Appendix

Table of Contents

A Contributions

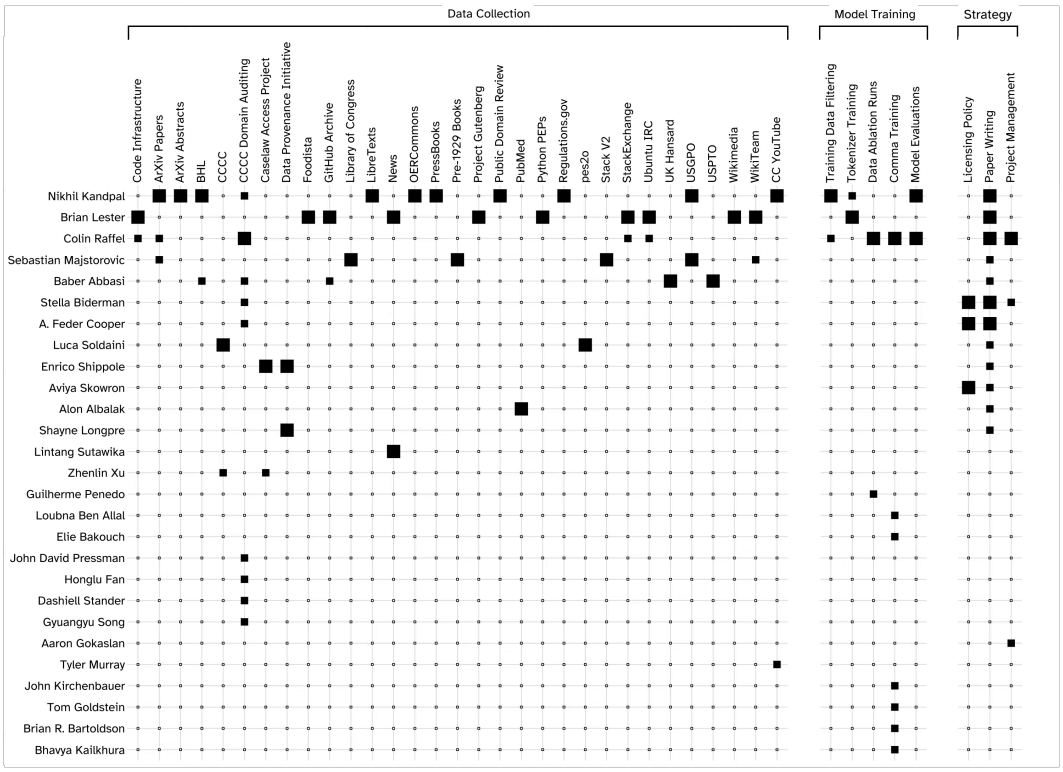

Figure 4: **Author contributions to the Common Pile and Comma.** Large squares indicate a major contribution and small squares indicate a supporting contribution.

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. 56, 183, 167] 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 distributed under open licenses.

peS2o To ensure broad coverage across many scientific disciplines, we include a version of peS2o [166] restricted to openly licensed articles. pes2o is derived from S2ORC [102], a corpus of openly licensed abstract and full-text papers that have been converted to a structured format using Grobid [1]. 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 [85]. 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 I.

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 nXML 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 [11] 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 dumps of all of their content, since July of 2024, StackExchange has stopped publishing XML 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 the 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 appeared to be missing "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 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. Parents 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 [75], 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 [130], 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 [179].

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 [104, 107]. 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 [112]. 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 [56], Dolma [167], and RedPajama [183]—include content from books. In the United States, books published prior to 1929 are in the public domain since 2024 due to copyright expiration. 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 have 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![140] dataset, which only includes books that are over 100 years old. Minimal preprocessing is applied to remove the Project Gutenberg header and footers, but 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 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 G 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 for 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 regexs.

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 We filter the Stack V2 [111] to only include code from openly licensed repositories, based 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 content. Finally, each video was transcribed to text using the Whisper speech recognition model [138].

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 [16]. Then, using a regular expression adapted from the C4Corpus project [66], 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 applied to the all text content rather than embedded media or a subset of the text on the domain. We extract the main content of these documents and remove boilerplate using Resiliparse [15]. 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 [167]; namely, we use C4-derived heuristics [142] to filter pages containing Javascript, Lorem Ipsum, and curly braces {}. We also apply all Gopher rules [141] to remove low-quality pages. We provide more details on the composition of this subset in Appendix H.

