Datasheet for the Pile

Stella Biderman, Kieran Bicheno, and Leo Gao

EleutherAI
contact@eleuther.ai

January 19, 2022

Abstract

This datasheet describes the Pile, a 825 GiB dataset of human-authored text compiled by EleutherAI for use in large-scale language modeling. The Pile is comprised of 22 different text sources, ranging from original scrapes done for this project, to text data made available by the data owners, to third-party scrapes available online.

1 Background on the Pile

The Pile is a massive text corpus created by EleutherAI for large-scale language modeling efforts. It is comprised of textual data from 22 sources (see below) and can be downloaded from the official website as well as from a community mirror. Each source dataset is at its core a textual work, and any non-textual data (including metadata) has been removed. While still preserving their internal order, the documents from all the sources have been randomly shuffled. For further information on the Pile, see Gao et al. [2020].

This document is not intended to be – and should not be used as – a substitute for a datasheet for the original versions of the component datasets. While it is accurate for the text data that we derived from each component dataset, the original source dataset may have other properties. This document is intended to inform people interested in using the Pile for natural language processing. People interested in using the original datasets should contact the data owners for information about the properties of the original data.

It is not always the case that the answer to the questions below are known with certainty. For example, while we have no reason to believe that personal identifying information (PII) is contained in most of the subsets of our dataset, it is always possible that someone wrote down PII in a document and uploaded it to arXiv. Due to the sheer scale of the data, it is impractical to systematically search through every text to validate that it is what it purports to be. We have endeavored to answer the questions below as best we can, and to be open and honest about the limitations of the accuracy of this document. Anyone who engages in research on or with the Pile is welcome to contact us to have their findings added to this document. Similarly, we welcome all comments, suggestions, or corrections.

Datasets contained in the Pile:

Pile-CC: The Pile-CC dataset is a sample from the Common Crawl WARCs that has been converted to text using jusText [Endrédy and Novák, 2013].

PubMed Central: The PubMed Central dataset is a subset of the PubMed online repository for biomedical articles run by the United States of America's National Center for Biotechnology Information (NCBI).

Books3: The Books3 component is a dataset of books derived from a copy of the contents of the Bibliotik private tracker made available by The Eye [Presser, 2020].

arXiv: The arXiv component is a subset of the ArXiv preprint repository for research papers that has operated since 1991.

Github: The GitHub component is an EleutherAI scrape of GitHub, a large dataset of open source code repositories.

OpenWebText2: The OpenWebText2 component is a web scrape dataset produced by EleutherAI and inspired by WebText [Radford et al., 2019] and OpenWebTextCorpus [Gokaslan and Cohen, 2019].

FreeLaw: The Free Law Project is US registered non-profit that provide access to millions of legal opinions and analytical tools for academic studies in the legal realm.

Wikipedia (en): The Wikipedia (en) dataset is taken from the Wikipedia site as a standard source of high-quality text for language modeling.

StackExchange: The StackExchange dataset is a dump of anonymized user-contributed content on the Stack Exchange network, a popular collection of websites centered around user-contributed questions and answers.

USPTO Backgrounds: The USPTO Backgrounds dataset is a set of background sections from patents granted by the United States Patent and Trademark Office, derived from its published bulk archives[1].

PubMed Abstracts: The PubMed Abstracts dataset comprises the abstracts of 30 million publications in the PubMed online repository for biomedical articles.

Project Gutenberg (PG-19): The Project Gutenberg dataset is a corpus of high-quality, classic literature.

OpenSubtitles: The OpenSubtitles dataset is an English language dataset of subtitles from movies and television shows gathered by Tiedemann [2016].

DM Mathematics: The DeepMind Mathematics dataset consists of a collection of mathematical problems such as algebra, arithmetic, calculus, number theory, and probability, formatted as natural language prompts [Saxton et al., 2019].

BookCorpus2: BookCorpus2 is an expanded version of the original BookCorpus [Zhu et al., 2015], a widely used language modeling corpus consisting of books written by "as of yet unpublished authors."

Ubuntu IRC: The Ubuntu IRC dataset is derived from the publicly available chatlogs[2] of all Ubuntu-related channels on the Freenode IRC chat server.

EuroParl: EuroParl [Koehn, 2005] is a multilingual parallel corpus originally introduced for machine translation but which has also seen use in several other fields of NLP [Groves and Way, 2006a, Van Halteren, 2008, Ciobanu et al., 2017].

YouTube Subtitles: The YouTube Subtitles dataset is a parallel corpus of text gathered from human generated closed-captions on YouTube.

PhilPapers: PhilPapers[3] is a dataset of open access philosophy publications from an international database maintained by the Center for Digital Philosophy at the University of Western Ontario.

NIH ExPORTER: The NIH Grant abstracts provides a bulk-data repository for awarded applications through the ExPORTER[4] service covering the fiscal years 1985-present.

HackerNews: Hacker News[5] is a link aggregator operated by Y Combiner, a startup incubator and investment fund.

Enron Emails: The Enron Emails dataset [Klimt and Yang, 2004] contains text from the contents of Enron's email servers unearthed during the investigation into that organization's accounting methods and is a valuable corpus for understanding the modality of email communications, which are typically not found in any of our other datasets.

2 Motivation For Dataset Creation

Why was the dataset created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

The Pile: The Pile was created for the purposes of training large-scale language models such as Radford et al. [2019], Rosset [2019], and Brown et al. [2020]. Training such models requires massive amounts of

---

[1]https://bulkdata.USPTO.gov/
[2]https://irclogs.ubuntu.com/
[3]https://philpapers.org/
[4]https://exporter.nih.gov/
[5]https://news.ycombinator.com

human-authored text data which, hitherto, was not available from a single source other than Pile-CC derived data. Below we discuss the motivation for creating each constituent collection. For data collections that were created specifically for the Pile, we also discuss the motivations for using the underlying data collection.

Pile-CC: The Pile-CC dataset was created to be included in the Pile. The underlying data comes from the Common Crawl, which was created to give people access to the wealth of information contained in the internet. Its creators were concerned that only data mining companies would be able to collect this data, and has the explicit aim of democratizing technology.

PubMed Central: The PubMed Central dataset was created to be included in the Pile. The underlying data comes from the PubMed Central database, which was created to allow public access to the results of research funded by the U.S. National Institute of Health and in compliance with the Consolidated Appropriations Act of 2008 (H.R. 2764).

Books3: The Books3 dataset was created as a resource for language modelling research. We included Books3 [Presser, 2020], a pre-existing collection of long-form books, because books are invaluable for long-range context modeling research.

arXiv: The arXiv dataset was created to be included in the Pile. We included arXiv in the hopes that it will be a source of high quality text and math knowledge, and benefit potential downstream applications to research in Math, CS, Physics, and Machine Learning.

