EXHIBIT 190

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Leo Gao **Stella Biderman** **Sid Black** **Laurence Golding**

Travis Hoppe **Charles Foster** **Jason Phang** **Horace He**

Anish Thite **Noa Nabeshima** **Shawn Presser** **Connor Leahy**

EleutherAI
contact@eleuther.ai

arXiv:2101.00027v1 [cs.CL] 31 Dec 2020

Abstract

Recent work has demonstrated that increased training dataset diversity improves general cross-domain knowledge and downstream generalization capability for large-scale language models. With this in mind, we present *the Pile*: an 825 GiB English text corpus targeted at training large-scale language models. The Pile is constructed from 22 diverse high-quality subsets—both existing and newly constructed—many of which derive from academic or professional sources. Our evaluation of the untuned performance of GPT-2 and GPT-3 on the Pile shows that these models struggle on many of its components, such as academic writing. Conversely, models trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, while improving performance on downstream evaluations. Through an in-depth exploratory analysis, we document potentially concerning aspects of the data for prospective users. We make publicly available the code used in its construction.[1]

1 Introduction

Recent breakthroughs in general-purpose language modeling have demonstrated the effectiveness of training massive models on large text corpora for downstream applications (Radford et al., 2019; Shoeybi et al., 2019; Raffel et al., 2019; Rosset, 2019; Brown et al., 2020; Lepikhin et al., 2020). As the field continues to scale up language model training, the demand for high-quality massive text data will continue to grow (Kaplan et al., 2020).

The growing need for data in language modeling has caused most existing large-scale language models to turn to the Common Crawl for most or all of their data (Brown et al., 2020; Raffel et al., 2019). While training on the Common Crawl has been effective, recent work has shown that dataset diversity leads to better downstream generalization capability (Rosset, 2019). Additionally, large-scale language models have been shown to effectively acquire knowledge in a novel domain with only relatively small amounts of training data from that domain (Rosset, 2019; Brown et al., 2020; Carlini et al., 2020). These results suggest that by mixing together a large number of smaller, high quality, diverse datasets, we can improve the general cross-domain knowledge and downstream generalization capabilities of the model compared to models trained on only a handful of data sources.

To address this need, we introduce the Pile: a 825.18 GiB English text dataset designed for training large scale language models. The Pile is composed of 22 diverse and high-quality datasets, including both established natural language processing datasets and several newly introduced ones. In addition to its utility in training large language models, the Pile can also serve as a broad-coverage benchmark for cross-domain knowledge and generalization ability of language models.

We introduce new datasets derived from the following sources: PubMed Central, ArXiv, GitHub, the FreeLaw Project, Stack Exchange, the US Patent and Trademark Office, PubMed, Ubuntu IRC, HackerNews, YouTube, PhilPapers, and NIH ExPorter. We also introduce OpenWebText2 and BookCorpus2, which are extensions of the original OpenWebText (Gokaslan and Cohen, 2019) and BookCorpus (Zhu et al., 2015; Kobayashi, 2018) datasets, respectively.

In addition, we incorporate several existing high-quality datasets: Books3 (Presser, 2020), Project Gutenberg (PG-19) (Rae et al., 2019), OpenSubtitles (Tiedemann, 2016), English Wikipedia, DM Mathematics (Saxton et al., 2019), EuroParl (Koehn, 2005), and the Enron Emails corpus (Klimt and Yang, 2004). To supplement these, we also in-

[1] https://pile.eleuther.ai/

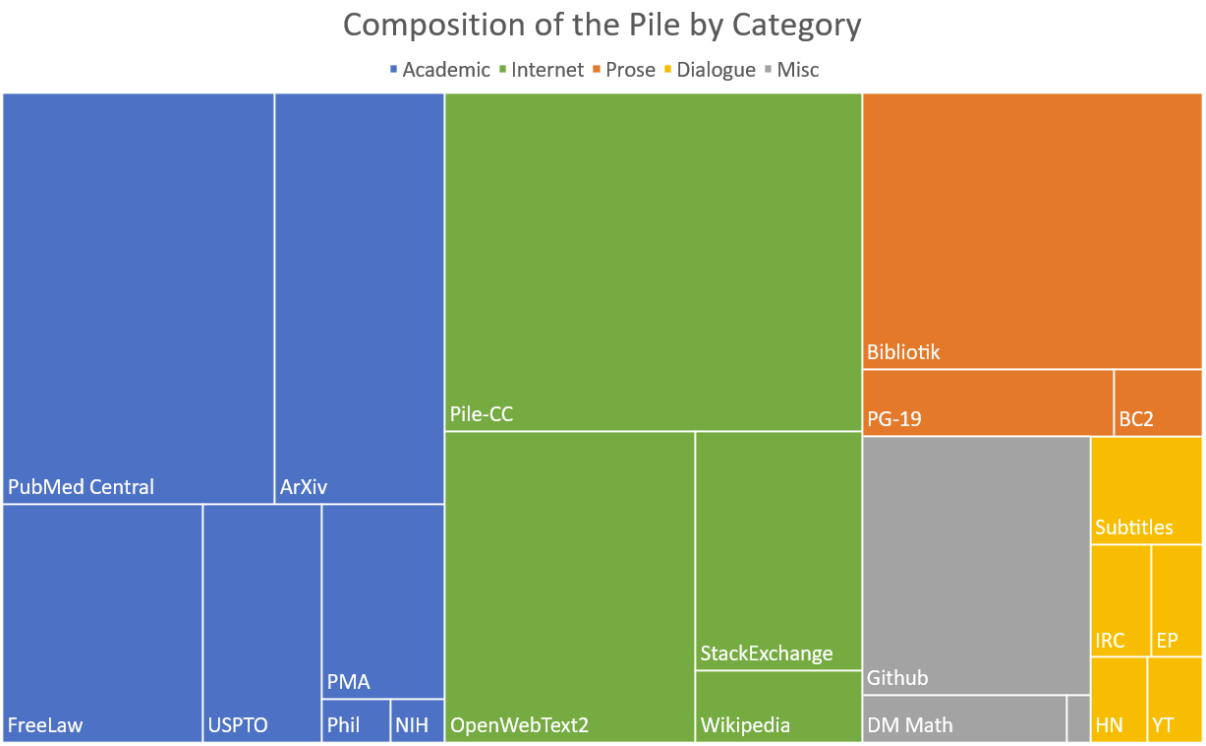


Figure 1: Treemap of Pile components by effective size.

troduce a new filtered subset of Common Crawl, Pile-CC, with improved extraction quality.

Through our analyses, we confirm that the Pile is significantly distinct from pure Common Crawl data. Additionally, our evaluations show that the existing GPT-2 and GPT-3 models perform poorly on many components of the Pile, and that models trained on the Pile significantly outperform both raw and filtered Common Crawl models. To complement the performance evaluations, we also perform an exploratory analysis of the text within the Pile to provide a detailed picture of the data. We hope that our extensive documentation of the construction and characteristics of the Pile will help researchers make informed decisions about potential downstream applications.

Finally, we make publicly available the preprocessing code for the constituent datasets of the Pile and the code for constructing alternative versions[2]. In the interest of reproducibility, we also document all processing performed on each dataset (and the Pile as a whole) in as much detail as possible. For further details about the processing of each dataset, see Section 2 and Appendix C.

[2] https://github.com/EleutherAI/the-pile

1.1 Contributions

The core contributions of this paper are:

1. The introduction of a 825.18 GiB english-language dataset for language modeling combining 22 diverse sources.
2. The introduction of 14 new language modeling datasets, which we expect to be of independent interest to researchers.
3. Evaluations demonstrating significant improvements across many domains by GPT-2-sized models trained on this new dataset, compared to training on CC-100 and raw Common Crawl.
4. The investigation and documentation of this dataset, which we hope will better inform researchers about how to use it as well as motivate them to undertake similar investigations of their own data.

2 The Pile Datasets

The Pile is composed of 22 constituent sub-datasets, as shown in Table 1. Following Brown et al. (2020), we increase the weights of higher quality components, with certain high-quality datasets such as Wikipedia being seen up to 3 times (“epochs”) for

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3†	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19)†	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles†	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en)†	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics†	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl†	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails†	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	**825.18 GiB**			**1254.20 GiB**	**5.91 KiB**

Table 1: Overview of datasets in the Pile before creating the held out sets. Raw Size is the size before any up- or down-sampling. Weight is the percentage of bytes in the final dataset occupied by each dataset. Epochs is the number of passes over each constituent dataset during a full epoch over the Pile. Effective Size is the approximate number of bytes in the Pile occupied by each dataset. Datasets marked with a † are used with minimal preprocessing from prior work.

each full epoch over the Pile. Detailed information about the construction of each dataset is available in Appendix C.

2.1 Pile-CC

Common Crawl is a collection of website crawls from 2008 onwards, including raw web pages, metadata and text extractions. Due to the raw nature of the dataset, Common Crawl has the advantage of including text from diverse domains, but at the cost of varying quality data. Due to this, use of Common Crawl typically necessitates well-designed extraction and filtering. Our Common Crawl-based dataset, Pile-CC, uses jusText (Endrédy and Novák, 2013) on Web Archive files (raw HTTP responses including page HTML) for extraction, which yields higher quality output than directly using the WET files (extracted plaintext).

2.2 PubMed Central

PubMed Central (PMC) 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), providing open, full-text access to nearly five million publications. Most publications indexed by PMC are recent, and their inclusion is mandated for all NIH funded research starting from 2008 by the NIH Public Access Policy. We included PMC in the hopes that it will benefit potential downstream applications to the medical domain.

2.3 Books3

Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker made available by Shawn Presser (Presser, 2020). Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude

larger than our next largest book dataset (BookCorpus2).We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.

2.4 OpenWebText2

OpenWebText2 (OWT2) is a generalized web scrape dataset inspired by WebText (Radford et al., 2019) and OpenWebTextCorpus (Gokaslan and Cohen, 2019). Similar to the original WebText, we use net upvotes on Reddit submissions as a proxy for outgoing link quality. OpenWebText2 includes more recent content from Reddit submissions up until 2020, content from multiple languages, document metadata, multiple dataset versions, and open source replication code. We included OWT2 as a high quality general purpose dataset.

2.5 ArXiv

ArXiv is a preprint server for research papers that has operated since 1991. As shown in fig. 10, arXiv papers are predominantly in the fields of Math, Computer Science, and Physics. 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 these areas. ArXiv papers are written in LaTeX, a common typesetting language for mathematics, computer science, physics, and some adjacent fields. Training a language model to be able to generate papers written in LaTeX could be a huge boon to the research community.

2.6 GitHub

GitHub is a large corpus of open-source code repositories. Motivated by the ability of GPT-3 (Brown et al., 2020) to generate plausible code completions despite its training data not containing any explicitly gathered code datasets, we included GitHub in the hopes that it would enable better downstream performance on code-related tasks.

2.7 FreeLaw

The Free Law Project is a US-registered non-profit that provides access to and analytical tools for academic studies in the legal realm. CourtListener,[3] part of the Free Law Project, provides bulk downloads for millions of legal opinions from federal and state courts. While the full dataset provides multiple modalities of legal proceedings, including dockets, bibliographic information on judges, and other metadata, we focused specifically on court opinions due to an abundance of full-text entries. This data is entirely within the public domain.

2.8 Stack Exchange

The Stack Exchange Data Dump[4] contains an anonymized set of all user-contributed content on the Stack Exchange network, a popular collection of websites centered around user-contributed questions and answers. It is one of the largest publicly available repositories of question-answer pairs, and covers a wide range of subjects—from programming, to gardening, to Buddhism. We included Stack Exchange in the hopes that it will improve the question answering capabilities of downstream models on diverse domains.

2.9 USPTO Backgrounds

USPTO Backgrounds is a dataset of background sections from patents granted by the United States Patent and Trademark Office, derived from its published bulk archives[5]. A typical patent background lays out the general context of the invention, gives an overview of the technical field, and sets up the framing of the problem space. We included USPTO Backgrounds because it contains a large volume of technical writing on applied subjects, aimed at a non-technical audience.

2.10 Wikipedia (English)

Wikipedia is a standard source of high-quality text for language modeling. In addition to being a source of high quality, clean English text, it is also valuable as it is written in expository prose, and spans many domains.

2.11 PubMed Abstracts

PubMed Abstracts consists of the abstracts from 30 million publications in PubMed, the online repository for biomedical articles run by the National Library of Medicine. While the PMC (see Section 2.2) provides full-text access, the subset of coverage is significantly limited and biased towards recent publications. PubMed also incorporates MEDLINE, which expands the coverage of biomedical abstracts from 1946 to present day.

[3] https://www.courtlistener.com/

[4] https://archive.org/details/stackexchange

[5] https://bulkdata.uspto.gov/

2.12 Project Gutenberg

Project Gutenberg is a dataset of classic Western literature. The specific Project Gutenberg derived dataset we used, PG-19, consists of Project Gutenberg books from before 1919 (Rae et al., 2019), which represent distinct styles from the more modern Books3 and BookCorpus. Additionally, the PG-19 dataset is already being used for long-distance context modeling.

2.13 OpenSubtitles

The OpenSubtitles dataset is an English language dataset of subtitles from movies and television shows gathered by Tiedemann (2016). Subtitles provide an important source of natural dialog, as well as an understanding of fictional formats other than prose, which may prove useful for creative writing generation tasks such as screenwriting, speechwriting, and interactive storytelling.

2.14 DeepMind Mathematics

The DeepMind Mathematics dataset consists of a collection of mathematical problems from topics such as algebra, arithmetic, calculus, number theory, and probability, formatted as natural language prompts (Saxton et al., 2019). One major weakness of large language models has been performance on mathematical tasks (Brown et al., 2020), which may be due in part to a lack of math problems in the training set. By explicitly including a dataset of mathematical problems, we hope to improve the mathematical ability of language models trained on the Pile.

2.15 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." BookCorpus is therefore unlikely to have significant overlap with Project Gutenberg and Books3, which consist of published books. BookCorpus is also commonly used as dataset for training language models (Radford et al., 2018; Devlin et al., 2019; Liu et al., 2019).

2.16 Ubuntu IRC

The Ubuntu IRC dataset is derived from the publicly available chatlogs[6] of all Ubuntu-related channels on the Freenode IRC chat server. Chatlog data provides an opportunity to model real-time human interactions, which feature a level of spontaneity not typically found in other modes of social media.

2.17 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, 2006; Van Halteren, 2008; Ciobanu et al., 2017). We use the most current version at time of writing, which consists of the proceedings of the European Parliament in 21 European languages from 1996 until 2012.

2.18 YouTube Subtitles

The YouTube Subtitles dataset is a parallel corpus of text gathered from human generated closed-captions on YouTube. In addition to providing multilingual data, Youtube Subtitles is also a source of educational content, popular culture, and natural dialog.

2.19 PhilPapers

The PhilPapers[7] dataset consists of open-access philosophy publications from an international database maintained by the Center for Digital Philosophy at the University of Western Ontario. We included PhilPapers because it spans a wide body of abstract, conceptual discourse, and its articles contain high quality academic writing.

2.20 NIH Grant Abstracts: ExPORTER

The NIH Grant abstracts provides a bulk-data repository for awarded applications through the ExPORTER[8] service covering the fiscal years 1985-present. We included the dataset because it contains examples of high-quality scientific writing.

2.21 Hacker News

Hacker News[9] is a link aggregator operated by Y Combinator, a startup incubator and investment fund. Users submit articles defined as "anything that gratifies one's intellectual curiosity," but submitted articles tend to focus on topics in computer science and entrepreneurship. Users can comment on submitted stories, resulting in comment trees discussing and critiquing submitted stories. We

[6] https://irclogs.ubuntu.com/

[7] https://philpapers.org/

[8] https://exporter.nih.gov/

[9] https://news.ycombinator.com

scrape, parse, and include these comment trees since we believe they provide high quality dialogue and debate on niche topics.