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 F. 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 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.

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

Collection	Dataset Identifier	Licenses
AgentInstruct	AgentInstruct-alfworld[164]	MIT License
HelpSteer	HelpSteer[182]	CC BY 4.0
Aya Dataset	aya-english[165]	Apache License 2.0
CommitPackFT	commitpackft-abap[165]	MIT License
CommitPackFT	commitpackft-agda[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-apl[165]	MIT License, ISC License
CommitPackFT	commitpackft-arc[165]	MIT License
CommitPackFT	commitpackft-aspectj[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-ats[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-blitzmax[165]	MIT License
CommitPackFT	commitpackft-bluespec[165]	MIT License
CommitPackFT	commitpackft-boo[165]	MIT License
CommitPackFT	commitpackft-brainfuck[165]	Apache License 2.0, BSD 2-Clause License, MIT License
CommitPackFT	commitpackft-bro[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-cartocss[165]	MIT License
CommitPackFT	commitpackft-chapel[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-clean[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-coldfusion[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-creole[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-crystal[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-dns-zone[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-dylan[165]	MIT License
CommitPackFT	commitpackft-eiffel[165]	MIT License
CommitPackFT	commitpackft-emberscript[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-fancy[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-flux[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-forth[165]	MIT License
CommitPackFT	commitpackft-g-code[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-gdscript[165]	Apache License 2.0, CC0 1.0, MIT License
CommitPackFT	commitpackft-genshi[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-graphql[165]	Apache License 2.0, BSD 3-Clause License, CC0 1.0, MIT License
CommitPackFT	commitpackft-harbour[165]	MIT License
CommitPackFT	commitpackft-hlsl[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-http[165]	Apache License 2.0, MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
CommitPackFT	commitpackft-idris[165]	MIT License, BSD 3-Clause License, BSD 2-Clause License
CommitPackFT	commitpackft-igor-pro[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-inform-7[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-ioke[165]	MIT License
CommitPackFT	commitpackft-isabelle[165]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-jflex[165]	MIT License
CommitPackFT	commitpackft-json5[165]	MIT License, BSD 3-Clause License, BSD 2-Clause License
CommitPackFT	commitpackft-jsonld[165]	Apache License 2.0, BSD 3-Clause License, CC0 1.0, MIT License
CommitPackFT	commitpackft-krl[165]	MIT License
CommitPackFT	commitpackft-latte[165]	MIT License
CommitPackFT	commitpackft-lean[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-lfe[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-lilypond[165]	MIT License
CommitPackFT	commitpackft-liquid[165]	Apache License 2.0, CC0 1.0, MIT License
CommitPackFT	commitpackft-literate-agda[165]	MIT License
CommitPackFT	commitpackft-literate-coffeescript[165]	MIT License
CommitPackFT	commitpackft-literate-haskell[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-llvm[165]	Apache License 2.0, BSD 3-Clause License, BSD 2-Clause License, MIT License
CommitPackFT	commitpackft-logos[165]	Apache License 2.0, BSD 3-Clause License, BSD 2-Clause License, MIT License, ISC License
CommitPackFT	commitpackft-lsl[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-maple[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-mathematica[165]	MIT License, CC0 1.0
CommitPackFT	commitpackft-metal[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-mirah[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-monkey[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-moonscript[165]	MIT License
CommitPackFT	commitpackft-mtml[165]	MIT License
CommitPackFT	commitpackft-mupad[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-nesc[165]	MIT License
CommitPackFT	commitpackft-netlinx[165]	MIT License
CommitPackFT	commitpackft-ninja[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-nit[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-nu[165]	Apache License 2.0, MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
CommitPackFT	commitpackft-ooc[165]	MIT License
CommitPackFT	commitpackft-openscad[165]	MIT License, CC0 1.0, BSD 2-Clause License
CommitPackFT	commitpackft-oz[165]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-pan[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-piglatin[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-pony[165]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-propeller-spin[165]	MIT License
CommitPackFT	commitpackft-pure-data[165]	MIT License
CommitPackFT	commitpackft-purebasic[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-purescript[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-ragel-in-ruby-host[165]	MIT License
CommitPackFT	commitpackft-rebol[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-red[165]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-rouge[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-sage[165]	MIT License
CommitPackFT	commitpackft-sas[165]	MIT License
CommitPackFT	commitpackft-scaml[165]	MIT License, BSD 2-Clause License