Github: The GitHub dataset was created to be included in the Pile. The data comes from the Github website[6] which was created for hosting offsite Git[7] software repositories

OpenWebText2: The OpenWebText2 dataset was created to be included in the Pile. We wanted a high-quality web scrape similar to WebText [Radford et al., 2019] and OpenWebTextCorpus [Gokaslan and Cohen, 2019].

FreeLaw: The FreeLaw dataset was created to be included in the Pile. The underlying data comes from the Free Law Project. According to the Free Law Project website, they have several complimentary purposes for collecting and publishing the data:

> to provide free, public, and permanent access to primary legal materials on the Internet for educational, charitable, and scientific purposes to the benefit of the general public and the public interest; to develop, implement, and provide public access to technologies useful for legal research; to create an open ecosystem for legal research and materials; to support academic research on related technologies, corpora, and legal systems; and to carry on other charitable activities associated with these purposes, including, but not limited to, publications, meetings, conferences, trainings, educational seminars, and the issuance of grants and other financial support to educational institutions, foundations, and other organizations exclusively for educational, charitable, and scientific purposes as allowed by law.

Wikipedia (en): The Wikipedia (en) dataset was created as a standard source of high-quality text for language modeling and a source of information in question-answer models.

StackExchange: The StackExchange dataset was created to be included in the Pile. The underlying data comes from a dump of StackExchange questions and answers which was created to provide people with the ability to analyze the data contained in the StackExchange network.

USPTO Backgrounds: The USPTO dataset was created to be included in the Pile. The underlying data comes from the US Patent and Trademark Office's website, and the data sets were created "[t]o advance research on matters relevant to intellectual property, entrepreneurship, and innovation," to "facilitate economic research on patents and trademarks" and to "support White House policy that champions transparency and access to government data under the "data.gov" umbrella of initiatives."

PubMed Abstracts: The PubMed Abstracts dataset was created to be included in the Pile. The underlying data comes from the PubMed publication index which was created as part of fulfilling the U.S. National Institute of Health's "charge of maintaining a record of biomedical research of grants, publications, and other scholarly works" and providing public access to taxpayer funded research.

---

[6]https://github.com/
[7]https://git-scm.com/

Project Gutenberg (PG-19): The PG-19 dataset was created by DeepMind from the Gutenberg Project, and is included as a source of high-quality classic literature and the source of many English idioms useful in downstream NLP projects.

OpenSubtitles: The OpenSubtitles dataset was create to serve as a "prime resource for the compilation of parallel corpora."[Tiedemann, 2016].

DM Mathematics: The DM Mathematics dataset was created by DeepMind to aid investigation of the ability of neural networks to learn to solve arithmetical problems.

BookCorpus2: The BookCorpus2 dataset was created for the Pile to aid in natural language processing research.

Ubuntu IRC: The Ubuntu IRC corpus was created as a "unique resource for research into building dialogue managers based on neural language models that can make use of large amounts of unlabeled data" [Lowe et al., 2016].

EuroParl: The EuroParl dataset was created to aid statistical machine translation research.

YouTube Subtitles: The YouTube Subtitles dataset was created to be included in the Pile.

PhilPapers: The PhilPapers dataset was created to be included in the Pile.

NIH ExPORTER: The NIH ExPORTER database was created to provide public access to biomedical research.

HackerNews: The HackerNews dataset was created to be included in the Pile.

Enron Emails: The Enron Emails dataset was created to aid investigation into the unethical accounting practices of Enron and the circumstances surrounding its collapse. The corpus is not the email data itself but text derived from it for what has become routine research into email and email users' behaviour.

Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

The Pile: The Pile has been used as a trainin dataset for a variety of large language models including GPT-Neo [Black et al., 2021], GPT-J [Wang and Komatsuzaki, 2021], Jurassic-1 [Lieber et al., 2021], Wu Dao [Tang, 2021], and GPT-NeoX [Andonian et al., 2021]. Other models, trained using custom subsets of the Pile, are listed below by subset. A number of papers study the properties of models trained on the Pile, including Mitchell et al. [2021], Peyrard et al. [2021], Matiana et al. [2021], Mukherjee et al. [2021], Magee et al. [2021], and Lee et al. [2021].

Pile-CC: Papers that train models on datasets that include the Pile-CC subset of the Pile include Luo et al. [2021], Kharya and Alvi [2021], Askell et al. [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

PubMed Central: None

Books3: Papers that train models on datasets that include the Books3 subset of the Pile include Luo et al. [2021], Wang et al. [2021], Kharya and Alvi [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

arXiv: Papers that train models on datasets that include the arXiv subset of the Pile include Askell et al. [2021], Kharya and Alvi [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

Github: Papers that train models on datasets that include the GitHub subset of the Pile include Askell et al. [2021], Kharya and Alvi [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

OpenWebText2: Papers that train models on datasets that include the OpenWebText2 subset of the Pile include Luo et al. [2021], Kharya and Alvi [2021]

FreeLaw: Papers that train models on datasets that include the FreeLaw subset of the Pile include Askell et al. [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

Wikipedia (en): Papers that train models on datasets that include the Wikipedia (en) subset of the Pile include Luo et al. [2021], Kharya and Alvi [2021]. While this data has likely been used by other researchers for various purposes, the only ones that we are aware of that is directly comparable are Radford et al. [2019], Brown et al. [2020].

StackExchange: Papers that train models on datasets that include the StackExchange subset of the Pile include Kharya and Alvi [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

USPTO Backgrounds: Papers that train models on datasets that include the USPTO Backgrounds subset of the Pile include Askell et al. [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

PubMed Abstracts: Papers that train models on datasets that include the PubMed Abstracts subset of the Pile include Kharya and Alvi [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

Project Gutenberg (PG-19): Project Gutenberg 1919 has been used extensively, including by [Zaheer et al., 2020, Choromanski et al., 2020, Roller et al., 2020, Huang and Yang, 2020, Irie et al., 2020]. Papers that train models on datasets that include the PG-19 subset of the Pile include Kharya and Alvi [2021].

OpenSubtitles: OpenSubtitles has been used extensively, including by [Wang, 2017, Sjöblom et al., 2018, Zilio et al., 2018, Gordon and Duh, 2020, Krišlauks and Pinnis, 2020]. Papers that train models on datasets that include the OpenSubstitles subset of the Pile include Luo et al. [2021], Askell et al. [2021]

DM Mathematics: DM Mathematics has been used extensively, including by [Cho et al., 2019, Qi and Wu, 2019, Talmor et al., 2020, Dinu et al., 2020, Firestone, 2020].

BookCorpus2: The BookCorpus dataset that BookCorpus2 is based on has been used extensively, including by [Karpathy and Fei-Fei, 2015, Kiros et al., 2015, Reed et al., 2016, Ba et al., 2016, Devlin et al., 2018]. BookCorpus2 was studied by Bandy and Vincent [2021]. Papers that train models on datasets that include the Pile-CC subset of the Pile include Kharya and Alvi [2021], Askell et al. [2021]

Ubuntu IRC: Papers that train models on datasets that include the Pile-CC subset of the Pile include Askell et al. [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

EuroParl: EuroParl has been extensively used, including by [Bouma, 2009, Koehn, 2009, Heafield, 2011, Navigli and Ponzetto, 2012, Cho et al., 2014].