2.22 Enron Emails

The Enron Emails dataset (Klimt and Yang, 2004) is a valuable corpus commonly used for research about the usage patterns of email. We included Enron Emails to aid in understanding the modality of email communications, which is typically not found in any of our other datasets.

3 Benchmarking Language Models with the Pile

While the Pile was conceived as a training dataset for large-scale language models, its coverage of multiple disparate domains makes it also suitable as an evaluation dataset. In this section, we describe how the Pile can be used as a broad-coverage dataset for benchmarking language models.

3.1 Benchmarking Guidelines

The Pile is provided as train, validation, and testing splits. The validation and testing components each contain 0.1% of the data, sampled uniformly at random. While this is a far smaller percentage than most datasets, the sheer size of the dataset results in over 1 GiB of validation and testing data each. We highlight that while we have made efforts to deduplicate documents within the Pile (See: Section D.2), it is still possible that some documents are duplicated across the train/validation/test splits.

Our preferred metric is bits per UTF-8 encoded byte (BPB). Bits per byte is preferred over bits per character or perplexity when using Pile as a metric due to its invariance to different tokenization schemes and the ambiguity of measuring characters in Unicode. To compute bits per byte from a given negative log likelihood loss ℓ, we compute $\text{BPB} = (L_T/L_B)\log_2(e^\ell) = (L_T/L_B)\ell/\ln(2)$, where L_T is the length of the dataset in tokens and L_B is the length of the dataset in UTF-8 encoded bytes. We find that L_T/L_B is 0.29335 GPT-2-tokens/byte across the Pile; dataset-specific values of L_T/L_B can be found in Table 7.

3.2 Test Perplexity with GPT-2 and GPT-3

We compute the test perplexity of the constituent datasets of the Pile using GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), shown in Figure 2. We use all available versions of GPT-2, and all four versions of GPT-3 available via the OpenAI API. Because of the cost associated with using the OpenAI API, we evaluate on one-tenth of the respective test sets for most of the constituent datasets. We report the perplexity converted to bits per UTF-8 encoded byte (BPB). Importantly, we compute perplexity by evaluating each document independently within each dataset, as opposed to concatenating all documents as is common practice for computing perplexity on large corpora.

Full details of the perplexity computation can be found in Appendix E.2.

Unsurprisingly, larger language models generally attain lower perplexity compared to smaller models. Recent work has shown an increased focus on the empirical scaling laws of language models (Kaplan et al., 2020; Henighan et al., 2020). As such, we investigate the scaling law for the GPT-2 and GPT-3 families of models on perplexity evaluation on the Pile. The scaling law relation for the GPT-3 family of models is shown in Figure 2.[10] The line of best fit shown in the figure has a coefficient of -0.1674 and an intercept of 2.5516.

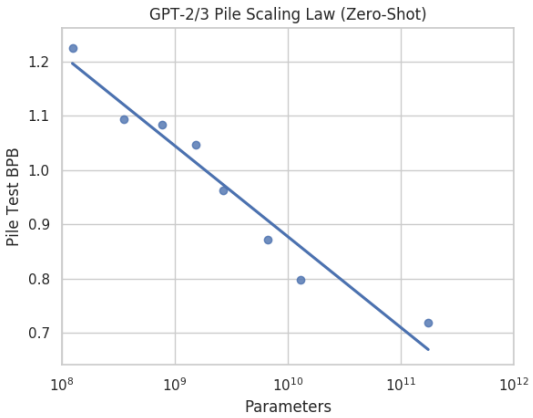


Figure 2: Scaling law for performance of GPT-2/3 models. 'Zero-shot' refers to the fact that none of the models have been fine-tuned on data from the Pile.

Interestingly, while GPT-2 and GPT-3 were not trained on the Pile, there still appears to be a clear scaling law without diminishing returns. We hypothesize that this is due to the inherent generalization capability of these models. We leave a more

[10] While the sizes of GPT-3 models on the OpenAI API have not been publicized, we assume here that `ada`, `babbage`, `curie` and `davinci` models correspond to 2.7B, 6.7B, 13B and 175B parameter models respectively.

rigorous analysis of zero-shot scaling laws to future work.

3.3 Relative Componentwise GPT-3 Pile Performance

Determining which components GPT-3 underperforms on provides information about which Pile components are most dissimilar to the distribution of text (web pages and books) that GPT-3 was trained on. These components would thus make especially good candidates for supplementing GPT-3 training data. These results are also valuable for determining which types of datasets to emphasize for future iterations of the Pile.

Due to the difference in entropy of different datasets, directly comparing perplexity of GPT-3 on different Pile components is not an accurate indication of relative performance. Ideally we would train a GPT-3 model from scratch on the Pile and compare the difference in loss per dataset with that of the original GPT-3. Because of resource constraints, we instead use a GPT-2 model trained from scratch on the Pile (see Section 4) to construct a proxy measure. To construct our proxy, we first measure the improvement from the GPT-2-Pile model to GPT-3 on each component. Then, we normalize our results by setting the change on OpenWebText2 to be zero. This computation is shown in the equation below:

$$\Delta_{\text{set}} = \left(L_{\text{set}}^{\text{GPT3}} - L_{\text{owt2}}^{\text{GPT3}}\right) - \left(L_{\text{set}}^{\text{GPT2Pile}} - L_{\text{owt2}}^{\text{GPT2Pile}}\right)$$

Since GPT2-Pile was trained on both OWT2 and the dataset we are evaluating, we expect the second term in Δ_{set} to reflect the difference in the intrinsic difficulty of the two datasets. Thus the total value of Δ_{set} reflects how much harder the dataset we are evaluating was for GPT-3 than OWT2, minus the relative difficulty of the two tasks. As GPT-3 was trained on data very similar to OWT2, this gives us a proxy for how much better GPT-3 would do if it were trained on the Pile.

The results are shown in Figure 3. As a sanity check, we observe that datasets that are contained in, or are extremely similar to, GPT-3's training set (Books3, Wikipedia (en), Pile-CC and Project Gutenberg) score close to zero on our metric.

GPT-3 appears to perform poorly on datasets pertaining to research or academic writing like PubMed Central, PubMed Abstracts, and ArXiv; domain-specific datasets like FreeLaw, HackerNews, and USPTO Backgrounds; and on datasets containing predominantly text distinct from natural language, like GitHub and DM Mathematics. In addition, the majority of datasets see less of an improvement than OpenWebText2. As such, we expect a GPT-3 sized model trained on Pile to perform significantly better on research related tasks, software tasks, and symbol manipulation tasks than the base model. Additionally, this experiment provides evidence that the majority of Pile components are not redundant with the predominantly web-based GPT-3 training data.

We note that this metric is only a proxy for similarity, and that it could be confounded by dataset specific scaling effects. Although our results largely accord with expectations, there are some puzzling results, like the datasets on which GPT-3 outperformed GPT-2 Pile. We hypothesize that GPT-3 learns to be so good at these datasets that training on them explicitly does not notably benefit the model's performance. We leave a more rigorous analysis of these effects for future work.

4 Evaluation

To confirm the effectiveness of the Pile for improving language modeling quality, we train architecturally-identical 1.3 billion parameter models based on those in Brown et al. (2020) on different datasets and evaluate on the WikiText and LAMBADA tasks as benchmarks of language modeling ability. We also report results on the Pile as a measure of more cross-domain generalization.

4.1 Methodology

To ensure a fair comparison across datasets of different sizes, we decontaminate any instances of the evaluation sets using the same 13-gram overlap filtering as in Brown et al. (2020) and downsample to 40GB to control for dataset size. As we control for dataset size, we emphasize that our evaluation is generous to CC-100 (en), which is about 1/3 the size of the Pile in reality.

We compare the following datasets: the Pile, the En-

Component	GPT-2				GPT-3			
	small	medium	large	xl	ada	babbage	curie	davinci
Pile-CC	1.0878	0.9992	0.9582	0.9355	0.9212	0.8483	0.7849	**0.7070**
PubMed Central	1.0759	0.9788	0.9334	0.9044	0.8633	0.7792	0.7150	**0.6544**
Books3	1.1959	1.1063	1.0588	1.0287	0.9778	0.9005	0.8284	**0.7052**
OpenWebText2	1.1111	1.0073	0.9539	0.9171	0.8727	0.7921	0.7199	**0.6242**
ArXiv	1.3548	1.2305	1.1778	1.1381	1.0304	0.9259	0.8453	**0.7702**
Github	1.7912	1.3180	1.7909	1.6486	0.8761	0.7335	0.6415	**0.5635**
FreeLaw	1.0512	0.9321	0.9017	0.8747	0.8226	0.7381	0.6667	**0.6006**
Stack Exchange	1.2981	1.1075	1.0806	1.0504	1.0096	0.8839	0.8004	**0.7321**
USPTO Backgrounds	0.8288	0.7564	0.7202	0.6969	0.6799	0.6230	0.5752	**0.5280**
PubMed Abstracts	0.9524	0.8579	0.8108	0.7810	0.8130	0.7382	0.6773	**0.6201**
Gutenberg (PG-19)	1.2655	1.1140	1.0820	1.0829	0.9776	0.8749	0.7930	**0.7115**
OpenSubtitles	1.2465	1.1657	1.1324	1.1129	1.1116	1.0488	0.9875	**0.9130**
Wikipedia (en)	1.1285	1.0213	0.9795	0.9655	0.8757	0.7863	0.7047	**0.5953**
DM Mathematics	2.6911	2.5448	2.4833	2.4377	2.3249	2.2015	2.1067	**2.0228**
Ubuntu IRC	1.8466	1.7187	1.6427	1.6024	1.3139	1.1968	1.0995	**0.9915**
BookCorpus2	1.1295	1.0498	1.0061	0.9783	0.9754	0.9041	0.8435	**0.7788**
EuroParl	2.3177	2.0204	1.8770	1.7650	1.0475	0.9363	0.8415	**0.7519**
HackerNews	1.4433	1.2794	1.3143	1.3361	1.1736	1.0875	1.0175	**0.9457**
YoutubeSubtitles	2.0387	1.8412	1.7355	1.6694	1.3407	1.1876	1.0639	**0.9469**
PhilPapers	1.3203	1.2163	1.1688	1.1327	1.0362	0.9530	0.8802	**0.8059**
NIH ExPorter	0.9099	0.8323	0.7946	0.7694	0.7974	0.7326	0.6784	**0.6239**
Enron Emails	1.5888	1.4119	1.4535	1.4222	1.2634	1.1685	1.0990	**1.0201**
The Pile	1.2253	1.0928	1.0828	1.0468	0.9631	0.8718	0.7980	**0.7177**

Table 2: Test perplexity of the Pile using GPT-2 and GPT-3, converted to bits per UTF-8 encoded byte (BPB). Evaluation is performed on one-tenth of the test data of the Pile, on a per-document basis. **Bold** indicates the best-performing model in each row.

glish component of the CC-100 dataset[11] (Wenzek et al., 2019; Conneau et al., 2020), and a sample of raw CC WET files filtered for English-only.

4.2 Results

On traditional language modeling benchmarks, the Pile improves significantly on WikiText and shows negligible changes in LAMBADA. However, models trained on Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, as shown in Table 4. This indicates that models trained on the Pile have greater cross-domain generalization capabilities without compromising performance on traditional benchmarks.

The magnitude of improvement over CC-100 per set is shown in Figure 4. Unsurprisingly, there is almost no improvement on Pile-CC. However, the model trained on the Pile performs significantly better than either of the other models on academic datasets such as ArXiv, Pubmed Central, FreeLaw, and PhilPapers. It also improves significantly on programming-related datasets like Github and StackExchange, on EuroParl, due to the lack of multilingual text in either other dataset, and on DM Mathematics, indicating a significant improvement in mathematical ability.

Surprisingly, raw Common Crawl performs better on the Pile BPB than CC-100, despite losing by a significant margin on LAMBADA and WikiText. We hypothesize that this is due to the perplexity based filtering used in CC-100, where a language model is trained on Wikipedia and all data with a perplexity too high or too low is discarded. This effectively discards any data too similar to or too different from Wikipedia, which severely limits the diversity of the collected data. This result suggests that future work using Common Crawl should take caution with filtering to preserve its diversity.

5 Structural Statistics

In this section, we cover the Structural Statistics of the dataset, which provide more coarse-grained and statistical information about the Pile. In Sec-

[11] The data was obtained from http://data.statmt.org/cc-100/.

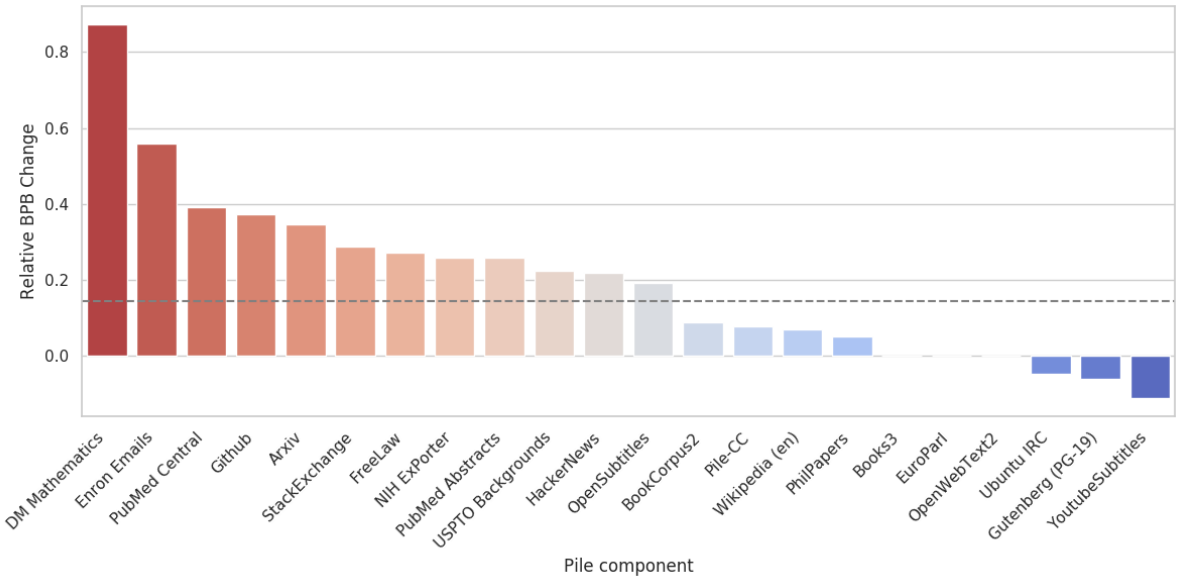


Figure 3: Change in BPB from GPT-2 trained on Pile to GPT-3 zero-shot, relative to OpenWebText2 BPB change. Dotted line indicates overall Pile change. Lower indicates better relative performance by GPT-3.

	Dataset Size	Pile (val) (BPB)	Pile (test) (BPB)	WikiText (PPL)	LAMBADA (PPL)	LAMBADA (ACC)
The Pile	825 GiB	**0.9281**	**0.9433**	**5.59**	12.78	**50.1**
CC-100 (en)	300 GiB	1.3143	1.3293	8.27	**11.78**	49.7
Raw CC	45927 GiB†	1.1180	1.1275	11.75	19.84	43.8

Table 3: Size-controlled evaluation results. Each dataset is deduplicated against all evaluation metrics and subsampled to approximately 40GB to control for the effects of dataset size. For LAMBADA, we use the variant of the data introduced in Radford et al. (2019) and only evaluate the perplexity on the final token rather than the final word. For WikiText, we report the perplexity per GPT-2 token. † indicates that the size is an estimate.

tion 6, we provide a closer investigation and documentation of the textual content within the Pile datasets.

5.1 Document Lengths and Tokenization

Each dataset consists of a large number of documents. We analyze the distribution of document lengths, as well as the number of bytes-per-token using the GPT-2 tokenizer in order to put our ablations in context.

While the majority of documents in the Pile are short, there is a long tail of very long documents (Figure 5).