CommitPackFT	commitpackft-scilab[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-slash[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-smt[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-solidity[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-sourcepawn[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-squirrel[165]	MIT License
CommitPackFT	commitpackft-ston[165]	MIT License
CommitPackFT	commitpackft-systemverilog[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-unity3d-asset[165]	Apache License 2.0, BSD 3-Clause License, BSD 2-Clause License, MIT License, ISC License, CC0 1.0
CommitPackFT	commitpackft-uno[165]	MIT License
CommitPackFT	commitpackft-unrealscript[165]	MIT License
CommitPackFT	commitpackft-urweb[165]	MIT License, BSD 3-Clause License
CommitPackFT	commitpackft-vcl[165]	Apache License 2.0, BSD 3-Clause License, MIT License
CommitPackFT	commitpackft-xbase[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-xpages[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-xproc[165]	Apache License 2.0, MIT License
CommitPackFT	commitpackft-yacc[165]	MIT License, ISC License, BSD 2-Clause License
CommitPackFT	commitpackft-zephir[165]	MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
CommitPackFT	commitpackft-zig[165]	MIT License
Dolly 15k	dolly-brainstorming[165]	CC BY-SA 3.0
Dolly 15k	dolly-classification[165]	CC BY-SA 3.0
Dolly 15k	dolly-closedqa[165]	CC BY-SA 3.0
Dolly 15k	dolly-creative_writing[165]	CC BY-SA 3.0
Dolly 15k	dolly-infoextract[165]	CC BY-SA 3.0
Dolly 15k	dolly-openqa[165]	CC BY-SA 3.0
Dolly 15k	dolly-summarization[165]	CC BY-SA 3.0
DialogStudio	ds-ABCD[165]	Apache License 2.0, MIT License
DialogStudio	ds-ATIS[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-ATIS-NER[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-AirDialogue[165]	Apache License 2.0
DialogStudio	ds-AntiScam[165]	Apache License 2.0, CC0 1.0
DialogStudio	ds-BANKING77[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-BANKING77-OOS[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-BiTOD[165]	Apache License 2.0
DialogStudio	ds-CLINC-Single-Domain-OOS-banking[165]	Apache License 2.0, CC BY 3.0
DialogStudio	ds-CLINC-Single-Domain-OOS-credit_cards[165]	Apache License 2.0, CC BY 3.0
DialogStudio	ds-CLINC150[165]	Apache License 2.0, CC BY-SA 3.0
DialogStudio	ds-CaSiNo[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-CoQA[165]	Apache License 2.0, MIT License
DialogStudio	ds-CoSQL[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-ConvAI2[165]	Apache License 2.0
DialogStudio	ds-CraigslistBargains[165]	Apache License 2.0, MIT License
DialogStudio	ds-DART[165]	Apache License 2.0, MIT License
DialogStudio	ds-DSTC8-SGD[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-DialogSum[165]	Apache License 2.0, MIT License
DialogStudio	ds-Disambiguation[165]	Apache License 2.0, MIT License
DialogStudio	ds-FeTaQA[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-GECOR[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-GrailQA[165]	Apache License 2.0
DialogStudio	ds-HDSA-Dialog[165]	Apache License 2.0, MIT License
DialogStudio	ds-HH-RLHF[165]	Apache License 2.0, MIT License
DialogStudio	ds-HWU64[165]	Apache License 2.0, CC BY-SA 3.0
DialogStudio	ds-HybridQA[165]	Apache License 2.0, MIT License
DialogStudio	ds-KETOD[165]	Apache License 2.0, MIT License
DialogStudio	ds-MTOP[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-MULTIWOZ2_2[165]	Apache License 2.0, MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
DialogStudio	ds-MulDoGO[165]	Apache License 2.0, CDLA Permissive 1.0
DialogStudio	ds-MultiWOZ_2.1[165]	Apache License 2.0, MIT License
DialogStudio	ds-Prosocial[165]	Apache License 2.0, MIT License
DialogStudio	ds-RESTAURANTS8K[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-SGD[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-SNIPS[165]	Apache License 2.0
DialogStudio	ds-SNIPS-NER[165]	Apache License 2.0
DialogStudio	ds-SParC[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-SQA[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-STAR[165]	Apache License 2.0, MIT License
DialogStudio	ds-Spider[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-TOP[165]	Apache License 2.0, CC BY-SA
DialogStudio	ds-TOP-NER[165]	Apache License 2.0, CC BY-SA
DialogStudio	ds-Taskmaster1[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-Taskmaster2[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-Taskmaster3[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-ToTTo[165]	Apache License 2.0, CC BY-SA 3.0
DialogStudio	ds-TweetSumm[165]	Apache License 2.0, CC0 1.0
DialogStudio	ds-WOZ2_0[165]	Apache License 2.0
DialogStudio	ds-WebQSP[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-WikiSQL[165]	Apache License 2.0, BSD 3-Clause License
DialogStudio	ds-WikiTQ[165]	Apache License 2.0, CC BY-SA 4.0
DialogStudio	ds-chitchat-dataset[165]	Apache License 2.0, MIT License
DialogStudio	ds-wizard_of_internet[165]	Apache License 2.0, CC BY 4.0
DialogStudio	ds-wizard_of_wikipedia[165]	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[58]	CC BY-SA 3.0
Flan Collection (Chain-of-Thought)	fc-cot-stream_creak[122]	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[45]	CC BY 4.0
Flan Collection (Flan 2021)	fc-flan-e2e_nlg[120]	CC BY-SA 4.0
Flan Collection (Flan 2021)	fc-flan-natural_questions[87]	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[147]	CC BY-SA 4.0
Flan Collection (Flan 2021)	fc-flan-squad_v2[147]	CC BY-SA 4.0
Flan Collection (Flan 2021)	fc-flan-trec[98]	CC0 1.0
Flan Collection (Flan 2021)	fc-flan-true_case[98]	CC0 1.0
Flan Collection (Flan 2021)	fc-flan-wiki_lingua_english_en[88]	CC BY 3.0
Flan Collection (Flan 2021)	fc-flan-winogrande[158]	Apache License 2.0, CC BY 4.0