YouTube Subtitles: While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

PhilPapers: While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

NIH ExPORTER: Papers that train models on datasets that include the NIH ExPORTER subset of the Pile include Kharya and Alvi [2021]. While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

HackerNews: While this data has likely been used by other researchers for various purposes, we are unaware of any uses that would be directly comparable.

Enron Emails: The Enron Emails have been used extensively, including by [Diesner et al., 2005, Kossinets et al., 2008, Tang et al., 2008, Crammer et al., 2009, Madjarov et al., 2012].

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

The Pile: Answers differ by the subset. For each component document, see below.

Pile-CC: The Pile-CC dataset could be used for statistical analysis and comparisons of online dialects in NLP, or structural analysis for non-ML research disciplines.

PubMed Central: The PubMed dataset will allow researchers to test the efficacy of a large-scale general model that includes technical medical information in communication, explanation, and data-mining fields. It could also be used to ensure the legal and regulatory instruments that mandate the dataset's creation are being followed.

Books3: The Books3 dataset could be used as a batch of templates for copyright enforcement checks, as well as allowing models based on this corpus to be used by researchers in the fields of literature and creative arts.

arXiv: The arXiv dataset could be mined for potential answers to significant academic pursuits, assuming many submissions remain broadly unread. Including a deep set of technical papers in the corpora and

its downstream models will give researchers a tool with both general and specific understanding of their domain.

Github: Including a large base of code in a structured way gives downstream models the potential for aiding researchers in creating code for their theoretical work without needing to know how to code themselves.

OpenWebText2: The OpenWebText2 dataset's inclusion allows downstream research tools access to insights not formally captured in a traditional setting like arXiv or PubMed.

FreeLaw: The FreeLaw dataset could be used as a rolling track of the linguistic accessibility of judgments and legal arguments. AI digital assistants or researcher tools created with this corpus would also have an in-built lexicon for understanding legal requirements, such as their own Terms of Use.

Wikipedia (en): The Wikipedia dataset gives a powerful question-and-answer capability for researchers in a broad set of domains and a curated source of fact-checking capacity.

StackExchange: The StackExchange dataset, especially when combined with the GitHub corpus, gives researchers the potential ability to create self-diagnosing and self-repairing models. This also gives the Pile a basis for creating technical support chatbots.

USPTO Backgrounds: The USPTO dataset's inclusion gives researchers a basis for downstream problem-solving models thanks to the large body of problem-based situational awareness built into The Pile.

PubMed Abstracts: The PubMed Abstracts dataset gives researchers, especially in the medical field, a strong basis for building models for deep data-mining and pattern detection in collections of data too large to be dealt with by humans.

Project Gutenberg (PG-19): Training on the PG-19 dataset gives downstream models access to a large number of idiomatic language. The dataset also provides a basis for researchers to track linguistic changes in the English language over time.

OpenSubtitles: The OpenSubtitles data set gives researchers access to better-structured common-language understanding in their downstream projects.

DM Mathematics: The DeepMind Mathematics data set would give researchers in the educational space a platform for testing smaller corpora on their ability to handle mathematical Q&A capabilities against those given by The Pile.

BookCorpus2: BookCorpus2 has a secondary role for researchers as a benchmark for the efficacy of models based on more focused corpora, as BookCorpus2 adds depth of training to the otherwise wide capabilities of The Pile.

Ubuntu IRC: The Ubuntu IRC data set could be used for building chatbots or training self-diagnostic models for computers.

EuroParl: The EuroParl texts add a layer of multi-lingual benchmarking for researchers testing results between specific modals and general models such as those based on The Pile.

YouTube Subtitles: The YouTube subtitles dataset also gives those researchers using the Pile a multi-lingual capability among the predominantly English remainder of the corpus.

PhilPapers: The PhilPapers dataset could be mined for an exhaustive map of intellectual domains by researchers involved in ontology-based projects.

NIH ExPORTER: The NIH ExPORTER dataset gives downstream models a framework by which researchers can add an understanding of government and NGO processes to their models.

HackerNews: HackerNews provides The Pile and downstream models a non-technical bridge between domain specific lexicons and everyday speech in a way that researchers can exploit in projects such as data assistants.

Enron Emails: The Enron emails could be used to study the activities of human agents during a large-system collapse and provide researchers using downstream tools to interface more naturally with email-based projects.

Who funded the creation of the dataset?

The Pile: The Pile was created by EleutherAI. This dataset was created by individuals working on their own time without funding. All components that required processing beyond their orignal form were also

processed by EleutherAI members on their own time without any funding.

Pile-CC: The data is sourced from Common Crawl, a non-profit 501(c)(3) organization founded by Gil Elbaz. The data from Common Crawl was processed by EleutherAI into Pile-CC.
PubMed Central: The U.S. National Center for Biotechnology Information (NCBI), a branch of the National Institutes of Health (NIH).
Books3: Books3 was created by Shawn Presser, an independent ML developer.
arXiv: EleutherAI scraped arXiv ourselves.
Github: EleutherAI scraped Github ourselves.
OpenWebText2: EleutherAI created OpenWebText2 ourselves.
FreeLaw: The Free Law Project, a 501(c)(3) non-profit founded by Michael Lissner and Brian Carver.
Wikipedia (en): The Wikimedia Foundation, a 501(c)(3) organization founded by Jimmy Wales, was the source of the data and DeepMind, an Alphabet subsidiary, created the dataset.
StackExchange: EleutherAI scraped the StackExchange Network ourselves.
USPTO Backgrounds: The U.S. Office of the Chief Economist.
PubMed Abstracts: The U.S. National Institute of Health.
Project Gutenberg (PG-19): Project Gutenberg has received funding from a variety of sources, including the University of Illinois, Carnegie Melon University, and University of North Carolina at Chapel Hill. We believe that its costs are currently paid for by UNC Chapel Hill, but are not certain. The form of the dataset we used was created by DeepMind, an Alphabet subsidiary [Rae et al., 2019].
OpenSubtitles: The site is largely, if not entirely, volunteer-created but paid for by members paying for 'VIP Membership'.
DM Mathematics: DeepMind, an artificial intelligence company and research laboratory owned by Alphabet Inc. (the parent company of Google inc.)
BookCorpus2: The underlying data was created by individual "as of yet unpublished authors". EleutherAI processed the BookCorpus2 dataset ourselves.
Ubuntu IRC: Ubuntu itself is funded by Canonical Ltd., a private company founded by Mark Shuttleworth, and the conversations that constitute the dataset were on the Freenode IRC server, funded through donations and volunteers.
EuroParl: We believe that this dataset was funded the School of Informatics of the University of Edinburgh, Scotland, but were unable to confirm this fact. For further information, contact Groves and Way [2006b].
YouTube Subtitles: EleutherAI scraped YouTube for subtitles ourselves.
PhilPapers: EleutherAI scraped PhilPapers ourselves. The underlying data collection was funded by the Centre for Digital Philosophy at the University of Western Ontario.
NIH ExPORTER: EleutherAI scraped PhilPapers ourselves. The underlying data collection was funded by the U.S. National Institute of Health and other agencies of the U.S. Department of Health and Human Services (ACF, AHRQ, CDC, HRSA, FDA), and the VA.
HackerNews: EleutherAI scraped the Hacker News website ourselves. Hacker News is funded by YCombinator.
Enron Emails: The U.S. Federal Energy Regulatory Commission (FERC).