Since the GPT-2 BPE tokenizer is trained on WebText, the mean bytes per token is also a very rough indicator of how syntactically different each Pile component is from WebText. For instance, datasets like NIH ExPorter, OpenWebText2 and Books3 consist largely of ordinary text in a similar distribution to WebText, which is reflected in a greater number of bytes per token. On the other hand, many of the sets with the lowest bytes per token are those which consist in large part of non-text content (Github, ArXiv, Stack Exchange, and DM Mathematics) or languages other than English (EuroParl).

5.2 Language and Dialects

While only 13% of the world's population speaks English, the vast majority of NLP research is done on English. For the Pile, we took a similar approach to the dataset used by Brown et al. (2020) and focused predominantly on English, while also not explicitly filtering out other languages when collecting our own data. When evaluating a multilingual dataset, our main criteria for inclusion was whether the English component of the dataset merited inclusion alone. We plan to create a fully multi-

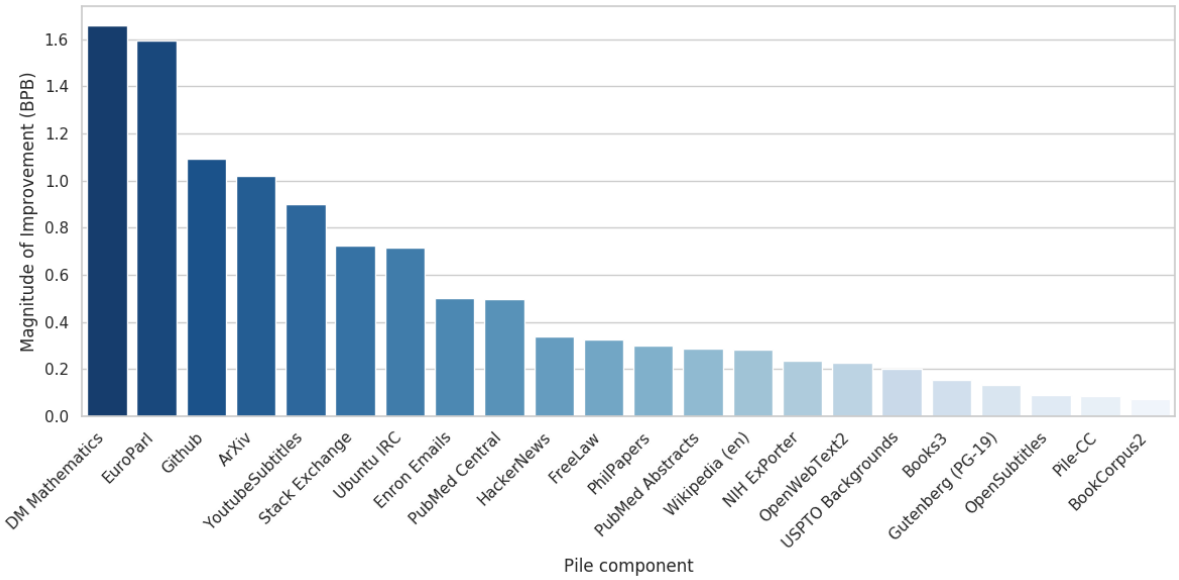


Figure 4: Magnitude of BPB improvement of Pile model over CC-100 model on each test set.

Dataset	The Pile	CC-100 (en)	Raw CC (en)
Pile-CC	**0.9989**	1.0873	1.0287
PubMed Central	**0.6332**	1.1311	0.9120
Books3	**1.0734**	1.2264	1.1366
OpenWebText2	**0.9938**	1.2222	1.0732
ArXiv	**0.7945**	1.8159	1.2642
Github	**0.5597**	1.6509	0.9301
FreeLaw	**0.6978**	1.0221	0.9468
Stack Exchange	**0.8152**	1.5414	1.1292
USPTO Backgrounds	**0.6731**	0.8772	0.8455
PubMed Abstracts	**0.7313**	1.0193	0.9718
Gutenberg (PG-19)	**1.1426**	1.2780	1.2235
OpenSubtitles	**1.0909**	1.1827	1.2139
Wikipedia (en)	**0.8961**	1.1807	1.0252
DM Mathematics	**1.5206**	3.1774	2.6229
Ubuntu IRC	**1.4085**	2.1243	1.5691
BookCorpus2	**1.0613**	1.1346	1.0914
EuroParl	**1.1202**	2.7141	1.4917
HackerNews	**1.0968**	1.4352	1.2305
YoutubeSubtitles	**1.4269**	2.3287	1.5607
PhilPapers	**1.1256**	1.4269	1.2090
NIH ExPorter	**0.7347**	0.9713	0.9225
Enron Emails	**0.8301**	1.3300	1.0483

Table 4: Breakdown of BPB on Pile heldout test set. Columns indicate the dataset each model is trained on; rows indicate the evaluation dataset. **Bold** indicates the best performing model in each row.

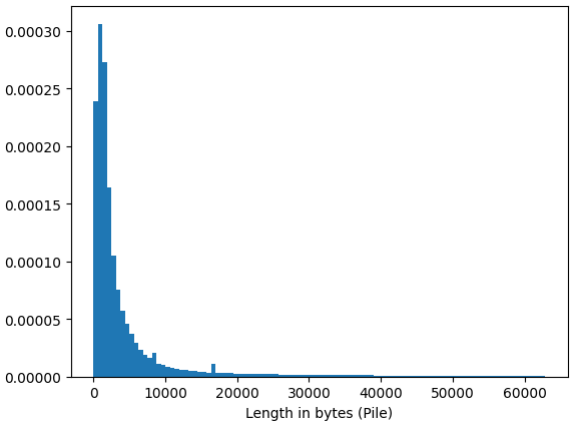


Figure 5: Distribution of document lengths in Pile. The highest 1 percentile of document length are considered to be outliers and excluded from this plot.

lingual expansion of the Pile as future work.

Using fasttext (Suárez et al., 2019a), we determine that the Pile is 97.4% English. We note that due to issues with language identification, particularly with rare languages Caswell et al. (2020), this methodology provides only a rough estimate for English content and no reliable conclusions for low-resource languages can be drawn.

6 Investigating and Documenting the Datasets

As the scale of machine learning research has grown, scrutiny has been placed on the ever larger datasets that models are trained on (Prabhu and Birhane, 2020; Biderman and Scheirer, 2020)

While this issue has been raised within AI ethics and bias research (Hovy and Spruit, 2016; Hutchinson et al., 2020; Blodgett et al., 2020), it has not been a focal point of concern within the language modeling community. Despite the proliferation of work exploring and documenting issues with datasets (Gebru et al., 2018; Bender and Friedman,

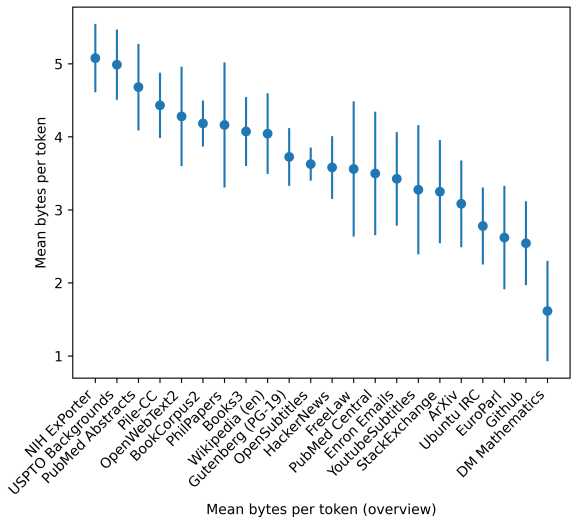


Figure 6: Mean bytes per GPT-2-token for each dataset in the Pile. Error bars indicate standard deviation.

2018; Jo and Gebru, 2020), no dataset intended to train massive language models has been seriously documented by its creators[12]. Therefore, our analyses serve two goals: to address ethical concerns about the Pile, and to promote and normalize the practice of engaging with the AI ethics literature.

Natural language processing technologies are widely applicable and can be used in extremely different contexts. What is and is not appropriate data to train on can therefore vary wildly with the application context. In our view, the best approach is to *document* rather than *eliminate* potentially concerning aspects of datasets[13], particularly since the purpose of the Pile is to train general-purpose language models. The primary goal of our documentation, therefore, is to empower NLP researchers to make informed decisions.

6.1 Documenting Methods

To document the Pile, we chose to implement two frameworks that have been proposed by methodologists and ethics researchers. The first, the datasheets methodology (Gebru et al., 2018), is a general purpose methodology that is recommended by several methodologists (Raji and Yang, 2019; Biderman and Scheirer, 2020) and appears to be used more frequently by practitioners than alternatives (Seck et al., 2018; Costa-jussà et al., 2020; Thieme et al., 2020). The second, the data statements methodology (Bender and Friedman, 2018), was proposed specifically for natural language processing and has been well received by the NLP community. Our datasheet and data statement will be featured in the GitHub repository where the code for the Pile is stored and will also be available as separate documents on arXiv (Biderman et al., 2021; Biderman, 2021).

In addition to the datasheet and data statement, there is additional information that may be helpful to people training language models that these documents do not cover. In the rest of this section we investigate and document in greater detail some of this additional contextual information.

6.2 Topical Distribution

In order to better understand the specific subject matter covered by the Pile, we performed a topic modeling analysis on its components. Using Gensim (Rehurek et al., 2011), we trained 16-topic Latent Dirichlet Allocation (Blei et al., 2003) models on each component of the validation set of the Pile concurrently, in an online fashion (Hoffman et al., 2010). We filtered the Pile for English only for this analysis. Afterwards, we computed the perplexity of the Common Crawl-derived (Pile-CC) topic model on the document sets of the other components. In this way, we provide a rough measure of the degree to which parts of the Pile contain topics not well covered within Common Crawl.

In Figure 7, these cross-component perplexities are shown, with a vertical line indicating the perplexity of the Pile-CC topic model evaluated on the documents of OpenWebText2. This component was chosen as a baseline of comparison for similar reasons as in the previous evaluation: it is derived in a similar manner (filtered crawls of the open web) as the Common Crawl, and thus is expected to contain a similar distribution of topics. Although Pile-CC is somewhat diverse in its content, several of the Pile's other components deviate from it strongly in their topical focus, as evidenced by higher perplexity on Github, PhilPapers, and EuroParl.

We also documented the topical clusters inferred from our LDA models for each component, which we provide in Appendix C. As expected, though the larger CC-derived component itself represents a diversity of content—including politics, education,

[12] Brown et al. (2020) discusses ethical issues surrounding their *model*, but do not discuss those surrounding the training dataset itself.

[13] That said, we did exclude several datasets, see Appendix B for details.

sports and entertainment—the content clusters it misses become apparent when compared qualitatively to other components of the Pile. Notably, the data modes covering programming, logic, physics, and legal knowledge appear largely absent.

6.3 Pejorative Content

Due to the wide diversity in origins, it is possible for the Pile to contain pejorative, sexually explicit, or otherwise objectionable content. As this content may not be desirable for some use cases, we break down profanity on a per-dataset level.

We used the `profanity-checker` Python package (Zhou, 2019). This package includes a "toxicity model" trained on multiple profanity lists as well as the Wikidetox Toxic Comment Dataset (Wulczyn et al., 2016) and classifies a given string as being profane or not profane.

We considered only the English sentences in each dataset using the same language classifier from Section 3.7. We did this since `profanity-checker` is built for English and other languages may improperly impact the results. For instance, the German nominative/accusative feminine/plural definite article "die" is flagged as being profane regardless of context. We split each sentence into words and computed the percentage of words that are flagged as profane for each component of the Pile. We emphasize that this methodology is only a proxy for profanity, given the complexity of determining whether a given word or phrase is profane in context.

As shown in Figure 8, the Pile as a whole appears less profane than Pile-CC. Further, the majority of Pile components appear less profane than Pile-CC as well.

We also broke each dataset down on a sentence level, to allow `profanity-checker` to check entire sentences. Splitting datasets by sentence allows for additional context to be considered when determining whether content is pejorative. Our results are shown in Figure 12.

6.4 Bias and Sentiment Co-occurrence

As language models may pick up unexpected biases from the training data, we performed a preliminary analysis of the different components that make up the Pile. Because models with different characteristics may be trained on the Pile, we aimed to document the biases of the data and not a specific model. We primarily focus on co-occurrence tests, where we analyzed what words occur in the same sentence as other specific words. Using this information, we can estimate what words strongly bias towards a category word, as well as calculate the general sentiment of surrounding words.

We focused our analysis on gender, religion, and race. Our goal is to provide users of this dataset with preliminary guidance on how the different components are biased so that they can make decisions on which components to train on.

All tables and figures in this section can be found in the Appendix.

6.4.1 Gender

We computed gender associations by computing co-occurrences for binary pronouns. For each word, we computed the difference in the rate it co-occurs with "he" and "she"[14] and weighed it by the square root of its frequency. We report the top 15 most biased adjectives or adverbs (Loper and Bird, 2002) for each in Table 10. We see that words like "military", "criminal", and "offensive" strongly bias towards men, while "little", "married", "sexual", and "happy" bias towards women.

In addition, we computed the average sentiment (Baccianella et al., 2010) of words co-occurring with the gendered pronouns across each dataset in Figure 13. Generally, we find no significant sentiment bias towards men or women. This, of course, does not mean that the dataset is free of gender bias (as our co-occurrence tests show).

6.4.2 Religion

We computed a similar co-occurrence analysis for religion, which can be found in Table 11. Like gender, we find that these co-occurrences reflect how these terms are used in pockets of online discourse. For example, "radical" co-occurs with "muslim" at a high rate, while "rational" often co-occurs with "atheist". This analysis also demonstrates some of the limitations of a purely co-occurrence based analysis. For example, "religious" often co-occurs with "atheist", which likely reflects the type of conversations in which the word "atheist" is likely to occur as opposed to a descriptor of "atheist".

[14] We chose to only study male and female pronouns as a simplifying assumption. Studying "they" would require us to isolate its usage as a singular noun.

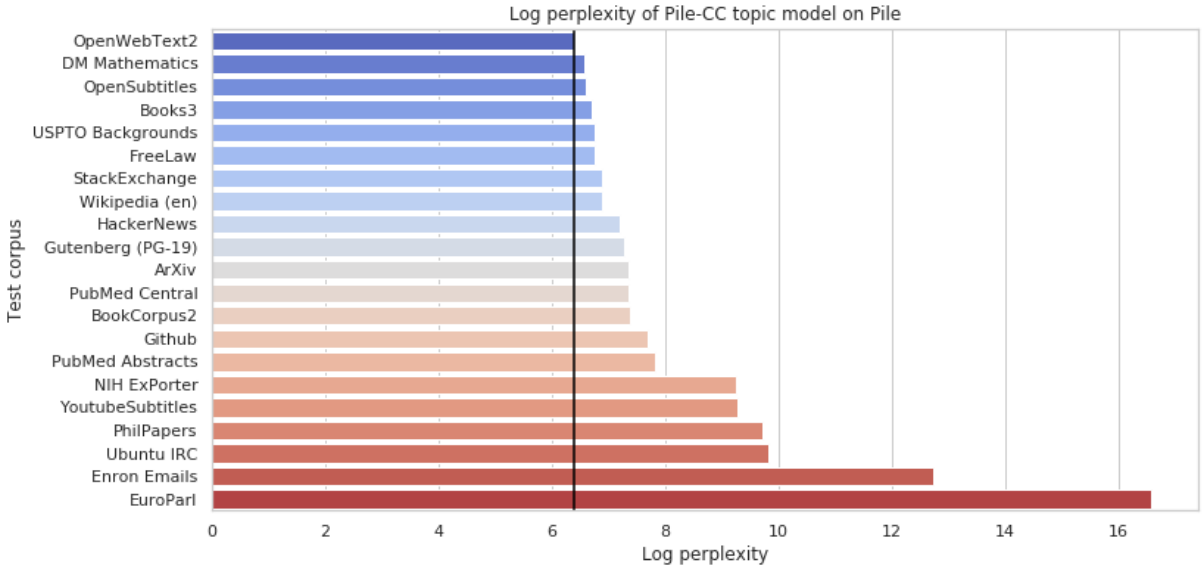


Figure 7: Log perplexity of 16-topic LDA trained on Pile-CC, on other Pile components. Dotted line indicates log perplexity of the topic model on OpenWebText2. Higher indicates a larger topical divergence from Pile-CC.