Continued on next page

Collection	Dataset Identifier	Licenses
Flan Collection (Flan 2021)	fc-flan-wnli[96]	CC BY 4.0
Flan Collection (Flan 2021)	fc-flan-word_segment[96]	CC0 1.0
Flan Collection (Flan 2021)	fc-flan-wsc[96]	CC BY 4.0
Flan Collection (P3)	fc-p3-adversarial_qa[12]	CC BY-SA 3.0
Flan Collection (P3)	fc-p3-cos_e[144]	BSD 3-Clause License
Flan Collection (P3)	fc-p3-dbpedia_14[95]	CC BY-SA 3.0
Flan Collection (P3)	fc-p3-hotpotqa[188]	Apache License 2.0, CC BY-SA 4.0
Flan Collection (P3)	fc-p3-quarel[153]	CC BY 4.0
Flan Collection (P3)	fc-p3-quartz[169]	CC BY 4.0
Flan Collection (P3)	fc-p3-quoref[43]	CC BY 4.0
Flan Collection (P3)	fc-p3-web_questions[14]	CC BY 4.0
Flan Collection (P3)	fc-p3-wiki_bio[91]	CC BY-SA 3.0
Flan Collection (P3)	fc-p3-wiki_hop[91]	CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-adversarial_qa[12]	CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-adverserial_qa[12]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-air_dialogue[185]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-ancora_ca_ner[185]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-anem[121]	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[151]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-atomic[73]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-bard[53]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-cedr[161]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-circa[110]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-clue_cmrc2018[41]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-coached_conv_pref[139]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-copa_hr	BSD 2-Clause License
Flan Collection (Super-NaturalInstructions)	fc-sni-crows_pairs[119]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-cuad[69]	CC BY 4.0