3   Dataset Composition

What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

The Pile: Instances of the dataset are textual documents of a variety of contents. The instances come from the categories described below.

Pile-CC: Instances are webpages.
PubMed Central: Instances are academic medical papers.

Books3: Instances are published books.
arXiv: Instances are preprints of academic papers, primarily in mathematics, computer science, physics, and statistics.
Github: Instances are code files.
OpenWebText2: Instances are webpages.
FreeLaw: Instances are legal documents.
Wikipedia (en): Instances are pages of Wikipedia (en).
StackExchange: Instances are questions posted on StackExchange, along with highly upvoted answers.
USPTO Backgrounds: Instances are abstracts of US Patent applications.
PubMed Abstracts: Instances are abstracts of papers in the PubMed archive, primarily published medical papers.
Project Gutenberg (PG-19): Instances are books published prior to 1919.
OpenSubtitles: Instances are the subtitles for a movie.
DM Mathematics: Instances are groups of related mathematics problems.
BookCorpus2: Instances are books.
Ubuntu IRC: Instances are chatlogs, chunked by week.
EuroParl: Instances are minutes from meetings of the European Parliament. A large proportion of the data contains the same text repeated in different languages. Copies of the same text in different languages are separate documents.
YouTube Subtitles: Instances are subtitles for YouTube videos.
PhilPapers: Instances are preprints of academic papers, primarily in philosophy.
NIH ExPORTER: Instances are medical academic papers.
HackerNews: Instances are conversation threads from the Hacker News Network.
Enron Emails: Instances are emails between employees of Enron from January 6, 1998 until February 4, 2004.

How many instances are there in total (of each type, if appropriate)?

The number of documents reported are after deduplication. Note that the Pile has had its datasets weighted. The sizes reported here are the raw sizes. For the effective sizes in the Pile, see Figure 1 of Gao et al. [2020].
The Pile: $211,043,181$ documents (unweighted), totaling 825.18 GiB.

Pile-CC: $54,953,117$ documents, totaling 227.12 GiB.
PubMed Central: $3,098,931$ documents, totaling 90.27 GiB.
Books3: $196,640$ documents, totaling 100.96 GiB.
arXiv: $1,264,405$ documents, totaling 56.21 GiB.
Github: $19,021,454$ documents, totaling 95.16 GiB.
OpenWebText2: $17,103,059$ documents, totaling 62.77 GiB.
FreeLaw: $3,562,015$ documents, totaling 51.15 GiB.
Wikipedia (en): $6,033,151$, totaling 6.38 GiB
StackExchange: $15,622,475$ documents, totaling 32.20 GiB.
USPTO Backgrounds: $5,883,037$ documents, totaling 22.90 GiB.
PubMed Abstracts: $15,518,009$ documents, totaling 19.26 GiB.
Project Gutenberg (PG-19): $28,602$ documents, totaling 10.88 GiB.
OpenSubtitles: $446,612$ documents, totaling 12.98 GiB.
DM Mathematics: $1,014,997$ documents, totaling 7.75 GiB.
BookCorpus2: $17,868$ documents, totaling 6.30 GiB.
Ubuntu IRC: $10,605$ documents, totaling 5.52 GiB.
EuroParl: $69,814$ documents, totaling 4.59 GiB.
YouTube Subtitles: $173,651$ documents, totaling 3.73 GiB.
PhilPapers: $33,990$ documents, totaling 2.38 GiB.
NIH ExPORTER: $939,668$ documents, totaling 1.89 GiB.
HackerNews: $831,198$ documents, totaling 3.90 GiB.

Enron Emails: 517,401 documents, totaling 0.88 GiB.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

The Pile: Instances are text files, processed for readability and to remove autogenerated text, garbage generated during parsing, and code fragments picked up from websites. Some of this junk undoubtedly made it through our processing and is embedded in the text files.

Is there a label or target associated with each instance? If so, please provide a description.

The Pile: No.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

The Pile: Not as far as we are aware.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

The Pile: Not as far as we are aware.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The Pile: Answers differ by the subset. For each component document, see below. Many of the underlying datasets grow over time, in which case the date of collection is included. A more complete explanation can be found in Gao et al. [2020], Appendix C. For all components, we have no idea how representative it is of the relevant whole reference class.

Pile-CC: A tiny fraction of the entire Common Crawl was included, chosen arbitrarily and heavily filtered as detailed in Gao et al. [2020].
PubMed Central: All articles in PubMed Central (as of June 2020) are contained in the Pile.
Books3: The entire Books3 dataset is contained in the Pile.
arXiv: We downloaded the TeX sources of all papers on arXiv up to the July 2020 dump (the last file included in our data is `arXiv_src_2007_068.tar`) via arXiv's S3 Bulk Source File Access[8], and used `pandoc 1.19.2.4` to convert these source files to Markdown, discarding any papers which had errors during the conversion process.
Github: The underlying data taken from Github is filtered for only small repositories and only files smaller than 100KB, and subsampled to 95 GiB from the original 600 GiB of files matching the previous criteria. For full details, see Gao et al. [2020].
OpenWebText2: To produce the dataset, URLs and their associated metadata were first extracted from all Reddit submissions up to April 2020. URLs were deduplicated, with each unique URL featuring a list of associated submissions metadata, and an aggregate score. URLs with an aggregate score of less then 3 were removed. The links were then scraped and processed with Newspaper scraper. Deduplication was performed at the document level using in memory MinHashLSH through the DataSketch library.

---

[8] https://arxiv.org/help/bulk_data_s3

FreeLaw: We retained the subset consisting of court opinions and excluded dockets, people, retention votes, and citation data as they were broadly administrative. Court opinions consisted of either plaintext or html. For opinions provided in html, all formatting was discarded and the raw text with extracted using `BeautifulSoup`.

Wikipedia (en): The entire `wikipedia/20200301.en` dataset[9] is included in the Pile.

StackExchange: To construct the dataset, we download and parse every Stack Exchange database dump as of July 1, 2020 to plaintext files. We opt to extract the top three answers with at least three upvotes, discarding all other responses. We only include the plain text question and response and do not incorporate any metadata.