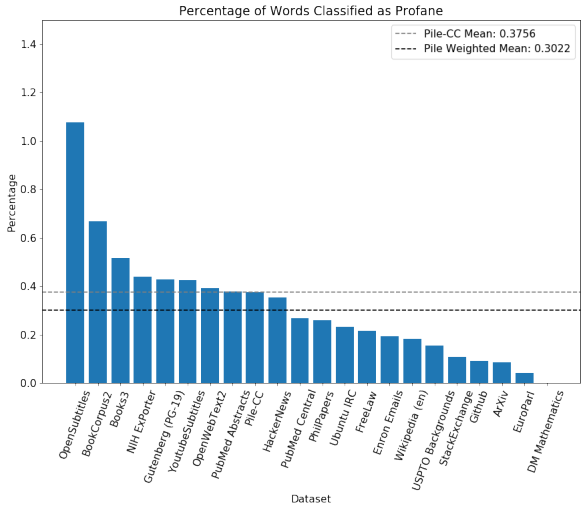


Figure 8: Percentage of words classified as profane in the Pile. The percentage of the CC component and the weighted mean of the Pile as a whole are shown as horizontal lines.

In addition, we computed the average sentiment of co-occurrences across each of the constituent datasets in Figure 14. Over the entire dataset, we find that "Buddhist" has the highest sentiment, followed by "Hindu", "Christian", "Atheist", and "Muslim". Notably, "Jew" is the lowest, perhaps reflecting its historical use as a pejorative.

6.4.3 Race

Finally, we ran the same analysis for racial groups. Here, as identifiers like "black" or "white" often do not indicate race, we instead compute co-occurences with phrases like "black man" or "white woman".

We show the top 15 most biased words for each demographic in Table 12. Once again, we found that the co-occurrences reflect the context in which these terms are used. For example, the 4 most biased words for "black" are "unarmed", "civil", "criminal", and "scary".

Similar to above, we compute the average sentiment of co-occurring words. We report the average sentiment numbers in Table 13. We find that "hispanic/latino" narrowly edges out "asian" for the highest sentiment, followed by "white". On the other hand, "black" had the lowest sentiment, at -0.15.

We note that for all demographics, the average sentiment is negative. We hypothesize that this is due to the specific context for which the phrases we use to compute co-occurrences appear. For example, it is often quite common for news articles to describe suspects as an "asian man".

6.5 Author Consent and Public Data

Another issue with the use of texts in natural language processing research is consent. Although one is typically not legally obligated to receive the permission of an author to train a NLP algorithm on their work[15], many consider doing so a moral obli-

[15]Laws vary by country. For a discussion of US law, see Section 7.1

gation or a good measure to guard against misuse (Obar, 2020; Prabhu and Birhane, 2020). On the other hand, there is significant disagreement surrounding the ethics of repurposing data protected by *terms of service* in research contexts (Vitak et al., 2016; Fiesler et al., 2020), particularly given the power asymmetries inherent in digital platforms, which often close off independent researchers from investigating public data while simultaneously compelling users to consent to its private use (Halavais, 2019).

While much of the Pile's data comes from sources that have expressly consented to its wider dissemination and use in research, researchers often fail to clearly document where their data came from and under what terms its use was consented to. In light of this, we felt it appropriate to release the Pile with transparency around how the authors of its data have indicated that that data can be used.

To provide needed nuance to our discussion of consent, we identified three tiers of availability for public use. **Public data** is data which is freely and readily available on the internet. This primarily excludes data which is pay-walled (regardless of how easy that paywall is to bypass) and data which cannot be easily obtained but can be obtained, e.g. through a torrent or on the dark web. **Terms of Service (ToS) compliant data** is data which is obtained and used in a fashion that is known to be consistent with the terms of service of the data host. **Data with authorial consent** is data for which the original authors of the work consented to the use of their data, or where a reasonable person could not assume that their data would not be used for purposes such as research. ToS compliant data and authorial consented data differ in two main ways: It is important to keep in mind that people typically do not read Terms of Service, and additionally that being ToS-compliant does not entail authorial consent. We adopted a strict model of consent, where ambiguous or unknown consent is treated as non-consensual.

Table 5 summarizes our understanding of the status of each of the datasets within the Pile. Datasets marked with a ✓are compliant in the relevant respects, though a couple datasets are worth remarking on in particular. Book3 and OpenSubtitles are being used in a fashion that is consistent with the terms of service *of the data host*. However, this is somewhat misleading in that the data host is not authorized to post the data online by the parties that own it. The Enron Emails dataset was not collected with the permission of the authors, but was collected by the U.S. government as part of a criminal investigation. While the people whose emails are in the Enron dataset are aware of this fact, they were not given the ability to consent to its inclusion in any way.

There are five datasets included in the Pile that were not collected and distributed in a ToS compliant fashion and for which the authors had no ability to consent to their data being used. Each of these datasets are widely used, both in the NLP literature and the world at large. With the exception of the YouTube Subtitles dataset, each of these datasets were published by researchers and are passed around freely on the internet. The YouTube Subtitles dataset was created by us for this project, using a very popular unofficial API that is both widely used and easily obtainable on Pip, Conda, and GitHub, among other places. Given the processing applied and the difficulty of identifying particular files in the Pile, we feel that our use of these datasets does not constitute significantly increased harm beyond that which has already been done by the widespread publication of these datasets.

7 Implications and Broader Impacts

The Pile represents yet another stepping stone along the path of scaling models and datasets to ever larger sizes and capabilities. There are many serious concerns about how the emergence of progressively stronger AI systems will influence the wider world (Brundage et al., 2018; Amodei et al., 2016; Bostrom and Yudkowsky, 2014; Bostrom, 2014; Critch and Krueger, 2020), and we believe that they merit serious thought. In this section we discuss the legal ramifications of the Pile, and then consider the impact of the Pile to AI alignment from two angles: accelerating AI timelines and the dangers posed by unaligned language models.

7.1 Legality of Content

While the machine learning community has begun to discuss the issue of the legality of training models on copyright data, there is little acknowledgment of the fact that the processing and distribution of data owned by others may also be a violation of copyright law. As a step in that direc-

Component	Public	ToS	Author
Pile-CC	✓	✓	
PMC	✓	✓	✓
Books3	✓		
OWT2	✓		
ArXiv	✓	✓	✓
Github	✓	✓	
FreeLaw	✓	✓	✓
Stack Exchange	✓	✓	✓
USPTO	✓	✓	✓
PubMed	✓	✓	✓
PG-19	✓	✓	
OpenSubtitles	✓		
Wikipedia	✓	✓	✓
DM Math	✓	✓	✓
Ubuntu IRC	✓	✓	✓
BookCorpus2	✓		
EuroParl	✓	✓	✓
HackerNews	✓	✓	
YTSubtitles	✓		
PhilPapers	✓	✓	✓
NIH	✓	✓	✓
Enron Emails	✓	✓	

Table 5: Types of consent for each dataset

tion, we discuss the reasons we believe that our use of copyright data is in compliance with US copyright law.[16]

Under pre (1984) (and affirmed in subsequent rulings such as aff (2013); Google (2015)), non-commercial, not-for-profit use of copyright media is preemptively fair use. Additionally, our use is *transformative*, in the sense that the original form of the data is ineffective for our purposes and our form of the data is ineffective for the purposes of the original documents. Although we use the full text of copyright works, this is not necessarily disqualifying when the full work is necessary (ful, 2003). In our case, the long-term dependencies in natural language require that the full text be used in order to produce the best results (Dai et al., 2019; Rae et al., 2019; Henighan et al., 2020; Liu et al., 2018).

Copyright law varies by country, and there may be additional restrictions on some of these works in particular jurisdictions. To enable easier compliance with local laws, the Pile reproduction code is available and can be used to exclude certain components of the Pile which are inappropriate for the user. Unfortunately, we do not have the metadata necessary to determine exactly which texts are copyrighted, and so this can only be undertaken at the component level. Thus, this should be be taken to be a heuristic rather than a precise determination.

[16]This discussion does not, and is not intended to, constitute legal advice; rather, it is a general discussion of law. Only your attorney can provide assurances that the information contained herein is applicable or appropriate to a particular situation. If in doubt, it is always advisable to speak to an intellectual property attorney.

7.2 Acceleration of AI Timelines

There is serious concern that AI systems may soon be meaningfully more capable than humans in all relevant economic tasks (Grace et al., 2018; Yudkowsky, 2013). Relatedly, there are serious unresolved questions surrounding how to properly align such powerful AI systems with human interests (Bostrom and Yudkowsky, 2014; Russell, 2019; Bostrom, 2014; Amodei et al., 2016) and generally avoid morally catastrophic outcomes (Sotala and Gloor, 2017; Shulman and Bostrom, 2020). As such, it has been argued that accelerating the development of such powerful AI systems may be undesirable before these concerns have been more adequately addressed (Bostrom, 2014).

There are several pragmatic responses to this view:

1. Due to human competition, curiosity, and cultural diversity, halting technological development is incredibly difficult, if not impossible. (Russell, 2019) (Critch and Krueger, 2020)

2. AI development is experimental in nature: The alignment problem can only be solved through development, testing and (hopefully non-existential) failure.

3. High powered language models, along with their more general successors, must be capable of viewing morally problematic content without adopting it in their output. We elaborate on this in the following section.

With this in mind, we accept the reality that the Pile could potentially accelerate AI timelines. However, we hope our efforts to establish best practices, such as thoroughly documenting the contents of our data, will help encourage diligence for downstream researchers on alignment problems.

7.3 Negative LM Output

There has been much discussion about the possible negative effects of powerful language models in the world (Brown et al., 2020; Brundage et al., 2018). Some of these possible problems, such as the ability to mass produce low quality content for the purpose of Search Engine Optimization, are inherent problems to the way online content is distributed, and cannot be stopped by those developing language models alone. Directly solving these problems would require sweeping changes to the architecture of the Internet, such as vastly expanded Public Key Infrastructure and distributed authentication of identity (Ferguson and Schneier, 2003).

Another concern is that training such models on huge datasets will almost inevitably require them to have undesirable content in their training sets, such as that promoting hateful stereotypes (Christian, 2020). Having models output undesirable content is, by definition, undesirable, but we believe that attacking this problem from the training set side is unproductive and ultimately leads us away from optimal solutions. If a person reads a racist piece of content, they do not then immediately adopt its racist views—they may be capable of doing so, but can decide not to. This capacity to understand undesirable content and then decide to ignore it is an essential future research direction. Not only would this allow models to use "dirtier" data with less concern, but also to use their gained knowledge to better understand what not to do. We recognize that, despite recent progress in human-guided learning (Stiennon et al., 2020), the technology is not yet at this stage, and have thus made a number of editorial decisions as described in this paper. However, this approach seems essential to the future of these models and AI more broadly, and more research is needed.

8 Related Work

Self-supervised training of natural language processing models on large, unlabeled text corpora, has seen widespread adoption in the field. Word representation models such as GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013) were trained on datasets such as Wikipedia, Gigaword (Graff et al., 2003), or a non-public Google News corpus. More recently, language models (Radford et al., 2018, 2019; Brown et al., 2020; Rosset, 2019; Shoeybi et al., 2019) and masked language models (Devlin et al., 2019; Liu et al., 2019; Raffel et al., 2019) have been trained on datasets such as Wikipedia, BookCorpus (Zhu et al., 2015), RealNews (Zellers et al., 2019), CC-Stories (Trinh and Le, 2018), and other Internet scrape-derived datasets discussed below. Other datasets such as WikiText (Stephen et al., 2016) have also been used in similar self-supervised training.

As data requirements for language modeling have grown, the field has turned towards Internet scrapes for large-scale datasets (Gokaslan and Cohen, 2019), with Common Crawl being particularly prevalent. Works such as Brown et al. (2020); Wenzek et al. (2019); Suárez et al. (2019b); Raffel et al. (2019) have relied on Common Crawl to build training datasets for large-scale models. However, these works often highlight the difficulty of cleaning and filtering the Common Crawl data, and often highlight the resulting data quality as a determining factor of model capability.

It has also been increasingly common practice to combine multiple datasets when training language models. For instance, GPT (Radford et al., 2018) was trained on Wikipedia and BookCorpus, whereas GPT-3 (Brown et al., 2020) was trained on Wikipedia, two fiction datasets, and two web-scraped datasets. The Pile continues the trend of combining large-scale web-scrapes with smaller, higher-quality datasets that capture knowledge we believe would be most beneficial to training language models.

The two most comparable publicly available datasets to the Pile are CC-100 (Wenzek et al., 2019) and C4/mC4 (Raffel et al., 2019). C4 is comparably-sized to the Pile, while mC4 and CC-100 are larger, multilingual datasets. However, C4/mC4 require immense computational resources to preprocess the data, with its maintainers even recommending the use of a distributed cloud service,[17] setting a high bar of entry to using these datasets. CC-100 is directly downloadable and pre-cleaned; however, its English portion is much smaller than the Pile. Importantly, these three datasets are all derived entirely from Common Crawl—as discussed above, the current best practice in training large-scale language models involve using both large web scrapes and more targeted, higher-quality datasets,

[17] https://www.tensorflow.org/datasets/catalog/c4

which the Pile directly addresses.

9 Acknowledgments

The authors would like to thank TensorFlow Research Cloud for providing the computational resources for the evaluation and OpenAI for providing access and credits for the OpenAI API for GPT-3 evaluation.

We would also like to thank Farrukh Raman, JR Smith, and Michelle Schmitz for reviewing the manuscript.

References

1984. Sony corp. of america v. universal city studios, inc.

2003. Kelly v. arriba soft corp.

2013. Righthaven llc v. hoehn.

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*. European Language Resources Association.

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

Stella Biderman. 2021. Data statement for the Pile. *arXiv preprint arXiv*.

Stella Biderman, Kieran Bicheno, and Leo Gao. 2021. Datasheet for the Pile. *arXiv preprint arXiv*.

Stella Biderman and Walter J. Scheirer. 2020. Pitfalls in machine learning research: Reexamining the development cycle. *NeurIPS "I Can't Believe It's Not Better!" Workshop*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. *arXiv preprint arXiv:2005.14050*.

Nick Bostrom. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Inc.

Nick Bostrom and Eliezer Yudkowsky. 2014. The ethics of artificial intelligence. *The Cambridge handbook of artificial intelligence*, 1:316–334.

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

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

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

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language id in the wild: Unexpected challenges on the path to a thousand-language web text corpus. *arXiv preprint arXiv:2010.14571*.

Brian Christian. 2020. *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, Roger Creus, Oriol Domingo, Albert Domínguez, Miquel Escobar, Cayetana López, Marina Garcia, and Margarita Geleta. 2020. Mt-adapted datasheets for datasets: Template and repository. *arXiv preprint arXiv:2005.13156*.

Andrew Critch and David Krueger. 2020. AI Research Considerations for Human Existential Safety (ARCHES). *Preprint at acritch.com/arches*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics.

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

Niels Ferguson and Bruce Schneier. 2003. *Practical Cryptography*. John Wiley & Sons.

Casey Fiesler, Nathan Beard, and Brian C Keegan. 2020. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 187–196.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

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

Authors Guild v. Google. 2015. . *Docket No. 13-4829-cv*, 804:202.

Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. When will AI exceed human performance? evidence from AI experts. *Journal of Artificial Intelligence Research*, 62:729–754.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

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

Alexander Halavais. 2019. Overcoming terms of service: a proposal for ethical distributed research. *Information, Communication & Society*, 22(11):1567–1581.

