results for "Comma v0.1 2T" as the average of the 10 evenly spaced intermediate checkpoints of the phase 2 cooldown using the same per-parameter averaging strategy as in Section 4.4.

We compare to the following budget-matched (7 billion parameter, 2 trillion token) base models: OLMo Twin (specifically OLMo-7B-Twin-2T) [64], Llama 2 [182], DeepSeekLLM [16], and Trillion [69]. The performance of each of these models on our chosen benchmarks is provided in Table 12. Notably, we find that the 2 trillion token variants of Comma are competitive with OLMo, Llama 2, and DeepSeekLLM, with especially strong performance on MMLU, BoolQ, ARC-E, and the coding tasks. However, the 2T variant of Comma generally underperforms Trillion. Trillion is a recently released model with a sophisticated data filtering pipeline that removes most data with the aim of training a performant model on relatively small token budget. We emphasize that the Comma v0.1 2T result here is likely *not* a best-case 2T-token run using the Common Pile v0.1 due to excessive repetition, and better performance could likely be attained through a 2T-specific mixture and curriculum. However, these results further support the strength of the Common Pile as a source of pre-training data but further emphasize the need to scale up data collection efforts in support of more sophisticated data filtering and longer training durations.

Table 12: Comparison between Comma v0.1, extended Comma runs to 2T tokens, and a variety of baseline models trained form 2T tokens across a variety of knowledge, reasoning, and coding benchmarks.

Model	ARC-C	ARC-E	MMLU	BoolQ	HS	OBQA	CSQA	PIQA	SIQA	HEval	MBPP	Avg.
OLMo Twin	45.2	67.5	28.2	71.7	73.4	48.0	61.8	77.9	48.5	18.2	27.5	51.6
Llama 2	48.5	69.5	45.8	80.2	76.2	48.4	62.8	76.7	50.8	26.1	28.5	55.8
Comma v0.1 2T	45.8	71.8	49.8	78.6	64.4	46.2	64.0	72.5	52.3	44.2	41.5	57.4
DeepSeekLLM	49.5	67.7	48.5	71.7	74.1	52.0	66.6	77.8	51.6	43.1	43.8	58.8
Trillion	57.2	77.0	62.8	83.7	80.4	54.4	69.4	78.9	56.0	74.8	52.7	67.9

Table 13: Overview of the data mixing used to up/down-weight individual sources for the Stage 1 of the Curriculum ablation run (Appendix O.1). In this table we omit the size columns for brevity. Refer to main mixture tables to compare the sizes of different subsets (Tables 7 and 8).

Source	**Repeats**	**Tokens (Billions)**	**Percentage**
USPTO	1.4125	233.5	66.81%
Pre-1929 Books	5.65	65.4	18.71%
Stack V2 (HTML)	11.3	12.8	3.65%
USGPO	1.41	12.8	3.65%
Library of Congress	1.41	12.6	3.59%
Biodiversity Heritage Library	1.41	12.52	3.58%
Total	**–**	**349.4**	**100%**

Table 14: Overview of the data mixing used to up/down-weight individual sources for the Stage 3 of the Curriculum ablation run (Appendix O.1). In this table we omit the size columns for brevity. Refer to main mixture tables to compare the sizes of different subsets (Tables 7 and 8).

Source	**Repeats**	**Tokens (Billions)**	**Percentage**
Stack V2	1	63.8	18.519%
Database of Open Access Books	6	18	5.230%
Wikimedia	6	86.1	24.981%
StackExchange	2.5	56.1	16.259%
peS2o	1	45.6	13.241%

Continued on next page

Source	Repeats	Tokens (Billions)	Percentage
CC Common Crawl	3	43.6	12.638%
ArXiv Papers	5	24.4	7.063%
Data Provenance Initiative	6	5.1	1.485%
PressBooks	6	0.87	0.251%
LibreTexts	6	0.54	0.157%
News	6	0.37	0.108%
Foodista	6	0.12	0.036%
OERCommons	6	0.08	0.023%
PEPs	6	0.02	0.005%
Public Domain Review	6	0.01	0.003%
Total	**–**	**344.7**	**100%**