USPTO Backgrounds: The United States Patent and Trademark Office (USPTO) has published bulk archives of the full text of all patents granted in the US from 1976 to September 2020. From these archives, we extract the Background sections, along with key grant-specific metadata, such as the inventor, assignee, and classification information.

PubMed Abstracts: About one-third of the articles in the dataset were missing or contained a malformed title or abstract and were excluded. Additionally, PubMed Central contains full-text resources to many recent publications; any publications which already appear in PMC are excluded from this set.

Project Gutenberg (PG-19): The entire Project Gutenberg 1919 dataset is contained in the Pile.

OpenSubtitles: All datapoints tagged as English by Tiedemann [2016]. We discarded any provided metadata.

DM Mathematics: The entire DM Mathematics dataset is contained in the Pile.

BookCorpus2: The original BookCorpus consists of 11,038 books. However, due to issues with availability of the original BookCorpus, as well as the possibility of collecting a larger version, we decided to collect our own version of BookCorpus using a similar methodology as Kobayashi [2018]. Our version of BookCorpus contains 17,868 books instead, due to the expanding nature of the underlying data.

Ubuntu IRC: We processed all logs from July 5, 2004 through September 1, 2020. All system messages, such as joins, disconnects, nick changes, etc. were discarded, but actions (i.e using `/me`) were kept.

EuroParl: We download the data in bulk from [10]. We remove all basic tag information and only retain the name of each document as a title. For example, `<SPEAKER ID=77 LANGUAGE="NL" NAME="Pronk">` becomes `Pronk`, and then extract the body of each document, discarding those that are shorter than 200 characters.

YouTube Subtitles: The entire YouTube Subtitles dataset is contained in the Pile.

PhilPapers: The entire PhilPapers dataset (as of [date]) is contained in the Pile.

NIH ExPORTER: The NIH provides a bulk-data repository for awarded applications through the ExPORTER service covering the fiscal years 1985–present. These data come from the NIH, but also other other Health and Human Services agencies (ACF, AHRQ, CDC, HRSA, FDA), and the VA. Additionally, the NIH provides a legacy data format named CRISP for awarded applications during the fiscal years 1970–2009. We merged both the ExPORTER and CRISP data to form a consolidated dataset of awarded applications. Entries were deduplicated based off their application ID, and excluded if their abstract text was missing or too short. Small grants, especially administrative ones, consisted solely of short boilerplate. For this reason, we further deduplicated on abstract text. All grants types were considered, including new applications (Application Type Code 1) and renewals (Application Type Code 2) as the text differed enough to provide novel input. The text was then minimally parsed to remove administrative boilerplate, (ex. most old awards contain some variation of "description: (provided by applicant)"). In total, there were 939,668 grant application abstracts added.

HackerNews: The entire HackerNews dataset (as of [date]) is contained in the Pile.

Enron Emails: The entire Enron Emails dataset is contained in the Pile.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The Pile: Yes. The Pile comes with recommended train/validation/test splits. The validation and test datasets are 0.1% of the data each, or about 1.4 GiB. They were chosen at random from the entire

---

dataset.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

The Pile: In the Pile, some components are deliberately upsampled (see Gao et al. [2020]). In terms of the constituent datasets, we have sought to deduplicate instances in Pile-CC and OpenWebText2, as detailed in Gao et al. [2020]. All datasets were machine processed, and therefore likely to contain unknown noise.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The Pile: All datasets in the Pile are self-contained.

4 Collection Process

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The Pile: Detailed information about how each dataset was collected and processed can be found in the appendix of Gao et al. [2020]. The actual code used can be found on GitHub.

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

The Pile: Answers differ by the subset. For each component document, see below.

Pile-CC: Data in the Pile-CC dataset were scraped from websites by the Common Craw and then downloaded directly from the Common Craw by EleutherAI.
PubMed Central: Papers in the PubMed Central dataset were uploaded to PubMed Central by the authors and then downloaded directly from the official databases by EleutherAI.
Books3: Books in the Books3 dataset were collected and processed in a variety of ways, not all of which have been publicly disclosed. Contact Shawn Presser or the Eye for further information.
arXiv: Papers in the arXiv dataset were uploaded by the authors, and then scraped directly from arXiv by EleutherAI.
Github: Code in the GitHub dataset were uploaded by the authors, and then scraped directly from GitHub by EleutherAI.
OpenWebText2: Webpages in the OpenWebText2 dataset were scraped from their original webpages by EleutherAI.
FreeLaw: We do not know how the cases in the FreeLaw dataset were collected. Contact the Free Law Project for further details. Cases were downloaded from the Free Law Project's bulk downloader by EleutherAI.

Wikipedia (en): Pages in the Wikipedia (en) dataset were obtained by DeepMind from official Wiki-Media dumps, uploaded to Kaggle, and then downloaded directly by EleutherAI.

StackExchange: Data in the StackExchange dataset were downloaded from official StackExchange dumps by EleutherAI.

USPTO Backgrounds: Data in the USPTO Backgrounds dataset were submitted by the patent applicants to the U.S. Patent and Trademark Office and then obtained from the U.S. Patent and Trademark Office directly.

PubMed Abstracts: Papers in the PubMed Abstracts dataset were uploaded by the authors, and then downloaded directly from PubMed by EleutherAI.

Project Gutenberg (PG-19): Books in the Project Gutenberg 1919 dataset were collected in a variety of ways, some of them not publicly known. Contact Project Gutenberg and DeepMind for more information.

OpenSubtitles: We do not know how the data for OpenSubtitles was collected.

DM Mathematics: Data in the DM Mathematics dataset were algorithmically generated by DeepMind, an artificial intelligence company and research laboratory owned by Alphabet Inc. (the parent company of Google inc.).

BookCorpus2: Books in the BookCorpus2 dataset were downloaded from SmashWords by EleutherAI.

Ubuntu IRC: Chat logs in the Ubuntu IRC dataset were archived by Ubuntu and downloaded directly by EleutherAI.

EuroParl: The proceedings of the European Parliament were transcribed by professional translators and transcribers, and downloaded directly by EleutherAI.

YouTube Subtitles: Subtitles in the YouTube Subtitles dataset were added by the video owners or volunteers, and then scraped directly from YouTube by EleutherAI.

PhilPapers: Papers in the PhilPapers dataset were uploaded by the authors, and then downloaded directly from PhilPapers by EleutherAI. Machine-readable entries and non-english entries were kept, but entries which could not be parsed by pdfbox were ignored.

NIH ExPORTER: Documents in the NIH ExPORTER dataset were uploaded by their original authors, and then downloaded directly from the ExPORTER database by EleutherAI.

HackerNews: Comments in the HackerNews dataset were posted by their authors and scraped from the Y Combinator API directly by EleutherAI.

Enron Emails: Emails in the Enron Emails dataset were collected by U.S. federal agents in the course of their investigations into the Enron fraud. The data was later collected and prepared by the CALO Project.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The Pile: All sampled datasets were sampled deterministically. How specific datasets were sampled is detailed in Gao et al. [2020], with code publicly available on GitHub.