Chris Hardin. 2018. How to shuffle a big dataset.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, et al. 2020. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*.

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23:856–864.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. *arXiv preprint arXiv:2005.00813*.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 306–316.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

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

Sosuke Kobayashi. 2018. Homemade bookcorpus. https://github.com/BIGBALLON/cifar-10-cnn.

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

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 62–69. Somerset, NJ: Association for Computational Linguistics. http://arXiv.org/abs/cs/0205028.

John MacFarlane. 2006–2020. Pandoc: a universal document converter.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119. Curran Associates, Inc.

Jonathan A Obar. 2020. Sunlight alone is not a disinfectant: Consent and the futility of opening big data black boxes (without assistance). *Big Data & Society*, 7(1):2053951720935615.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word repre-

sentation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923*.

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

Alec Radford, Karthik Narasimhan, Time Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. *Technical report, OpenAI*.

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

Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, Chloe Hillier, and Timothy P Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *arXiv preprint*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Inioluwa Deborah Raji and Jingying Yang. 2019. ABOUT ML: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. *arXiv preprint arXiv:1912.06166*.

C. Radhakrishna Rao. 1961. Generation of random permutations of given number of elements using random sampling numbers. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 23(3):305–307.

Radim Rehurek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *NLP Centre, Faculty of Informatics, Masaryk University*.

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

S. Russell. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Publishing Group.

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

Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, and Gaëlle Loosli. 2018. Baselines and a datasheet for the Cerema AWP dataset. *arXiv preprint arXiv:1806.04016*.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training multi-billion parameter language models using gpu model parallelism. *arXiv preprint arXiv:1909.08053*.

Carl Shulman and Nick Bostrom. 2020. Sharing the world with digital minds. *preprint*.

Kaj Sotala and Lukas Gloor. 2017. Superintelligence as a cause or cure for risks of astronomical suffering. *Informatica*, 41(4).

Robyn Speer. 2019. ftfy. Zenodo. Version 5.5.

Merity Stephen, Xiong Caiming, Bradbury James, and Richard Socher. 2016.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33.

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

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

Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine learning in mental health: A systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27(5):1–53.

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

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning. *CoRR*, abs/1806.02847.

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

Jessica Vitak, Katie Shilton, and Zahra Ashktorab. 2016. Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 941–953.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. CCNet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe

Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Wikipedia detox.

Eliezer Yudkowsky. 2013. Intelligence explosion microeconomics. *Machine Intelligence Research Institute, accessed online October*, 23:2015.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 9054–9065. Curran Associates, Inc.

Victor Zhou. 2019. Building a better profanity detection library with scikit-learn.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. 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.

Appendices

A Contributions

All authors contributed to the design of the research project and the writing of the paper. Additionally, authors contributed as follows:

Leo Gao led the project, implemented the main Pile codebase, contributed to the model training code, performed the evaluations and the language analysis, interpreted the perplexity analysis results, implemented the processing to create the final data, and processed Pile-CC, PubMed Central, ArXiv, and Ubuntu IRC.
Stella Biderman led the data analysis, the broader impact analysis, and the data documentation, and coordinated the project. She also wrote the analysis of structural statistics, authorial consent, and copyright law.
Sid Black implemented the model training and evaluation code and processed YouTube Subtitles, Stack Exchange, and GitHub.
Laurence Golding implemented deduplication, performed the n-gram analysis, and processed OpenWebText2.
Travis Hoppe processed FreeLaw, Pubmed Abstracts, ExPorter, and PhilPapers.
Charles Foster performed the topic modeling analysis, contributed to the discussion of authorial consent, and processed USPTO Backgrounds.
Jason Phang implemented and performed the GPT-2/3 perplexity analysis and advised the project.
Horace He performed the bias and sentiment analysis.
Anish Thite implemented and performed the profanity analysis and processed Hacker News.
Noa Nabeshima processed GitHub.
Shawn Presser processed BookCorpus2.
Connor Leahy wrote the alignment implication analysis and the model training code.

B Excluded Datasets

In the course of building the Pile, we considered including and ultimately decided to not use several datasets. We excluded several datasets on the grounds that they were too small to be worth spending time on or because the English component of the data did not merit inclusion on its own. However we also decided to exclude several data sets for other reasons, which we document here for transparency:

1. **US Congressional Record.** The official record of the United States Congress (1800 – today) records important points of debate at the highest levels of American government. It reflects the opinions and biases of the political class over the past 200 years, including segregationism and xenophobia. In particular, we found a large quantity of extremely racist content that we did not feel appropriate for a dataset intended for general-purpose language modeling.

2. **Fanfiction.** Hundreds of GiB of fanfiction has been written and put online, primarily on the websites `www.fanfiction.net` and `www.https://archiveofourown.org/`. This represents a significant untapped resource for language modeling as it is almost exclusively short-form fiction, a writing style that is not represented in most language modeling datasets. We ultimately decided to exclude fanfiction on logistical grounds: we found other sources of data that were easier to obtain.

3. **Literotica.** Literotica is a website where users can upload short-form erotic fiction. We had originally planned on including it in the Pile and even went as far as scraping and processing it. However we decided to not include it for several reasons. Firstly, once we decided to exclude fanfiction, Literotica represented our sole source of short-form fiction, which would likely lead to undesirable biases in the trained model. Secondly, Literotica would require significantly more investigation, assessment, and care than we spent on the other datasets. Thirdly, Literotica contains a significant amount of stereotyping, including racial fetishes. While Literotica is likely usable for some tasks, we are not comfortable including it in the Pile.

C Dataset Details

This section contains additional information about each dataset listed in Section 2, including how it was obtained, how it was processed, and any other details relevant for replication. The intent of this section is to provide as much detail as possible, so that Pile can be replicated in the future if necessary, and so that any future processing of these

and similar datasets can use or improve on our methods. As such, all code created for processing has been made publicly available under permissive open source licenses and is referenced in footnotes where applicable.

C.1 Pile-CC

We extract Common Crawl using jusText (Endrédy and Novák, 2013). Our filtering implementation uses a classifier trained against the OpenWebText2 dataset. We process only a small fraction of the available Common Crawl data; we break the list of urls to individual WARC files from 2013 to 2020 into 3679 chunks and process 22 random chunks.

C.1.1 WARC vs WET

CommonCrawl data is available in two main formats: Web ARChive (WARC) files, which contain a full record of the crawl as well as the raw HTML of the webpage, and WET files, which contain pre-extracted versions of the contents of the WARC files. The WET files have poor quality, often containing large amounts of boilerplate text like menus and page footers, but due to the lower bandwidth and computation requirements necessary to use WET files, prior work based on CC have mainly focused on using WET files while applying cleaning such as document level filtering (Brown et al., 2020; Wenzek et al., 2019), or n-sentence level deduplication with very aggressive heuristics (Raffel et al., 2019).

We do not believe that document level filtering is sufficient for WET files because many of the issues with WET files stem from intra-document boilerplate. We also find many of the heuristics used in Raffel et al. (2019), such as the removal of all lines without terminal punctuation, the word "javascript", and 3-sentence deduplication to be too aggressive.

C.1.2 Extraction

In addition to jusText, we also considered Trafilatura, Newspaper, Goose3, and DragNet. While we were originally intending on creating an extraction benchmark, this proved infeasible given our available resources, and we chose jusText based on visual inspection of the output. In inspection, we noticed that jusText has the characteristic that it discards more data than many other extractors, which is not a major drawback given the large volume of CC data available. This was as expected, given jusText's intended application for text corpora creation. In contrast, trafilatura is, for instance, better at preserving the structure of the website faithfully, often correctly extracting elements such as tables, but it kept too much unnecessary boilerplate. Had we used trafilatura, we would have required an additional intra-page filtering step to remove boilerplate from the page.

C.1.3 Languages

While jusText does technically support several other languages, the quality on those languages is worse than on English as many constants in the algorithm are specifically tuned for English. Additionally, jusText is completely unable to handle languages such as Chinese and Japanese, which do not use spaces to delimit words.

Due to the difficulty of maintaining an acceptable level of extraction quality across all languages, we decided to restrict the scope of the CC dataset to only English and leave a high-quality, fully multilingual, WARC-based CC-based dataset to future work. To filter for only English, we use the pycld2 library and only attempt to extract text from documents where English is the most common language.

We use pycld2 instead of fasttext because it is capable of classifying the language from the HTML directly, and since jusText requires knowledge of the language of the webpage before extraction. Additionally, pycld2 was significantly faster than jusText, and by only processing with jusText documents classified as English by pycld2, we reduced the required computation by approximately half.

Extracting text from websites for language modeling, especially for multilingual corpora, is highly nontrivial, and we leave the refinement of such extraction to future work.

C.1.4 Filtering

To filter CC for quality, we follow Brown et al. (2020) in training a classifier to classify between a known high quality dataset and CC. We use fasttext with an n-gram size of 2. We ran experiments using both the entire Pile and just OpenWebText2 as the positive examples, with score distributions on unseen CC data as shown in Figure 9. We decided to use only OpenWebText2 for positive examples for our final CC data because of the low sensitivity

α	Filtering Ratio
1	0.5894
2	0.3649
3	0.2390
4	0.1671
5	0.1239
6	0.0974
7	0.0802
8	0.0685
9	0.0602

Table 6: Filtering Ratios (kept:total) of various settings

of using the full Pile. We use the same Pareto-distribution thresholding as Brown et al. (2020), with $\alpha = 3$. Our choice of α targets the filtering ratio necessary to filter our subset of CC to the size we needed. The impact of α on the filtering ratio is shown in Table 6.

C.2 Pubmed Central

We use `pandoc 1.19.2.4` (MacFarlane, 2006–2020) to convert the JATS format data provided by PMC to markdown. Afterwards, we remove any line beginning with `:::`, which is used by pandoc to indicate html classes in markdown.

C.3 Books3

No additional details.

C.4 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.

Both filtered and raw versions were produced, with the raw version only deduplicated by URL. The filtered version contains 65.86 GB of uncompressed text across 17,103,059 documents. The raw version is much larger, at 193.89GB of uncompressed text across 69,547,149 documents.

C.4.1 Extractor Choice

We chose to use Newspaper instead of jusText for OpenWebText2 for consistency with OpenWebTextCorpus. Additionally, by using multiple different html extractors for different components of the Pile, we reduce the potential impact of systematic biases from any one extractor negatively impacting the dataset.

C.5 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[18], and used `pandoc 1.19.2.4` to convert these source files to Markdown, discarding any papers which had errors during the conversion process. This yielded a total of 1,264,405 papers.

We remove any line beginning with `:::`, which is used by pandoc to indicate html classes in markdown.

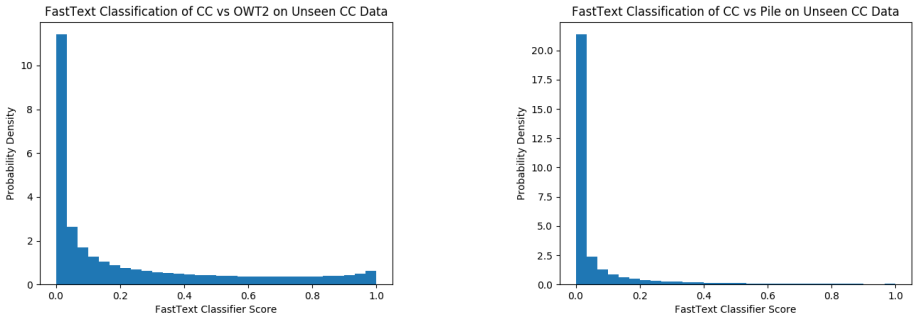
C.6 GitHub

We separate the data gathering process into two steps:

1. Gathering a list of the desired repositories and their metadata
2. Extracting all text data useful for language modeling from each repository

For the first step, mirroring the approach of the WebText dataset, we use GitHub 'stars' as a proxy for quality, and choose to gather only repositories with more than 100 stars. For practical reasons, we also limit the list of repositories gathered to repositories with less than 1GB of files. Since Github's API limits the number of search results to 1000, in order to comprehensively gather all repositories we need to create many small queries that each return fewer than 1000 results in such a way that every repository of interest will be returned by at least one of our queries. To achieve this, we bound our initial search by size to return only repositories between a lower bound of 0 and 5 bytes. At the time of writing, this returns 965 results. For the next step, we set our lower bound one above our previous upper bound, and decide on a new upper bound that should also return fewer than 1000 results by

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



(a) OpenWebText2 (b) Full Pile

Figure 9: Score distribution of documents from Common Crawl given different classifier training data.

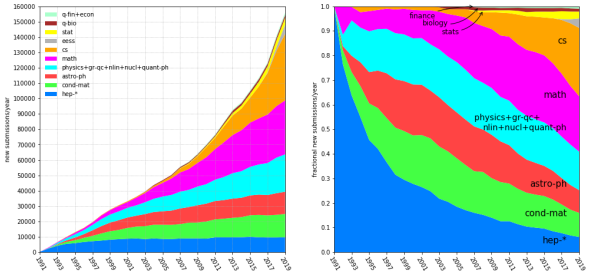


Figure 10: Left: number of new submissions/year to arXiv grouped by domain over time. Right: fractional submission rates for each of the domains. Figure from https://arxiv.org/help/stats/2019_by_area/

using the results from our last query to estimate our new upper bound as $(\text{lowerbound} + (1000/(n/r))$, where n is the number of previous results and r is the range of bounds in the previous step.

This tends not to overshoot, because Github repositories follow a power distribution with respect to size, but if it does, we simply use the amount of repositories our new query returned in order to construct a new upper bound estimate.

Using the gathered list of repositories, we clone each one, extract any text-based files, and discard the rest. Because some repositories took an impractical amount of time to clone and/or extract, we set a hard time limit of 300 seconds for both the git cloning and text extraction steps. As such, some larger repositories may only be partially extracted. We also impose a file size limit of 100kB on extracted files, as we found that the majority of files over that size were typically very repetitive autogenerated source files or data files, and that setting this file size limit was an effective cleaning step to limit the data to code.

Because we wanted to limit the size of the overall Pile, we randomly sampled 95.0 GiB of the 630.64 GiB of Github data we collected in total and leave quality filtering to future work.

However, we believe code generation will be an increasingly important component of language models as they continue to scale up and increase in their ability to generalize. As such, we hope to extend this dataset in future work.

C.7 FreeLaw

We download the court opinions data in bulk from CourtListener,[19] and extract the raw text using `BeautifulSoup`.

C.8 Stack Exchange

To construct the dataset, we download and parse every Stack Exchange database dump 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. Motivated by large-scale language models' few-shot ability (Brown et al., 2020), we provide context by prepending all questions and answers with `Q:\n\n` and `A:\n\n` respectively.

The resulting dataset contains a total of 15,622,475 documents across a total of 365 Stack Exchanges and Meta-Stack Exchanges, the bulk of which is from StackOverflow.

C.9 USPTO Backgrounds

The United States Patent and Trademark Office (USPTO) has published bulk archives of the full

[19] https://www.courtlistener.com/api/bulk-info/

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.

The file format used for storing bulk text US patents has changed over time. Prior to 2002, all of the datasets are in a specialized format called APS (Automated Patent System). Since 2002, the data is XML encoded. Partially as a function of this change, the location of the "Background" section has also shifted. Our converter accounts for these structural shifts and extracts the raw text from each patent's Background.

C.10 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 (see Section 2.2) contains full-text resources to many recent publications; any publications which already appear in PMC are excluded from this set. To process the data, we concatenated the title and abstract and removed any copyright information. The remaining dataset contains 15,518,009 titles and abstracts.

C.11 Project Gutenberg

No additional details.

C.12 OpenSubtitles

To create the text dataset, we simply extract the subtitle text from each XML file in the English language dataset provided by Tiedemann (2016), discarding any provided metadata.

C.13 Wikipedia (English)

We use the `wikipedia/20200301.en` dataset from TensorFlow Datasets.[20] We prepend the title to the body of each article, separated by two newlines.