Continued on next page

Collection	Dataset Identifier	Licenses
Flan Collection (Super-NaturalInstructions)	fc-sni-defeasible_nli_atomic[155]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-disfl_qa[65]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-e_snli[21]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-gap[184]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-hotpotqa[188]	Apache License 2.0, CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-human_ratings_of_natural_-language_generation_outputs[188]	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[52]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-jigsaw[52]	CC0 1.0
Flan Collection (Super-NaturalInstructions)	fc-sni-librispeech_asr[126]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-logic2text[27]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-numeric_fused_head[48]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-offenseval_dravidian[48]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-open_pi[171]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-paper_reviews_data_set[171]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-poem_sentiment[163]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-propara[42]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-quarel[153]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-quartz[169]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-quoref[43]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-ro_sts_parallel[47]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-schema_guided_dstc8[148]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-scitail[84]	Apache License 2.0
Flan Collection (Super-NaturalInstructions)	fc-sni-scitailv1.1[84]	Apache License 2.0

Continued on next page

Collection	Dataset Identifier	Licenses
Flan Collection (Super-NaturalInstructions)	fc-sni-semeval_2020_task4[84]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-sms_spam_collection_v.1[84]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-splash[84]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-squad2.0[147]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-squad_1.1[146]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-strategyqa[58]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-universal_dependencies___-english_dependency_treebank[58]	CC BY-SA 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-web_questions[14]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-wiki_hop[91]	CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-wikitext[114]	CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-winograd_wsc[96]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-winomt[168]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-winowhy[192]	MIT License
Flan Collection (Super-NaturalInstructions)	fc-sni-wsc; enhanced_wsc[192]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-wsc_fiexed[192]	CC BY-SA 3.0
Flan Collection (Super-NaturalInstructions)	fc-sni-xcopa[134]	CC BY 4.0
Flan Collection (Super-NaturalInstructions)	fc-sni-xquad[7]	CC BY-SA 4.0
Open Assistant	oasst-en[86]	Apache License 2.0, CC BY 4.0
Open Assistant OctoPack	oasst-en-octopack[86]	Apache License 2.0, CC BY 4.0
Open Assistant v2	oasst2-en[86]	Apache License 2.0
OIG	oig-unified_canadian_-parliament[86]	Apache License 2.0
OIG	oig-unified_cuad[86]	Apache License 2.0, CC BY 4.0
OIG	oig-unified_grade_school_math_-instructions[86]	Apache License 2.0, MIT License
OIG	oig-unified_nq[86]	Apache License 2.0, CC BY-SA 3.0
OIG	oig-unified_sqlv1[86]	Apache License 2.0, CC BY-SA 4.0
OIG	oig-unified_sqlv2[86]	Apache License 2.0, CC BY-SA 4.0
OIG	oig-unified_squad_v2_more_-neg[86]	Apache License 2.0, CC BY-SA 4.0

Continued on next page

Collection	Dataset Identifier	Licenses
Tasksource Instruct	tsi-balanced_copa[83]	BSD 2-Clause License
Tasksource Instruct	tsi-breaking_nli[59]	CC BY-SA 4.0
Tasksource Instruct	tsi-cladder[59]	MIT License
Tasksource Instruct	tsi-condaqa[149]	Apache License 2.0
Tasksource Instruct	tsi-conj_nli[157]	MIT License
Tasksource Instruct	tsi-defeasible_nli-atomic[155]	MIT License
Tasksource Instruct	tsi-defeasible_nli-snli[155]	MIT License
Tasksource Instruct	tsi-dynasent-dynabench.dynasent.r1.all-r1[135]	CC BY 4.0
Tasksource Instruct	tsi-dynasent-dynabench.dynasent.r2.all-r2[135]	CC BY 4.0
Tasksource Instruct	tsi-fever_evidence_related-mwong_-_fever_related[176]	CC BY-SA 4.0
Tasksource Instruct	tsi-few_nerd-supervised[44]	CC BY-SA 4.0
Tasksource Instruct	tsi-fig_qa[100]	MIT License
Tasksource Instruct	tsi-fracas[55]	MIT License
Tasksource Instruct	tsi-hyperpartisan_news[55]	CC BY 4.0
Tasksource Instruct	tsi-lex_glue-case_hold[23]	Apache License 2.0
Tasksource Instruct	tsi-lonli[173]	MIT License
Tasksource Instruct	tsi-moral_stories-full[50]	MIT License
Tasksource Instruct	tsi-neqa[50]	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[194]	Apache License 2.0
Tasksource Instruct	tsi-truthful_qa-multiple_choice[99]	Apache License 2.0
Tasksource Instruct	tsi-vitaminc-tals__vitaminc[162]	MIT License
Tasksource Instruct	tsi-winowhy[192]	MIT License
Tasksource Symbol-Tuning	tsy-breaking_nli[59]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-cladder[59]	MIT License
Tasksource Symbol-Tuning	tsy-condaqa[149]	Apache License 2.0
Tasksource Symbol-Tuning	tsy-conj_nli[157]	MIT License
Tasksource Symbol-Tuning	tsy-defeasible_nli-atomic[155]	MIT License
Tasksource Symbol-Tuning	tsy-defeasible_nli-snli[155]	MIT License