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

The Pile: We do not know how data collectors were compensated for any of the datasets we did not collect ourselves. People involved with collecting new datasets for the Pile were compensated with an invitation to be an author of Gao et al. [2020].

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The Pile: All data in the Pile was collected prior to September 1st, 2020. All data was collected between June 1st 2020 and September 1st 2020, which for the overwhelming majority of data does not overlap with the time that the data was created at.

Pile-CC: The earliest date of contents in Pile-CC is unknown.

PubMed Central: PubMed Central began when the NIH started collating papers in 2000, though it contains literature going back to the 1800s [pmc, 2020].

Books3: Not all details about the creation of Books3 have been publicly disclosed. Contact Shawn Presser or the Eye for further information.

arXiv: We collected all papers from when arXiv launched on August 14, 1991, up to July 1, 2020.

Github: We collected all code from when GitHub launched on April 10, 2008, up to July 1, 2020.

OpenWebText2: The earliest date of contents in OpenWebText2 is unknown.

FreeLaw: The Free Law project has been collecting data since its founding in 2000, but some contents may be older.

Wikipedia (en): Wikipedia Stack Exchange launched January 15, 2001, but some contents may be older.

StackExchange: Stack Exchange launched in 2010, but some contents may be older.

USPTO Backgrounds: USPTO Backgrounds probably contains documents going back to the founding of the US Patent and Trademark Office on January 2, 1975.

PubMed Abstracts: This dataset took only a matter of hours to collate.

Project Gutenberg (PG-19): No timeframe for the collection of the PG-19 dataset is given by the authors in their paper [Rae et al., 2019], but the use of Project Gutenberg content would represent a recent crawl of old (pre-1919) books [Rae et al., 2019].

OpenSubtitles: We have no idea how old the contents of OpenSubtitles are.

DM Mathematics: DM Mathematics is a dataset of formal mathematics and therefore doesn't become "out of date."

BookCorpus2: We have no idea how old the contents of BookCorpus2 are.

Ubuntu IRC: No timeframe for the collection of or creation of this corpus was put forward in the relevant research paper [Lowe et al., 2016]. Previous datasets created using the Ubuntu IRC put 2004-07-05 as the start date for the IRC channels themselves [Uthus and Aha, 2013].

EuroParl: The EuroParl corpus contains transcripts from 1996 to 2012.

YouTube Subtitles: We obtain subtitles for videos ranging from when YouTube launched on February 14, 2005 until July 1, 2020.

PhilPapers: We collected the publication subset of PhilPapers which spans the years from 2009–2020.

NIH ExPORTER: The NIH provides a bulk-data repository for awarded applications through the ExPORTER service covering the fiscal years 1985–present. These data come from the NIH, but also other other Health and Human Services agencies (ACF, AHRQ, CDC, HRSA, FDA), and the VA. Additionally, the NIH provides a legacy data format named CRISP for awarded applications during the fiscal years 1970–2009. We merged both the ExPORTER and CRISP data to form a consolidated dataset of awarded applications.

HackerNews: We collect the first 24531712 posts on HackerNews. This corresponds to a date range of approximately 10/09/2006 to 09/20/2020.

Enron Emails: The Enron Emails were originally collected by the U.S. Department of Justice during their investigation beginning January 9, 2002 [agencies, 2006] and took two weeks to collate in May 2002 by Joe Bartling [Bartling, 2015].

5 Data Preprocessing

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

The Pile: The data was extensively preprocessed as documented in Gao et al. [2020].

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

The Pile: Yes. Access to the raw data can be obtained from the GitHub repo.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

The Pile: Yes. The primary goal of the data processing is to create an extremely high quality dataset for language modeling. The Pile's success is evidenced both by its widespread adoption for training language models [Lieber et al., 2021, Tang, 2021, Askell et al., 2021] and by studies of the dataset and the models trained on it such as Peyrard et al. [2021], Mukherjee et al. [2021], Mitchell et al. [2021].

6 Dataset Distribution

How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The Pile: The data is distributed through several sources. It can be downloaded directly from the EleutherAI GitHub and from the Eye. It is archived by the Eye as well as on personal storage.

When will the dataset be released/first distributed? What license (if any) is it distributed under?

The Pile: The dataset was released on January 1st, 2021. It is licensed under the MIT License.

Are there any copyrights on the data?

The Pile: Some of the documents in the datasets that the Pile is based on are copyrighted. In particular, Books3 is almost entirely comprised of copyrighted works, and a substantial portion of arXiv and PhilPapers are as well. Other datasets, such as PubMed Central and GitHub contain documents that may be under limited licensing, but are not copyrighted as far as we are aware.

All data contained in the Pile has been heavily processed to aid in language modeling research and no copyrighted text is contained in the Pile in its original form. As far as we are aware, under U.S. copyright law use of copyrighted texts in the Pile falls under the "fair use" doctrine, which allows for the unlicensed use of copyright-protected works in some circumstances.

Copyright law varies by country, and there may be additional restrictions on some of these works in your country. If you are in doubt, it is always advisable to speak to an intellectual property attorney. If you wish to exclude some components of the Pile for legal (or any other) reason, you can compile a custom remix of the datasets using the code on the EleutherAI GitHub to do so.

Are there any fees or access/export restrictions?

The Pile: There are no fees, access restrictions, or import restrictions associated with the Pile.

7 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The Pile: The Pile is supported and maintained by EleutherAI. The data is hosted by the Eye, and has several community maintained mirrors.

Will the dataset be updated? If so, how often and by whom?

> **The Pile:** EleutherAI does not plan to update the Pile. We may release a "Pile Version 2" which will contain texts from a variety of languages as well as updated scrapes of the data sources that increase over time. However, in the event that such a dataset is created, it will be a separate dataset.

How will updates be communicated? (e.g., mailing list, GitHub)

> **The Pile:** Not applicable.

If the dataset becomes obsolete how will this be communicated?

> **The Pile:** If the dataset becomes obsolete, this will be communicated via our GitHub and website, as well as through various social media platforms (Twitter, Reddit, Discord).

Is there a repository to link to any/all papers/systems that use this dataset?

> **The Pile:** We do not maintain one, although academic databases such as Google Scholar and Semantic Scholar contain this information.

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

> **The Pile:** We greatly encourage community participation and feedback on the Pile. We have made all of the code necessary for constructing the Pile from scratch public to enable easier augmentation and improvement of the Pile. However we do not accept submissions of new contributions to the dataset.

8 Legal and Ethical Considerations

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

> **The Pile:** No formal ethical review process was done because EleutherAI does not have an associated IRB. Ethical considerations were discussed throughout the data collection process and is documented in our paper [Gao et al., 2020].

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

> **The Pile:** We are not aware of any confidential data in the Pile, though it is always a possibility. We do know that we are not distributing any *previously inaccessible* confidential data, as all data contained in the Pile was already widely and publicly available on the internet.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why

> **The Pile:** The answer is "probably" for all components other than GitHub and DM Mathematics, for which the answer is "probably not."