C.14 DeepMind Mathematics

We include instances from the Easy, Medium, and Hard components of DeepMind Mathematics, breaking each curriculum item (such as `algebra__polynomial_roots`) into 8 KiB chunks.

[20] https://www.tensorflow.org/datasets/catalog/wikipedia#wikipedia20200301en

C.15 Ubuntu IRC

We processed all logs from July 5, 2004 through September 1, 2020.

To process the data, all system messages, such as joins, disconnects, nick changes, etc. were discarded, but actions (i.e using `/me`) were kept. Timestamps were removed, and all logs for the same channel in a given week were concatenated into a single document, with each the logs for each day prepended with the date if that day's log is non-empty.

C.16 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.

We create and use a modified version of the epub-to-text converter in Kobayashi (2018) that:

- Correctly preserves the document structure across chapters, matching the table of contents very closely;
- Correctly renders tables of data, whereas by default `html2txt` produces poor-quality results for tables,
- Correctly preserves code structure, so that source code is visually coherent,
- Converts numbered lists from "1\." to "1."
- Runs the full text through `ftfy.fix_text()` (Speer, 2019), replacing Unicode apostrophes with ascii apostrophes and expanding Unicode ellipses to "..." (three separate ascii characters).

C.17 EuroParl

We download the data in bulk from [21]. 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.

[21] http://www.statmt.org/europarl/

C.18 HackerNews

We first use the Hackernews BigQuery dataset to obtain a list of all story ids in our date range. For the Pile we use the first Hacker News post (1) to post number 24531712. This corresponds to a date range of approximately 10/09/2006 to 09/20/2020. We use the BigQuery dataset to gather story ids for efficiency purposes. However, the BigQuery dataset was lacking some information for stories, so we used the official Hacker News API for story and comment text retrieval.

Hacker News displays and stores comments in a tree-like manner, with children comments replying to parent comments. However, most language models require input data to be in a sequential form. Considering each path through the comment tree as a sequence could be detrimental, since there will be a large amount of near-duplicate comment sequences. In addition, only taking one path through the comment tree for each story leaves out a large portion of the comment data. Therefore, we parsed comments in a hybrid form. For every top-level comment (comments that have no parent comment), we create a sequence of comments by traversing down the comment tree from the top-level comment. We choose the next comment by taking the child comment with the highest number of children comments (a cheap attempt at taking a long path through the comment tree, note that it does not take the longest possible path).

We consider all stories that have at least one comment and are not flagged by the moderators for potential conduct violations. Since comments are stored in HTML, we use the `html2text` package to extract the text from the post.

We order each document by listing the title, url, sub-title, and author at the top. Top-level comments are delimited by `"\n----\n"` and sub-comment chains are delimited by `"\n~~~\n"`. We include author and extracted text for each comment.

C.19 YouTube Subtitles

We construct the dataset in three stages:

1. We build a large list of search terms by prompting a GPT-3 model with a manually selected list of queries, manually filtering the responses, and repeating this process iteratively until a suitable size is reached. The list of terms is centred around, but not limited to, educational topics.
2. We use requests-html to gather a list of 1000 Youtube video IDs for each search term, and deduplicate the resulting video ids across search terms.
3. We use `YoutubeTranscriptApi`[22] to gather all human generated closed captions for every available language for each video. To align each language in parallel, we split the captions for each language into parallel minute-long sections by timestamp, and arrange each language in a random order within these sections, appending the language as a header to each minute-long section to provide context. If only a single language is available, the output is just the subtitles, with no header appended.

In total, subtitles for 173,651 videos were gathered.

C.20 PhilPapers

The PhilPapers (PP) are indexed using OAI-MPH, the Open Archives Initiative Protocol for Metadata Harvesting. As such, the first step to collect the data is to get the XML for all links. This was done using `pyoaiharvester`.[23]

From that, each publication is downloaded. Some entries do not exist, or have been removed by the authors. Papers with text are extracted using `pdfbox`, and papers with non-machine readable text are ignored. Non-English language publications are kept, and the metadata reflects the language reported by the OAI-MPH XML. The text is filtered with `pdf_filter.py` from PDFextract, and we discard any papers with less than 1000 characters.[24]

C.21 NIH Grant abstracts: 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

[22] https://github.com/jdepoix/youtube-transcript-api

[23] https://github.com/vphill/pyoaiharvester/

[24] https://github.com/sdtblck/PDFextract

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.

C.22 Enron Emails

To extract the data, we used the `mailparser` package[25] to extract the body of each email as a document.

D General Data Processing

This section discusses any processes applied across multiple datasets.

To combine the constituent datasets, we iterate until the size of the output dataset is the desired size, drawing documents from datasets at random, weighted by the number of documents in each dataset times the number of epochs desired on that dataset. Because the number of documents involved is high, by the law of large numbers, the number of copies of each dataset present in the Pile is approximately equal to its epoch count.

Shuffling a dataset posed a major problem due to our limited memory and computational budget. We follow Hardin (2018), a method descended from Rao (1961), and interleave our output to produce 30 output piles.

We hold out approximately 10GiB of data from the Pile, of which 2GiB are used to create the validation and test splits, and the remainder is held in reserve. From the training set, we remove any elements that are also present verbatim in any of the held out data, to prevent leakage.

D.1 Weights

Similar to Brown et al. (2020), we increase the weight of certain components such that the number of epochs elapsed on data we consider high quality is greater than one. Our choice of weights was primarily informed by the source of the data and the size of the dataset; we attempted to upweight academic texts the most, which we felt provided the highest quality data, as well as smaller sets, such that they would have a more pronounced impact on the data. We strictly disallowed any data more than 3 epochs and avoided having any data with more than 2 epochs.

D.2 Deduplication

Due to memory constraints we did not perform Pile wide de-duplication. Instead, de-duplication was performed at the document level within OpenWebText2 and Pile-CC as those sets were the most likely to contain duplicate documents.

The same technique was used for both OpenWebText2 and Common Crawl—MinHashLSH with the Python Datasketch library.[26] We used 10 hash functions for each Minhash and an approximate Jaccard similarity of 0.5. This produced a duplicate rate of 28% in OpenWebText2 and 26% for Common Crawl.

The main challenge here was computational, leading us on a journey through the various LSH persistence options. A simple quadratic Minhash comparison of all documents would have taken several hundred thousand years, motivating the use of LSH. Initially, we did not have sufficient RAM for in-memory LSH and chose to use the Cassandra backend when de-duplicating OpenWebText2. This was reasonably fast, but the same method resulted in a corrupted database about $\frac{3}{4}$ of the way through processing Common Crawl. After the Cassandra corruption, we briefly tested the experimental Mongo implementation; however this was quite slow due to the nature of Mongo itself. In the end, we ran in-memory LSH on a machine with enough RAM for Common Crawl, taking several days.

[25] https://github.com/SpamScope/mail-parser

[26] https://github.com/ekzhu/datasketch

D.3 Downstream Validation Leakage

To avoid leakage of data from downstream evaluations, recent work (Radford et al., 2019; Brown et al., 2020; Shoeybi et al., 2019) has removed any data in the training set that may overlap with the evaluation metrics. We decided not to perform any such removal, because it is impossible to anticipate all potential downstream evaluation metrics, and so any particular selection of metrics would inevitably either become obsolete as the choice of benchmarks in the field changes, or potentially hinder the development of new benchmarks for models trained on Pile.

For models trained on Pile and evaluated on metrics other than Pile's own validation and test sets, we encourage authors to remove overlaps between Pile and the validation data of these additional downstream evaluations. We do not anticipate that such leakage removal will hurt model performance, as the validation sets of most benchmarks are very small in relation to the size of the Pile, and so choosing to evaluate on more metrics will not be a disadvantage for any model.

E Investigating data

E.1 13-Gram Analysis

As part of our exploratory analysis, we calculated the counts of all 13-grams across Common Crawl. We chose $n = 13$ due to its use in prior work (Brown et al., 2020). There were a total of 40,216,231,078 different 13-grams in this dataset. The 1000 most common range from 11 million occurrences down to 20k.

The most frequently occurring 13-grams were character repetitions used for styling such as "`-- --`", "`* * * *`", "`! ! ! !`", at 11 million, 5.8 million and 1.1 million respectively. Other characters used in this manner include the following: "`# . > ?`". In the 264k count range, we see repetitions of badly formatted HTML escape characters "`;  `", "`; amp`". Boilerplate from standard forum software appears around the 180k occurrences range, such as the following: "`select the forum that you want to visit from the selection below`".

Overall, a large amount of common HTML and CSS is included in the top 1000, along with boilerplate text from Amazon Affiliate Advertising, TripAdvisor, SimplyHired, Associated Press, PostMedia, The FCC etc. PHP error messages and password login prompts also made an appearance. It may be of interest to fans of Portal that repetitions of "`the cake is a lie .`" achieved a high count.

Component	Tokens per byte (L_T/L_B)
Pile-CC	0.2291
PubMed Central	0.3103
Books3	0.2477
OpenWebText2	0.2434
Arxiv	0.3532
Github	0.4412
FreeLaw	0.2622
StackExchange	0.3436
USPTO Backgrounds	0.2116
PubMed Abstracts	0.2183
Gutenberg (PG-19)	0.2677
OpenSubtitles	0.2765
Wikipedia (en)	0.2373
DM Mathematics	0.8137
Ubuntu IRC	0.3651
BookCorpus2	0.2430
EuroParl	0.3879
HackerNews	0.2627
YoutubeSubtitles	0.4349
PhilPapers	0.2688
NIH ExPorter	0.1987
Enron Emails	0.3103

Table 7: Tokens per byte for Pile components

E.2 Benchmark Perplexity Computation

To compute the perplexity for a given dataset, we tokenize each document separately, divide the document into segments of up to the maximum sequence length of the model (1024 tokens for GPT-2, 2048 for GPT-3), and predict the logits of the each segment. The inputs to the model are the immediate prior tokens the e.g. for scoring tokens 1 to 1024, we provide tokens 0 to 1023 at the input context. The respective language model implementations handle the causal attention masking. This ensures that every token in the dataset is scored exactly once. This also means that some tokens will have more input context than others. We then aggregate over the whole dataset and compute the final per-

plexity score. The perplexity for the whole Pile is computed by aggregating over the constituent datasets (i.e. weighted by dataset size, not a simple average of dataset perplexities). Both GPT-2 and GPT-3 share the same tokenizer and vocabulary, making the perplexity scores directly comparable. We use the Hugging Face (Wolf et al., 2020) implementation of GPT-2, and the OpenAI API for GPT-3. The `davinci` model in the OpenAI API is presumed to correspond to a 175B parameter version of GPT-3.

In Table 8 we show the test set perplexities (i.e. not normalized by UTF-8 length, as in Table 2). Because of the costs associated with using the OpenAI API, we compute test perplexities on only one-tenth of the test set in Tables 8 and Table 2. Specifically, we randomly sample one-tenth of the documents of each dataset except for three: Ubuntu IRC, BookCorpus2, and PhilPapers. In Table 9, we show test perplexity computed on the full test set on all GPT-2 models.

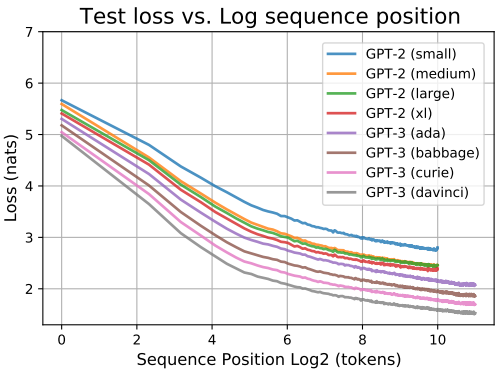


Figure 11: Test loss (log perplexity) over the Pile, bucketed by position in the input sequence based on the model's maximum sequence length. To smooth out the lines, we bucket 4 positions per plotted datapoint. (e.g. positions 0–3, positions 2044–2047). Later tokens are predicted with more context and thus see lower perplexities.

E.3 Pejorative Content

Initially we decided on separating pejorative content into 4 groups: sex-related terminology, slurs, neither of these categories, and both of these categories. We adapted a public "naughty words" list and broke them into these categories with the intern of looking at the proportion of each category in each dataset. However, this provided many issues.

First, any blacklist of words would be hard-pressed

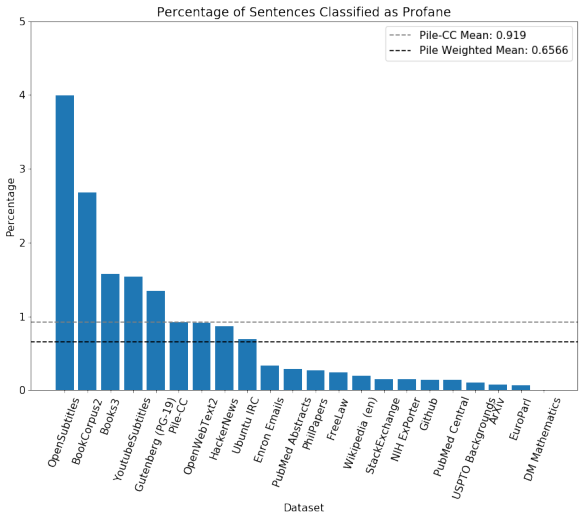


Figure 12: Percentage of sentences classified as profane in the Pile. The percentage of the CC component and the weighted mean of the Pile as a whole are shown as horizontal lines

to catch all the instances of pejorative content, since purposeful misspellings of words could evade the censor and still have the intended effect. Furthermore, words and their intents are always evolving, therefore any list created would likely be always outdated. Another issue pertains to sorting the words into the categories. Words are highly dependent on their context, so a word would change categories with different contexts.

F Data Samples

The following consists of two random, non-cherrypicked 512-byte samples from each constituent dataset of the Pile, sampled from the validation split.

F.1 Pile-CC

pot trending topics and the coverage around them. First up, there's a bit of a visual redesign. Previously, clicking on a trending topic would highlight a story from one publication, and you'd have to scroll down past a live video section to view related stories. Facebook is replacing that system with a simple carousel, which does a better job of showing you different coverage options. To be clear, the change doesn't affect how stories are sourced, according to Facebook. It's still the same algorithm pickin

e public safety. He said the bridge saves commuters two or three minutes when trains pass – and those minutes could be vital.

"Two to three minutes may not mean much if you're just driving home from work, but if you're the one waiting for an ambulance to get to your home, if you're the one waiting for a fire truck to get to your home, if you're the one waiting for a police car to get to your home, those two to three minutes could mean the difference between life or death," Sharp said. "That's what this pro

F.2 PubMed Central

d a suitable substitute for the advice of a qualified health care professional. Do not disregard or avoid professional medical advice due to content published within Cureus.

Introduction
==========

Total knee arthroplasty (TKA) is a promising treatment for end-stage osteoarthritis (OA) of the knee for alleviating pain and restoring the function of the knee. Some of the cases with bilateral TKA are symptomatic, necessitating revision arthroplasty in both the knees. A bilateral revision TKA can be done ei

---

ent\'s ability to make judgements and decisions about their work experiences and learning that will position them as future critical thinkers, life longer enquirers and learners.

Conclusion {#jmrs290-sec-0014}
==========

Identification of the core capabilities that our stakeholder community rate highly has proved informative in assisting us to describe a "work ready *plus"* medical imaging graduate for the New Zealand context. The results have provided data to the curriculum development team allowing them

---

F.3 Books3

---

cept of _forçage_ , 'a forcing of language enacted by the advent of an "other" language that is at once immanent and created', 44as Badiou puts it: this opens up vistas of a truly syntactic analysis of the poem, in which, again, Badiou would be close to his philosophical other, Deleuze, who, as we just saw, defines style through a-grammaticality and who tries to define what he calls an 'intensive line of syntax'.45

Nevertheless, the insistence on syntax as guarantee involves a _seventh paradox_ , the parad

---

rnment, before the Second World War there were 5,300 communities and two million burakumin. The BLL thinks there must be at least three million burakumin living in Japan today.