Continued on next page

Collection	Dataset Identifier	Licenses
Tasksource Symbol-Tuning	tsy-dynasent-dynabench.dynasent.r1.all-r1[135]	CC BY 4.0
Tasksource Symbol-Tuning	tsy-dynasent-dynabench.dynasent.r2.all-r2[135]	CC BY 4.0
Tasksource Symbol-Tuning	tsy-fever_evidence_related-mwong__fever_related[176]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-fracas[55]	MIT License
Tasksource Symbol-Tuning	tsy-hyperpartisan_news[55]	CC BY 4.0
Tasksource Symbol-Tuning	tsy-lonli[173]	MIT License
Tasksource Symbol-Tuning	tsy-recast-recast_factuality[173]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_-megaveridicality[173]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_ner[173]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_puns[173]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_sentiment[173]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_verbcorner[173]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-recast-recast_verbnet[173]	CC BY-SA 4.0
Tasksource Symbol-Tuning	tsy-tracie[194]	Apache License 2.0
Tasksource Symbol-Tuning	tsy-vitaminc-tals__vitaminc[162]	MIT License
Tasksource Symbol-Tuning	tsy-winowhy[192]	MIT License

D Additional Filtering Details

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 1. Additionally, statistics on the pre- and post-filtered sizes of each source can be found in Table 2.

Table 1: 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 2: 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
Database 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**

E Details on Comma Training Data Mixture

Inspired by the MixMin algorithm [177], 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 [131]. In doing so, we found that the amount of text in each source was poorly correlated text quality, motivating the use of heuristic mixing weights to up-/down-weight different sources in our pre-training mix. In Table 3 we list the pre-training mixture weights for each of the sources in the Common Pile.

Table 3: 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%**

F 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.

G 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.

H 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 4.

Table 4: 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

Continued on next page

Snapshot	Unicode Words	Documents
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
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

Continued on next page

Snapshot	Unicode Words	Documents
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**

I PeS2o Source Statistics

Additional statistics on the composition of the peS2o subset of the Common Pile can be found in Tables 5 to 7.

Table 5: Distribution of words, unicode characters, and tokens over train and validation splits in the peS2o subset.

	Train Split	Validation Split
Whitespace words	35.4 B	0.25 B
UTF-8 characters	188 B	1.3 B
Documents	6.25 M	39.1 K

Table 6: 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 7: Distribution of papers across 23 fields of study, as identified by the Semantic Scholar API [85]. 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

Continued on next page

Field of Study	Train Split	Validation Split
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

J 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. [131] to compare their data quality against the Common Pile. In Figure 5 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 8, 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.