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The Pile: Answers differ by the subset. For each component document, see below.

Pile-CC: We do not know the extent to which Pile-CC identifies any subpopulations, although we expect that it does.

PubMed Central: Many medical papers identify subpopulations in the course of their studies. We have confirmed the presence of papers that study race, gender, ability, nation of origin, religion, and sexual orientation.

Books3: Many books identify subpopulations in various ways. We have confirmed the presence of papers that study race, gender, ability, nation of origin, religion, and sexual orientation.

arXiv: We do not know the extent to which arXiv identifies any subpopulations. As the overwhelming majority of papers on arXiv do not deal with human data it is most likely statistically rare. However, it is possible.

Github: We have no reason to believe that any GitHub code identifies any subpopulations.

OpenWebText2: We do not know the extent to which OpenWebText2 identifies any subpopulations, although we expect that it does.

FreeLaw: Many legal documents identify subpopulations in various ways. We have confirmed the presence of papers that study race, gender, ability, nation of origin, religion, and sexual orientation.

Wikipedia (en): Wikipedia (en) identifies many subpopulations, and the way that this manifests and its relation to sociopolitical dynamics has been widely studied, including along the lines of race [Adams et al., 2019, Xing and Vetter, 2020], gender [Reagle and Rhue, 2011, Wagner et al., 2015, Hargittai and Shaw, 2015, Graells-Garrido et al., 2015], ability [Phillips, 2016, Derby, 2012], nation of origin [Rask, 2008, Lee and Chun, 2017], religion [Callahan and Herring, 2011, Ball, 2021], and sexual orientation [Eisner, 2013].

StackExchange: We do not know the extent to which StackExchange identifies any subpopulations. As the overwhelming majority of papers on StackExchange do not deal with human data it is most likely statistically rare. However, it is possible.

USPTO Backgrounds: We do not know the extent to which USPTO Backgrounds identifies any subpopulations. As the overwhelming majority of abstracts submitted to the US Patent and Trademark Office do not deal with human data it is most likely statistically rare. However it is possible.

PubMed Abstracts: Many medical papers identify subpopulations in the course of their studies. We have confirmed the presence of papers that study race, gender, ability, nation of origin, religion, and sexual orientation.

Project Gutenberg (PG-19): Many books identify subpopulations various ways. We have confirmed the presence of books that discuss race, gender, ability, nation of origin, religion, and sexual orientation.

OpenSubtitles: Many movies identify subpopulations various ways. We have confirmed the presence of papers that discuss race, gender, ability, nation of origin, religion, and sexual orientation.

DM Mathematics: The DM Mathematics dataset does not contain any data about people

BookCorpus2: Many books identify subpopulations various ways. We have confirmed the presence of papers that study race, gender, ability, nation of origin, religion, and sexual orientation.

Ubuntu IRC: We do not know the extent to which Ubuntu IRC identifies any subpopulations, although we expect that it does.

EuroParl: We do not know the extent to which EuroParl identifies any subpopulations, although we expect that it does.

YouTube Subtitles: We do not know the extent to which YouTube Subtitles identifies any subpopulations, although we expect that it does.

PhilPapers: We do not know the extent to which PhilPapers identifies any subpopulations, although we expect that it does.

NIH ExPORTER: Many medical papers identify subpopulations in the course of their studies. We have confirmed the presence of papers that study race, gender, ability, nation of origin, religion, and sexual orientation.

HackerNews: We do not know the extent to which HackerNews identifies any subpopulations, although we expect that it does.
Enron Emails: We do not know the extent to which Enron Emails identifies any subpopulations, although we expect that it does.

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

The Pile: Not applicable.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

The Pile: We do not know the extent to which this is the case, although we expect that it does.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The Pile: All data was collected from third parties, typically the original data hosts. For full details on the provenance of each dataset, see Gao et al. [2020].

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

The Pile: No.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

The Pile: The extent to which consent was obtained varies by dataset. For details about the provenance of each dataset, see Gao et al. [2020].

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

The Pile: No.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

The Pile: No.

References

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. *arXiv preprint*, 2020.

István Endrédy and Attila Novák. More effective boilerplate removal – the GoldMiner algorithm. In *Polibits*, 2013.

Shawn Presser. Books3. https://twitter.com/theshawwn/status/1320282149329784833, 2020.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

J. Tiedemann. Finding alternative translations in a large corpus of movie subtitles. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

Declan Groves and Andy Way. Hybridity in mt: Experiments on the Europarl corpus. In *Proceeedings of the 11th Annual conference of the European Association for Machine Translation (EAMT 2006)*, 2006a.

Hans Van Halteren. Source language markers in Europarl translations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 937–944, 2008.

Alina Maria Ciobanu, Liviu P Dinu, and Andrea Sgarro. Towards a map of the syntactic similarity of languages. In *International Conference on Computational Linguistics and Intelligent Text Processing*, pages 576–590. Springer, 2017.

Bryan Klimt and Yiming Yang. The Enron corpus: A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer, 2004.

C Rosset. Turing-NLG: A 17-billion-parameter language model by Microsoft. *Microsoft Blog*, 2019.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems, 2016.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. Gpt-neo: Large scale autoregressive language modeling with mesh-tensorflow. *If you use this software, please cite it using these metadata*, 2021.

Ben Wang and Aran Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model, 2021.

Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 2021.

Jie Tang. WuDao: pretrain the world. Keynote adress at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2021.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Phil Wang, and Samuel Weinbach. GPT-NeoX: Large scale autoregressive language modeling in pytorch, 2021. URL http://github.com/eleutherai/gpt-neox.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*, 2021.

Maxime Peyrard, Sarvjeet Singh Ghotra, Martin Josifoski, Vidhan Agarwal, Barun Patra, Dean Carignan, Emre Kiciman, and Robert West. Invariant language modeling. *arXiv preprint arXiv:2110.08413*, 2021.

Shahbuland Matiana, JR Smith, Ryan Teehan, Louis Castricato, Stella Biderman, Leo Gao, and Spencer Frazier. Cut the carp: Fishing for zero-shot story evaluation. *arXiv preprint arXiv:2110.03111*, 2021.

Rohan Mukherjee, Yeming Wen, Dipak Chaudhari, Thomas Reps, Swarat Chaudhuri, and Chris Jermaine. Neural program generation modulo static analysis. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Liam Magee, Lida Ghahremanlou, Karen Soldatic, and Shanthi Robertson. Intersectional bias in causal language models. *arXiv preprint arXiv:2107.07691*, 2021.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.

Ziyang Luo, Yadong Xi, Jing Ma, Xiaoxi Mao, and Changjie Fan. Analyzing the implicit position encoding ability of transformer decoder. 2021.