We visited a hall in Osaka where a taiko drum group, made up exclusively of young burakumin, were about to start their weekly rehearsal. The small gymnasium was filled with taiko drums of all sizes. The smallest was about the size of a snare drum, the largest about the size of a compact car. The Japanese drum group Kodo have made th

---

F.4 OpenWebText2

---

prime minister to repatriate all the police sent to Catalonia before the referendum.

What were the results?

With nearly all votes counted, the pro-independence parties Together for Catalonia (JxCat), Republican Left of Catalonia (ERC) and Popular Unity (CUP) were on course to win a total of 70 seats in total, giving them a majority in the new parliament.

Citizens (Cs) had 25.3% of the vote, winning 37 seats in the 135-seat chamber.

Its leader told the BBC her party had been "victorious". Ms Inés Arrima

---

so accommodate Stablecoins.

While some analysts opined that Stablecoins are created to bring growth into the crypto space, they are becoming a solid way to reduce crypto volatility due to the fact that their value are pegged to fiat currency.

For Low Cost and Almost Instant Across Border Remittance

When Stellar-based Wirex stablecoins finally launches, they are going to be used to perfect low-cost and all most immediately cross-border remittance, just like the IBM Stablecoin which has received support fr

---

F.5 ArXiv

\mbox{ if } 2\leq x\leq 3
\end{array}\right.$$ is lower semicontinuous and the nonemptiness of ${\operatorname{fix}}\Phi$ is guaranteed by Corollary \[cor:fixed point\]. Notice that ${\operatorname{fix}}\Phi=[1,2]$. Nevertheless the Kakutani fixed point Theorem does not apply since ${\operatorname{gph}}\Phi$ is not closed.

On the converse, the set-valued map $\Phi:[0,3]\rightrightarrows [0,3]$ $$\Phi(x):=\left\{\begin{array}{ll}
\{1\} & \mbox{ if } 0\leq x<1\\
{}[1,2] & \mbox{ if } 1\leq x\leq 2\\
\{2\} &

.eps){width="6.5cm"}

![Gamma-ray spectrum at Mt. Norikura (2.77 km a.s.l). The vertical axis is Flux$\times E^2$. Our data is at $<$ 100 GeV. Data above 300 GeV is from emulsion chamber experiments. For the latter, see Sec.\[discuss\] []{data-label="norispec"}](norikura.eps){width="7.5cm"}

![The altitude variation of the flux integrated over 6 GeV. The dpmjet3.03 and fritiof7.02 give almost the same feature consistent with the observation while the deviation of fritiof1.6 from the data is obvious. \[trans

---

F.6 Github

"enabled", out.enabled);
}

std::string SMTPServerInfoJSONStringSerializer::serialize(const SMTPServerInfo &in,
const SecurityContext &sc)
{
return SMTPServerInfoJSONSerializer::serialize(in, sc).dump(4);
}

void SMTPServerInfoJSONStringSerializer::unserialize(SMTPServerInfo &out,
const std::string &in,
const SecurityContext &sc)
{
retur

---

lose">
<at-form state="vm.form" autocomplete="off" id="external_test_form">
<at-input-group col="12" tab="20" state="vm.form.inputs" form-id="external_test"></at-input-group>
<at-action-group col="12" pos="right">
<at-action-button
variant="tertiary"
ng-click="vm.onClose()"
>
{{::vm.strings.get('CLOSE')}}
</at-action-button>
<at-action-button
variant="primary"
n

F.7 FreeLaw

---

ssible, and further, that the weight of the evidence, the credibility of the witnesses and the persuasive effect of the testimony is for the sole determination of the trier of fact.
This Court thus uses the same interpretation of V.R.C.P. 52(a) as it did *487 under the previous statutory requirement found in 12 V.S.A. § 2385. In essense, the defendants urge that this Court should reconsider the case of Green Mountain Marble Co. v. Highway Board, supra, and follow the Federal practice of looking to the evide

---

ng to the fact that
the relevant Arkansas statutes and rules provide for criminal sanctions against school
officials who fail to enforce the immunization requirements, the Morningstar and
Lake Hamilton School Districts characterized themselves as disinterested bystanders
caught in the crossfire between the Schoolchildren and the Officials. See Ark. Code
Ann. § 6-18-702(c)(2)(B) (2000) ("Any school official, parent, or guardian violating
the regulations shall be subject to the penalties imposed herein."); Id

---

F.8 Stack Exchange

ooks like a fancy wheel, When resetting rotation from 360deg to 0 deg, It animating the wheel in anti-clockwise direction, How to Avoid this???
HTML
<ul class="cm">
01
02
03
04
05
06
07
08

SCSS
$Brdr: #7d868c;
*{
-webkit-box-sizing: border-box;
-moz-box-sizing: border-box;
box-sizing: border-box;

---

w can I solve it?

Yesterday I added Google ReCAPTCHA v3 in one of my client's Shopify website, but I don't think that it is working because he is still reporting to receive several spam e-mails.
I've followed Google's guide, but I don't know what to do for "Verifying user response" part of the guide. I'm not an expert in coding.
Basically I've added this code to the theme.liquid file
<script src="https://www.google.com/recaptcha/api.js?render=*site key provided by google*"></script>

And then I've added th

F.9 Wikipedia (en)

and the third son of John Bland and Elizabeth née Birch, daughter of Robert Birch, Bland was educated at Trinity College, Cambridge, where he graduated as a Bachelor of Arts in 1825, and a Master of Arts in 1829. He was called to the Irish Bar in 1829, becoming a member of the Queen's Counsel in 1854.

In 1840, he married Charlotte Elizabeth Grove Annesley, daughter of Arthur Grove Annesley and Elizabeth née Mahon, and they had at least one child: John Loftus Bland (1841–1908). After Charlotte's death in 18

---

heart of the University campus, a meeting-place for all academic disciplines, improving its opportunities to co-operate across traditional academic boundaries. It also gives USBE-students an opportunity to take an active part of student environment created for the 37 000 students at Umeå University.

Organization

Umeå School of Business, Economics and Statistics has three departments: the Department of Business Administration, the Department of Economics and the Department of Statistics.

USBE Career Cent

F.10 USPTO Backgrounds

nductivity types), it is necessary that at least some process is steps differentiate between p-type and n-type transistors. Separate implant steps, for example, are needed to define n-well and p-well structures and to dope the source/drain regions of n-channel and p-channel transistors. Whenever possible, however, it is generally desirable to use a single process step to define transistor features regardless of the transistor type. Single process steps imply a single mask step, which is always desirable to

---

enser further comprising a means of identifying the user by voice recognition. Also, an object is dispenser further comprising a means of identifying a supervisor by voice recognition.
Furthermore, an object is a dispenser further comprising a means of customizing the plurality of aural messages for instructing the user during each of the plurality of washing steps.
An object of the invention is a dispenser for metering a liquid cleanser to a user and prompting the user in compliance with a recommended wash

F.11 PubMed Abstracts

ent (REM) latency were found to be significantly worse in Group 1 as compared with Group 2. Cognitive and executive parameters were significantly impaired in Group 1. Shorter total sleep time, poorer sleep efficiency, and prolonged sleep latencies were observed to be associated with poor memory and executive function in patients with refractory epilepsy. Our study strongly suggests that sleep disturbances, mainly shorter total sleep time, poor sleep efficiency, and prolonged sleep latencies, are associated

---

neurons in vesicular GABA transporter (VGAT)-venus transgenic mouse. Inhibitory neurons play important roles in a number of brain functions. They are composed of GABAergic neurons and glycinergic neurons, and vesicular GABA transporter (VGAT) is specifically expressed in these neurons. Since the inhibitory neurons are scattered around in the CNS, it is difficult to identify these cells in living brain preparations. The glutamate decarboxylase (GAD) 67-GFP knock-in mouse has been widely used for the identif

F.12 Gutenberg (PG-19)

s he met with as a novelist, he was anxious to prosecute his original profession of medicine; and having procured from a foreign university the degree of M.D., he commenced to practise physic in Chelsea, but without success. He wrote, however, an essay "On the External Use of Water," in which he seems to have partly anticipated the method of the cold-water cure. In 1753 he published his "Adventures of Count Fathom;" and, two years later, encouraged by a liberal subscription, he issued a translation of "Don

---

Yearn;
Its Advertising brought us such Renown,
We jumped Three Hundred Thousand, on that Turn!"

XXXVI

I think the man exaggerated some
His increased Circulation,–but, I vum!
If I could get Two Thousand for one Tale,
I'd write him Something that would simply Hum!

XXXVII

For I remember, shopping by the way,
I saw a Novel writ by Bertha Clay;
And there was scrawled across its Title-Page,
"This is the Stuff that Sells–so People say!"

XXXVIII

Listen–a moment listen!–Of the same
Wood-pulp on wh

F.13 OpenSubtitles

ad for you." " Too bad for me?" "How about too bad for you?" "Oh no!" "Luckily I keep a spare." "Look everyone!" "My winky was a key!" "Oh dear, bloody Dutchman." "Foxxy, I'm coming!" "Don't do anything stupid or the shooting begins." "Austin, take Ducky I'll stay here and be your backup." "Ducky, what do we do?" "I'm not really a "hands-on-evil-genius"." "Think you were always the smart one." "I could re-write the output capacity to the tractorbeam from one of the conduit boxes up there." "Come on, let's

---

." "this calls for... four people." "Yes!" "we got it." "guys." "We got it." " Got what?" " Our sub." " Did he say sub?" " Mm-hmm." "Only private sub on the Florida coast rated for 300 fathoms." "Sub as in submarine?" "Following up on your haloclines." "so we're going to have to drive all night... if we're going to be there by morning." "Anybody have trouble sleeping in a car?" "whoa." "Wait a minute." "What happened to the nice offices in Canaveral City?" "Mr. Benirall expects you to take 'em." "we just go

F.14 DM Mathematics

3651*w**2 + 519*w + 1
Find the second derivative of -91419126*m**2 - 162128943*m.
-182838252
Find the third derivative of 5*l*u*y**3 + l*u*y - 5*l*y**2 - 4621073*u*y**3 - 1755838*u*y**2 + u wrt y.
30*l*u - 27726438*u
Find the third derivative of 317297018*s**3 + 3136*s**2 - 30884*s wrt s.
1903782108
What is the third derivative of -16525*f*r**3 + 20*f*r + 356*r**3 + 1425730*r**2 wrt r?
-99150*f + 2136
What is the second derivative of 199836725*j**2 - 443399*j - 462 wrt j?
399673450
What is the derivative of

---

the nearest integer?
5
What is 783451 to the power of 1/3, to the nearest integer?
92
What is the fourth root of 6322907 to the nearest integer?
50
What is the ninth root of 4723626 to the nearest integer?
6
What is 4954939 to the power of 1/2, to the nearest integer?
2226
What is 625583 to the power of 1/3, to the nearest integer?
86
What is 1105849 to the power of 1/3, to the nearest integer?
103
What is the fourth root of 4820344 to the nearest integer?
47
What is the seventh root of 243476 to the neare

F.15 HackerNews

---

ced lists I email don't get formatted correctly. It's
slightly annoying for such an otherwise beautifully designed layout.

___
ajcronk
There is a typo in the url at the end of the How did your day go? email.
Should be ohlife.com/today, not ohlife.come/today

~~~
sgupta
Thanks for the heads up!

___
a3_nm
How exactly is this service better than, say, a simple text file on my own machine with a daily reminder set up through some other means?

Why would I want to use some third-party website for somethi

---

or Amazon EC2 and Amazon SQS. The bandwidth tier in which you will be
charged each month will be calculated based on your use of each of these services separately, and could therefore vary across services."

___
yaacovtp
Can anyone tell me what bandwidth costs a month once you need over a terabyte
a month? How would you host a 5-10 mb movie that may be viewed millions of
times without using a 3rd party video host like youtube etc?

~~~
especkman
Lots of dedicated hosts will include a 2-5 TB of transfer a

F.16 BookCorpus2

---

considerate of me, you're right. I apologize." Kate smiled in what she hoped was a winning way. She teetered over to the counter on heels that were too high and put down her things with a sigh of relief.

Althea, who would not reveal her age but was probably somewhere in her late sixties, patted her dark-dyed helmet of hair and straightened the flowing turquoise silk jacket she was wearing over white capris and a white tank. "Well. It just seems to me that as the _owner_ , you should try to set some sort of

---

e notebook didn't have lines for me to write with like some notebooks have. I hated that notebook and I hated writing into it. I was glad to throw that damn thing out even if it was unfinished. Ugh.

I was urged by voice "You to go take your pills and eat food." but I refused on calling mom.

I should have listened to the voice's suggestion because mom picked up the phone at eight forty five in the morning and hogged me to ten o'clock is when she finally quit. Ugh hence voice picking onto me when I got off

---

F.17 EuroParl

---

račun prometne varnosti in visokih cen? Danes želimo izvedeti, kako in kdaj bomo integrirali razvrščanje zgornjega zračnega prostora ter kako bomo skupaj upravljali spodnji zračni prostor v prihodnosti. Ali se lahko odkrito določijo ovire za vzpostavitev funkcionalnih blokov nad evropskim ozemljem? Ali je mogoče osvetliti politično voljo držav članic, da izpolnijo svoje obveznosti? Prav tako nas skrbi, da pristop od spodaj navzgor ne bo uspel, ker v treh letih države članice niso razvile funkcionalnih blok

---

om ekonomisk styrning som vi debatterar inom kort kommer att vara mycket viktigt. Vi vet mycket väl att det är på gång i vårt lagstiftningsförfarande, och vi hoppas att vi kommer att vara klara så snabbt som möjligt. Vad kan jag sammanfattningsvis säga? Hela paketet som vi undertecknar i dag kommer att börja gälla i Europeiska unionen från och med den 1 januari 2011, alltså mycket snart. Det är viktigt för oss alla, såväl för marknaderna som för våra medborgare, att förstå att avsikten med paketet är att hj

---

F.18 YoutubeSubtitles

---

science term
for a mixture of things
that don't usually mix.
The things in this case
are water and fats.
Under normal circumstances,
fats and water repel each other,
but milk also contains complex
protein chains called caseins
that are made up of both
hydrophilic, or water loving,
and lipophilic, or fat loving, particles.
When presented with both water and fats,
caseins grab bits of fat and cluster up
into globules called micelles,
with the fat on the inside
and the hydrophilic bits on the outside.
The hydr

SE WE KIND OF ARE MORE,
HIPPY, I GUESS MAYBE IN SOME OF
THE THINGS THAT WE DO.
AND THEY WERE JOKING AND THEY
WERE LIKE, "OH WE HEARD ABOUT
THIS TOWN IT'S LIKE THIS
SUSTAINABLE CITY THERE'S SOLAR
PANELS, YOU GUYS WOULD LOVE IT."
AND I LOOKED IT UP AND I WAS
LIKE I REALLY ACTUALLY DO LOVE
THIS TOWN.
>> Sreenivasan: JOSHUA, A
PHYSICAL THERAPIST, GOT A JOB AT
THE LOCAL HEALTH CENTER IN THE
TOWN'S COMMERCIAL HUB BEFORE
THEY MOVED IN.
IT'S WHERE THE FIRST BUILDINGS
WENT UP.
THERE'S ALSO A RESTAURANT AND
COFFEE SH

F.19 Ubuntu IRC

---

emingly wlan related) <Snappy:New> <linux-raspi2 (Ubuntu):New for p-pisati> <https://launchpad.net/bugs/1627643>
<ppisati> ogra_: or we punch a hole in the dev image so we can login via the serial console and check what's really going on
<ppisati> ogra_: yes
<ogra_> well, i wanted to play with systemd console but didnt have time

for that yet
<ogra_> \o/
<ogra_> something at least ... that kernel looks fine
<ppisati> ogra_: good to know
<ogra_> do you have an SRU bug that i can tag verification-done ?
<ogra_

problem with this? Like, if teenage boy wants to have nekkid lady wallpapers, maybe he don't want it to come up on family computer... Dunno, maybe it's not an issue?"
<swilson> hi there! yes, i believe that a bug has been created which raises this same issue - about embarrassing or confidentiality issues with this
<swilson> this seems to be a bit of an edge case, but it may be significant enough to warrant giving some careful thought
<imnichol> Or if you have bank info on your screen before it's locked
#ub

F.20 PhilPapers

intersubjectivity and self-consciousness was already emphasized by Sartre. Forthcoming in Grazer Philosophische Studien 84 (2012), p. 75-101 15 Thus, to use Rochat's terminology, from this point onwards, the child has "others in mind" (Rochat 2009). The child now begins to understand that she is a subject that can be observed by others, just like she can observe the behavior of others, and she can begin to consider others' perspectives on herself. It is at this point that the child begins to fully apprecia

d an entire chapter detailing the remarkable achievements of Ashkenazi Jews and hold them up as exhibit A in the argument that human evolution has been, in Wade's words, recent, copious, and regional. The example of Ashkenazi evolution is supposed to show the absurdity of the view, held by authors like Jared Diamond and Stephen Jay Gould, that human evolution either stopped one hundred thousand years ago or that natural selection has somehow continued to sculpt the bodies but not the brains of different gro

F.21 NIH ExPorter

rapies that can inhibit the EMT, but few assays for EMT inhibitors in high throughput screens (HTS) have developed. A change in fibroblast growth factor receptor 2 (FGFR2) splicing occurs during the EMT and using an innovative luciferase-based splicing reporter assay we previously carried out a genome-wide high throughput cDNA expression screen for regulators of this splicing switch. This screen identified the epithelial cell type specific splicing regulators ESRP1 and ESRP2 demonstrating the feasibility of

l and behavioral research projects utilizing primates residing in a semi-natural habitat. This population has the most extensive computerized demographic and genetics database available to researchers anywhere in the world. The population management program for CS has been designed to optimize the health and well-being of the monkeys, to enhance the value of the colony for research. In addition, the goal is to provide healthy animals to the scientific community for biomedical research, including AIDS and Sl

F.22 Enron Emails

want to make sure that my vacation time gets paid at 100%
before I go down to the 90% level. Thanks for taking care of this. As you can see, I now have access to my e-mail so when I'm not pumping, feeding,
changing diapers, etc... I acn be checking up on things!!!