Table 8: **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

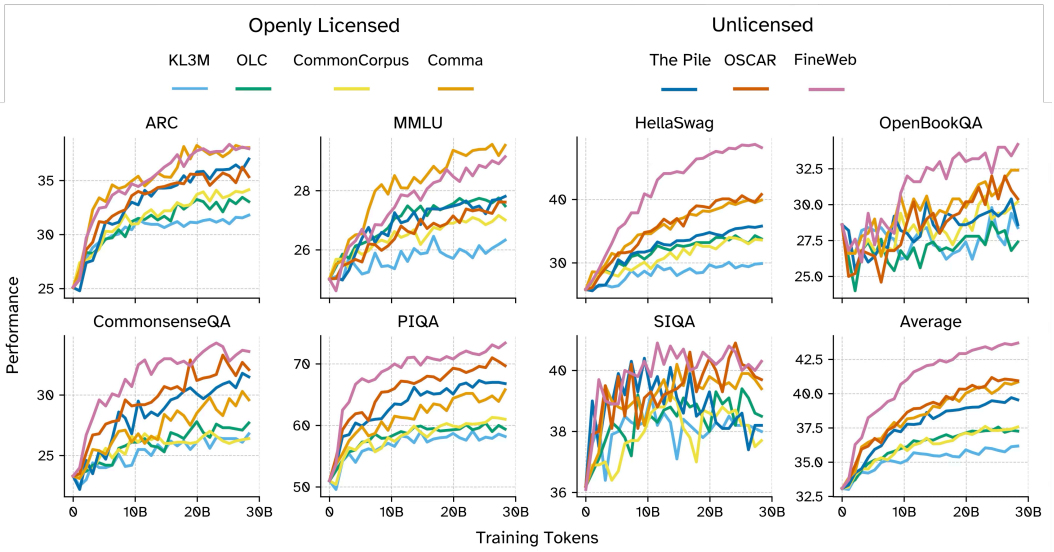

Figure 5: **Comma consistently outperforms models trained on other corpora of openly licensed text and outperforms the Pile on all but two tasks.** We train 1.7B parameter models on 28B tokens following Penedo et al. [131].

K Details on Comma's Cool-Down Data Mixture

Following Hu et al. [71], we end training with a "cool-down" where we train on 37.5B 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 9.

Table 9: 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%

Continued on next page

Source	Size (GB)	Repeats	Effective Size (GB)	Tokens (Billions)	Percentage
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%
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%**

L Additional Comma Results

We provide exact numerical results for Comma and baseline models across a variety of knowledge and reasoning tasks (Table 10) and coding benchmarks (Table 11). 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 10: Comparison between Comma and baseline models trained with similar resources (7 billion parameters, 1 trillion tokens) across a variety of knowledge and reasoning benchmarks.

Model	ARC-C	ARC-E	MMLU	BoolQ	HS	OBQA	CSQA	PIQA	SIQA	Avg.
RPJ-INCITE	42.8	68.4	27.8	68.6	70.3	49.4	57.7	76.0	46.9	56.4
Comma	52.8	68.4	42.4	75.7	62.6	47.0	59.4	70.8	50.8	58.9
MPT	46.5	70.5	30.2	74.2	77.6	48.6	63.3	77.3	49.1	59.7
OpenLLaMA	44.5	67.2	40.3	72.6	72.6	50.8	62.8	78.0	49.7	59.8
LLaMA	44.5	67.9	34.8	75.4	76.2	51.2	61.8	77.2	50.3	59.9
StableLM	50.8	65.4	45.2	71.7	75.6	48.2	57.2	77.0	48.2	59.9
Qwen3	57.2	74.5	77.0	86.1	77.0	50.8	66.4	78.2	55.0	69.1

Table 11: Comparison between Comma and baseline models trained with similar resources (7 billion parameters, 1 trillion tokens) on two coding benchmarks.

Model	HumanEval	MBPP	Avg.
RPJ-INCITE	11.1	15.9	13.5
LLaMA	19.9	27.9	23.9
StableLM	23.1	32.0	27.6
MPT	27.3	33.2	30.3
Open LLaMA	27.6	33.9	30.8
Comma	36.5	35.5	36.0
Qwen3	94.5	67.5	81

References

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

[2] 17 U.S. Code § 102. Subject matter of copyright: In general, December 1990. URL https://www.law.cornell.edu/uscode/text/17/102.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. pages 4623–4637, 2019.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Jonathan Berant, A. Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. pages 1533–1544, 2013.

[15] 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.

[16] 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.

[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. 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. 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 W. Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. pages 1026–1036, 2020.

[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. 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] 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. pages 5882–5888, 2019.

[42] 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. pages 1595–1604, 2018.

[43] Pradeep Dasigi, Nelson F. Liu, Ana Marasović, Noah A. Smith, and Matt Gardner. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. *ArXiv*, abs/1908.05803, 2019.

[44] 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.

[45] Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. pages 2368–2378, 2019.

[46] 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.

[47] 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. 2021.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] James Ferguson, Matt Gardner, Tushar Khot, and Pradeep Dasigi. Iirc: A dataset of incomplete information reading comprehension questions. pages 1137–1147, 2020.