Paresh Kharya and Ali Alvi. Using DeepSpeed and Megatron to train Megatron-Turing NLG 530B, the world's largest and most powerful generative language model, Oct 2021.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

Shufan Wang, Laure Thompson, and Mohit Iyyer. Phrase-bert: Improved phrase embeddings from bert with an application to corpus exploration. *arXiv preprint arXiv:2109.06304*, 2021.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33, 2020.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2020.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot, 2020.

Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413671. URL https://doi.org/10.1145/3394171.3413671.

K. Irie, A. Gerstenberger, R. Schlüter, and H. Ney. How much self-attention do we need? trading attention for feed-forward layers. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6154–6158, 2020. doi: 10.1109/ICASSP40776.2020.9054324.

Junshu Wang. *Information Extraction from TV Series Scripts for Uptake Prediction*. PhD thesis, Auckland University of Technology, 2017.

Eetu Sjöblom, Mathias Creutz, and Mikko Aulamo. Paraphrase detection on noisy subtitles in six languages, 2018.

Leonardo Zilio, Rodrigo Wilkens, and Cédrick Fairon. Passport: A dependency parsing model for portuguese. In Aline Villavicencio, Viviane Moreira, Alberto Abad, Helena Caseli, Pablo Gamallo, Carlos Ramisch, Hugo Gonçalo Oliveira, and Gustavo Henrique Paetzold, editors, *Computational Processing of the Portuguese Language*, pages 479–489, Cham, 2018. Springer International Publishing. ISBN 978-3-319-99722-3.

Mitchell A. Gordon and Kevin Duh. Distill, adapt, distill: Training small, in-domain models for neural machine translation, 2020.

Rihards Krišlauks and Mārcis Pinnis. Tilde at wmt 2020: News task systems. *arXiv preprint arXiv:2010.15423*, 2020.

Sungjae Cho, Jaeseo Lim, Chris Hickey, Jung Park, and Byoung-Tak Zhang. Simulating problem difficulty in arithmetic cognition through dynamic connectionist models. *arXiv preprint arXiv:1905.03617*, 2019.

Feng Qi and Wenchuan Wu. Human-like machine thinking: Language guided imagination, 2019.

Alon Talmor, Oyvind Tafjord, Peter Clark, Yoav Goldberg, and Jonathan Berant. Leap-Of-Thought: Teaching pre-trained models to systematically reason over implicit knowledge. *Advances in Neural Information Processing Systems*, 33, 2020.

Georgiana Dinu, Prashant Mathur, Marcello Federico, Stanislas Lauly, and Yaser Al-Onaizan. Joint translation and unit conversion for end-to-end localization, 2020.

Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1905334117. URL https://www.pnas.org/content/117/43/26562.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 3294–3302. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/f442d33fa06832082290ad8544a8da27-Paper.pdf.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Jack Bandy and Nicholas Vincent. Addressing" documentation debt" in machine learning research: A retrospective datasheet for bookcorpus. *arXiv preprint arXiv:2105.05241*, 2021.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40, 2009.

Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

Kenneth Heafield. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197, 2011.

Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217 – 250, 2012. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2012.07.001. URL http://www.sciencedirect.com/science/article/pii/S0004370212000793.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

Jana Diesner, Terrill L Frantz, and Kathleen M Carley. Communication networks from the enron email corpus "it's always about the people. enron is no different". *Computational & Mathematical Organization Theory*, 11(3):201–228, 2005.

Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 435–443, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401945. URL https://doi.org/10.1145/1401890.1401945.

Lei Tang, Huan Liu, Jianping Zhang, and Zohreh Nazeri. Community evolution in dynamic multi-mode networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 677–685, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605581934. doi: 10.1145/1401890.1401972. URL https://doi.org/10.1145/1401890.1401972.

Koby Crammer, Alex Kulesza, and Mark Dredze. Adaptive regularization of weight vectors. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, pages 414–422. Curran Associates, Inc., 2009. URL https://proceedings.neurips.cc/paper/2009/file/8ebda540cbcc4d7336496819a46a1b68-Paper.pdf.

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. An extensive experimental comparison of methods for multi-label learning. *Pattern Recognition*, 45(9):3084 – 3104, 2012. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2012.03.004. URL http://www.sciencedirect.com/science/article/pii/S0031320312001203. Best Papers of Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA'2011).

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.

Declan Groves and Andy Way. Hybridity in mt: Experiments on the Europarl corpus. 2006b.

Sosuke Kobayashi. Homemade bookcorpus. https://github.com/soskek/bookcorpus, 2018.

2020. URL https://www.ncbi.nlm.nih.gov/pmc/about/faq/.

David C Uthus and David W Aha. The Ubuntu chat corpus for multiparticipant chat analysis. Technical report, NAVAL RESEARCH LAB WASHINGTON DC, 2013.

Staff and agencies. Timeline: Enron, Jan 2006. URL http://www.theguardian.com/business/2006/jan/30/corporatefraud.enron.

Joe Bartling. The enron data set - where did it come from? *Bartling Forensic and Advisory*, 2015.

Julia Adams, Hannah Brückner, and Cambria Naslund. Who counts as a notable sociologist on wikipedia? gender, race, and the "professor test". *Socius*, 5:2378023118823946, 2019.

Jiawei Xing and Matthew Vetter. Editing for equity: Understanding instructor motivations for integrating cross-disciplinary wikipedia assignments. *First Monday*, 2020.

Joseph Reagle and Lauren Rhue. Gender bias in wikipedia and britannica. *International Journal of Communication*, 5:21, 2011.

Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. It's a man's wikipedia? assessing gender inequality in an online encyclopedia. In *Ninth international AAAI conference on web and social media*, 2015.

Eszter Hargittai and Aaron Shaw. Mind the skills gap: the role of internet know-how and gender in differentiated contributions to wikipedia. *Information, communication & society*, 18(4):424–442, 2015.

Eduardo Graells-Garrido, Mounia Lalmas, and Filippo Menczer. First women, second sex: Gender bias in wikipedia. In *Proceedings of the 26th ACM conference on hypertext & social media*, pages 165–174, 2015.

Murray G Phillips. Wikipedia and history: a worthwhile partnership in the digital era? *Rethinking History*, 20(4):523–543, 2016.

John Derby. Art education and disability studies. *Disability Studies Quarterly*, 32(1), 2012.

Morten Rask. The reach and richness of wikipedia: Is wikinomics only for rich countries? *First Monday*, 2008.

Youngwhan Lee and Heuiju Chun. Nation image and its dynamic changes in wikipedia. *Asia Pacific Journal of Innovation and Entrepreneurship*, 2017.

Ewa S Callahan and Susan C Herring. Cultural bias in wikipedia content on famous persons. *Journal of the American society for information science and technology*, 62(10):1899–1915, 2011.

Caroline Ball. Using wikipedia to explore issues of systemic bias and symbolic annihilation in information sources. Innovative Libraries, 2021.

Shiri Eisner. *Bi: Notes for a bisexual revolution.* Seal Press, 2013.