[53] Nancy Fulda, Nathan Tibbetts, Zachary Brown, and D. Wingate. Harvesting common-sense navigational knowledge for robotics from uncurated text corpora. pages 525–534, 2017.

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

[55] 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.

[56] 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.

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

[58] 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.

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

[60] 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.

[61] 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].

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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/.

[67] 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/.

[68] 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.

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

[70] 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.

[71] 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.

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

[73] 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. pages 6384–6392, 2020.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

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

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

[82] 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.

[83] 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.

[84] Tushar Khot, Ashish Sabharwal, and Peter Clark. Scitail: A textual entailment dataset from science question answering. pages 5189–5197, 2018.

[85] 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.

[86] 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.

[87] 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.

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

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

[90] 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.

[91] R. Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. pages 1203–1213, 2016.

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] H. Levesque, E. Davis, and L. Morgenstern. The winograd schema challenge. 2011.

[97] 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.

[98] Xin Li and D. Roth. Learning question classifiers. pages 1–7, 2002.

[99] Stephanie C. Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. pages 3214–3252, 2021.

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

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

[107] 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.

[108] 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.

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

[110] Annie Louis, D. Roth, and Filip Radlinski. "i'd rather just go to bed": Understanding indirect answers. *ArXiv*, abs/2010.03450, 2020.

[111] 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.

[112] 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.

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

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

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. pages 1953–1967, 2020.

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

[121] 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.

[122] 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.

[123] OpenAI. House of lords communications and digital select committee inquiry: Large language models. LLM0113, 2023.

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

[125] Pedro Javier Ortiz Su'arez, Benoit Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim, 2019. Leibniz-Institut f"ur Deutsche Sprache. doi: 10.14618/ids-pub-9021. URL http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

[126] 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.

[127] 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.

[128] 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.

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

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

[131] 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.

[132] Stephen R. Peterson. Expert report of stephen r. peterson, ph.d., 2024. URL https://storage.courtlistener.com/recap/gov.uscourts.tnmd.96652/gov.uscourts.tnmd.96652.67.19_1.pdf. Filed in *Concord Music Group v. Anthropic PBC*, No. 3:23-cv-01092 (M.D. Tenn.).

[133] 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.

[134] E. Ponti, Goran Glavavs, Olga Majewska, Qianchu Liu, Ivan Vulic, and A. Korhonen. Xcopa: A multilingual dataset for causal commonsense reasoning. pages 2362–2376, 2020.

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

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

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

[138] 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.

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

[140] 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.

[141] 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.

[142] 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.

[143] 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.

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

[145] 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.

[146] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. pages 2383–2392, 2016.

[147] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *ArXiv*, abs/1806.03822, 2018.

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

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

[150] 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.

[151] 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.

[152] 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.

[153] Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. Getting closer to ai complete question answering: A set of prerequisite real tasks. pages 8722–8731, 2020.

[154] 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.

[155] 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. pages 4661–4675, 2020.

[156] 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.

[157] Swarnadeep Saha, Yixin Nie, and Mohit Bansal. Conjnli: Natural language inference over conjunctive sentences. pages 8240–8252, 2020.

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

[159] 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.

[160] 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.

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

[162] Tal Schuster, Adam Fisch, and R. Barzilay. Get your vitamin c! robust fact verification with contrastive evidence. pages 624–643, 2021.

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

[164] 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.

[165] 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.

[166] 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.

[167] 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.

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

[169] Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. Quartz: An open-domain dataset of qualitative relationship questions. *ArXiv*, abs/1909.03553, 2019.

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

[171] 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.

[172] 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.

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

[174] 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.

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

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

[177] 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.

[178] 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.

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

[180] 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.

[181] 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.

[182] 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.

[183] 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.

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

[185] Wei Wei, Quoc V. Le, Andrew M. Dai, and Jia Li. Airdialogue: An environment for goal-oriented dialogue research. pages 3844–3854, 2018.

[186] 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.

[187] 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.

[188] 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. pages 2369–2380, 2018.

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

[190] 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.

[191] 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.

[192] Hongming Zhang, Xinran Zhao, and Yangqiu Song. Winowhy: A deep diagnosis of essential commonsense knowledge for answering winograd schema challenge. pages 5736–5745, 2020.

[193] 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.

[194] 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.