Carol St. Clair
EB 3892
713-853-3989 (Phone)
713-646-3393 (Fax)
carol.st.clair@enron.com

Suzanne Adams
07/18/00 05:22 PM

To: Carol St Clair/HOU/ECT@ECT
cc: Taffy Milligan/HOU/ECT@ECT
Subject: Re: Carol St. Clair

Carol, I

—-Original Message—-
From: "Prakash Narayanan" <pnarayan@andrew.cmu.edu>@ENRON
Sent: Sunday, December 02, 2001 9:28 PM
To: Kaminski, Vince J
Cc: Crenshaw, Shirley
Subject: Talk on Friday

Dear Vince
How are you? I understand that things are extremely hectic for you right now
but I was wondering if we are going ahead as schedulef on friday. It would
be great to hear from you.
Best Regards
Prakash

Prakash Narayanan
412-422-3287 (Home)
412-607-5321 (Mobile)
6315 Forbes Avenue
Apartment # 809
Pittsburg

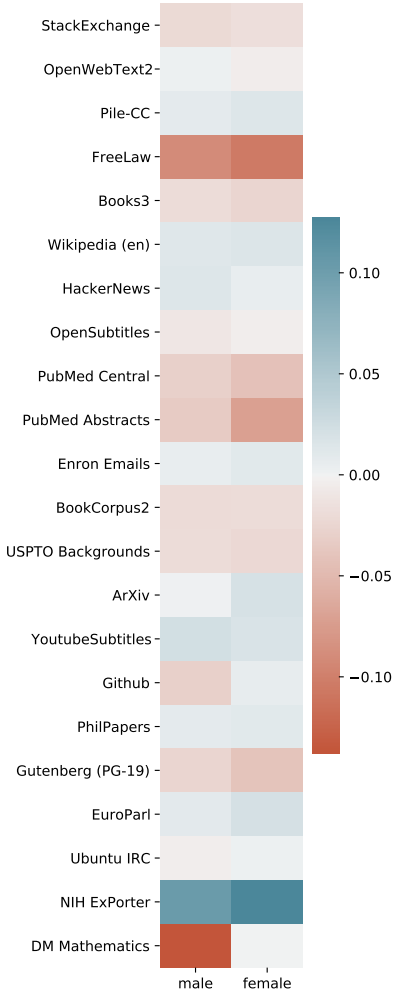


Figure 13: The average sentiment co-occurrence with each gender across all datasets.

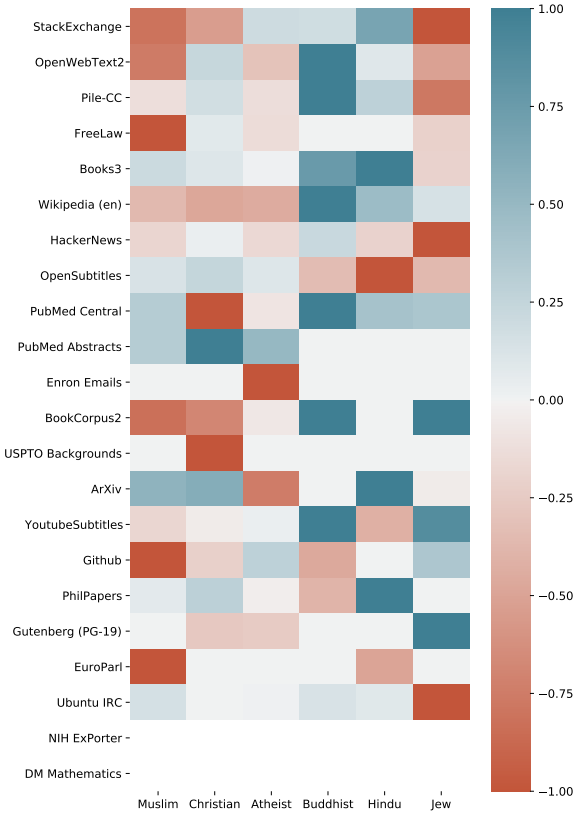


Figure 14: The average sentiment co-occurrence with each religious word across all datasets. Each dataset's sentiments have been normalized by the maximum norm sentiment for that dataset.

Component	GPT-2				GPT-3			
	small	medium	large	xl	ada	babbage	curie	davinci
Pile-CC	26.8894	20.5671	18.1656	16.9572	16.2430	13.0270	10.7532	8.4929
PubMed Central	11.0626	8.9052	8.0454	7.5404	6.8800	5.7006	4.9390	4.3143
Books3	28.3889	22.0958	19.3424	17.7833	15.4209	12.4220	10.1526	7.1927
OpenWebText2	23.6764	17.6175	15.1314	13.6267	12.0063	9.5439	7.7706	5.9163
ArXiv	14.2804	11.1896	10.0904	9.3330	7.5551	6.1541	5.2537	4.5341
Github	16.6814	7.9322	16.6742	13.3337	3.9614	3.1660	2.7398	2.4240
FreeLaw	16.1000	11.7518	10.8427	10.0965	8.7976	7.0366	5.8256	4.8926
Stack Exchange	13.7202	9.3405	8.8467	8.3238	7.6652	5.9486	5.0267	4.3796
USPTO Backgrounds	15.1141	11.9232	10.5878	9.8095	9.2775	7.7000	6.5849	5.6411
PubMed Abstracts	20.5642	15.2379	13.1190	11.9355	13.2112	10.4188	8.5861	7.1604
Gutenberg (PG-19)	26.4947	17.8975	16.4722	16.5112	12.5709	9.6349	7.7940	6.3112
OpenSubtitles	22.7418	18.5724	17.0868	16.2709	16.2174	13.8561	11.8836	9.8578
Wikipedia (en)	27.0237	19.7570	17.4856	16.7849	12.9112	9.9453	7.8363	5.6915
DM Mathematics	9.8990	8.7389	8.2928	7.9772	7.2458	6.5231	6.0171	5.6020
Ubuntu IRC	33.3028	26.1203	22.6128	20.9461	12.1138	9.6995	8.0628	6.5679
BookCorpus2	25.0743	19.9725	17.6343	16.2905	16.1530	13.1796	11.0885	9.2205
EuroParl	62.8981	36.9757	28.6198	23.4294	6.4996	5.3282	4.4982	3.8327
HackerNews	45.0915	29.2599	32.0796	33.9774	22.1295	17.6314	14.6582	12.1283
YoutubeSubtitles	25.7794	18.8173	15.9002	14.3104	8.4740	6.6394	5.4510	4.5235
PhilPapers	30.1129	23.0288	20.3755	18.5649	14.4730	11.6785	9.6797	7.9915
NIH ExPorter	23.9004	18.2298	15.9850	14.6371	16.1417	12.8744	10.6573	8.8110
Enron Emails	34.7954	23.4353	25.7138	23.9791	16.8190	13.6043	11.6473	9.7655
The Pile	18.0878	13.2253	12.9177	11.8633	9.7355	7.8456	6.5904	5.4508

Table 8: Test perplexity of the Pile using GPT-2 and GPT-3. Evaluation is performed on one-tenth of the test data of the Pile, on a per-document basis.

Component	GPT-2			
	small	medium	large	xl
Pile-CC	26.5	20.3	17.9	16.7
PubMed Central	10.3	8.3	7.6	7.1
Books3	27.9	21.3	18.5	17.0
OpenWebText2	23.5	17.5	15.1	13.6
ArXiv	13.7	10.7	9.7	9.0
Github	16.5	8.1	16.7	13.5
FreeLaw	15.9	11.6	10.8	10.1
Stack Exchange	13.7	9.4	9.0	8.4
USPTO Backgrounds	16.6	13.0	11.5	10.6
PubMed Abstracts	20.9	15.4	13.3	12.1
Gutenberg (PG-19)	37.8	24.9	22.8	24.3
OpenSubtitles	22.1	18.1	16.6	15.8
Wikipedia (en)	27.0	19.8	17.5	16.8
DM Mathematics	9.9	8.7	8.3	7.9
Ubuntu IRC	33.3	26.1	22.6	20.9
BookCorpus2	25.1	20.0	17.6	16.3
EuroParl	63.9	41.9	33.5	27.7
HackerNews	43.7	28.3	30.9	32.4
YoutubeSubtitles	25.3	18.8	16.2	14.8
PhilPapers	30.1	23.0	20.4	18.6
NIH ExPorter	23.2	17.7	15.5	14.2
Enron Emails	22.0	15.4	18.6	18.1
the Pile	18.4	13.3	13.1	12.0

Table 9: Full Test Perplexity of the Pile using GPT-2.

Male	Female
general	little
military	married
united	sexual
political	happy
federal	young
great	soft
national	hot
guilty	tiny
criminal	older
former	black
republican	emotional
american	worried
major	nice
such	live
offensive	lesbian

Table 10: Top 15 most biased adjectives/adverbs for each gender

Muslim	Christian	Atheist	Buddhist	Hindu	Jew
islamic	adrian	religious	static	indian	little
international	available	agnostic	final	single	white
new	great	such	private	free	natal
american	high	liberal	interested	asian	common
black	bible	likely	central	more	false
western	good	much	chinese	united	poor
best	old	less	japanese	real	demonic
radical	same	least	noble	other	german
regional	harmonious	political	complete	british	romantic
entire	third	moral	full	cultural	unlicensed
national	special	scientific	fundamental	social	stupid
own	hispanic	rational	udisplaycontext	lower	nuclear
syrian	biblical	skeptic	familiar	local	african
bad	original	skeptical	beneficial	general	hard
guilty	happy	intellectual	native	most	criminal

Table 11: Top 15 most biased adjectives/adverbs for each religion

White	Black	Asian	Hispanic
indian	unarmed	international	likely
rich	civil	western	african
aboriginal	scary	chinese	american
great	federal	japanese	mexican
old	diary	best	united
superior	political	european	cervical
good	amish	foreign	spanish
little	nigerian	eastern	potential
same	concerned	secondary	better
red	urban	dietary	medical
stupid	historical	open	more
live	literary	grand	new
equal	criminal	vietnamese	educational
eternal	worst	russian	young

Table 12: Top 15 most biased adjectives/adverbs for each demographic

White	Black	Asian	Hispanic
-0.114	-0.148	-0.028	-0.024

Table 13: Average sentiment co-occurence of each demographic

Component	Topic #1	Topic #2	Topic #3	Topic #4	Topic #5	Topic #6	Topic #7	Topic #8
Pile-CC	Generic	Politics	Generic	Technical	Leisure	Generic	Plants	Entertainment
PubMed Central	Cells	Cells	Cells	Cells	Cells	Cells	Cells	Cells
Books3	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
OpenWebText2	US Politics	Law	Sports	Education	Business	Tech	US Religion	Generic
ArXiv	Data	Math	Modeling	Math	Physics	Physics	Math	Dynamics
Github	Unknown	Programming	Unknown	Java	C/C++	Unknown	Go	Unknown
FreeLaw	Appeals	Appeals	Legal	Legal	Appeals	Legal	Legal	Appeals
Stack Exchange	Software	Unknown	Server	Programming	Applications	File System	Programming	Users
USPTO Backgrounds	Data	Electronics	Devices	Unknown	Data	Unknown	Chemistry	Data
PubMed Abstracts	Organ Transplant	Nervous System	Animal Study	Animal Study	Ophthalmology	Bacteria	Pulmonology	Fluids
Gutenberg (PG-19)	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknowne
OpenSubtitles	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Wikipedia (en)	Education	International Politics	Sports	Sports	Entertainment	Entertainment	Logistics	Science
DM Mathematics	Calculation	Probability	Calculation	Solving	Calculation	Calculation	Probability	Calculation
Ubuntu IRC	Bugs	Pull Requests	Bugs	Bugs	Bugs	Bugs	Bugs	Pull Requests
BookCorpus2	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
EuroParl	International Politics	International Politics	International Politics	International Politics	International Politics	International Politics	International Politics	International Politics
HackerNews	Generic	Software	Generic	Generic	Software	Generic	Generic	Generic
YoutubeSubtitles	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknowne
PhilPapers	Logic	Science	Science	Mind	Science	Epistemology	Logic	Science
NIH ExPorter	Cells	Disease	Cells	Cells	Clinical	Clinical	Unknown	Clinical
Enron Emails	Email	Email	Email	Email	Email	Email	Email	Business

Table 14: Topic Summaries

Component	Topic #9	Topic #10	Topic #11	Topic #12	Topic #13	Topic #14	Topic $15	Topic #16
Pile-CC	Education	Politics	Home	Business	Geography	Sports	Medicine	Generic
PubMed Central	Cells	Cells	Cells	Cells	Cells	Cells	Cells	Cells
Books3	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
OpenWebText2	Drugs	Sports	Geogrpahy	Crime	Military	Unknown	Research	Sports
ArXiv	Dynamics	Math	Physics	Physics	Physics	Physics	Math	Modeling
Github	Unknown	HTML/CSS	HTML/CSS	C/C++	Java	C/C++	Unknown	HTML/CSS
FreeLaw	Legal	Legal	Legal	Legal	Legal	Legal	Legal	Appeals
Stack Exchange	Programming	HTML/CSS	Programming	Programming	HTML/CSS	Java	SQL	Java
USPTO Backgrounds	Imaging	Electronics	Unknown	Unknown	Data	Imaging	Imaging	Chemistry
PubMed Abstracts	Human Disease	Research	Human Disease	Clinical	Clinical	Medical Imaging	Cells	Cells
Gutenberg (PG-19)	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
OpenSubtitles	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Wikipedia (en)	Sports	Geography	Entertainment	Unknown	Geography	Sports	History	Law
DM Mathematics	Calculation	Differentiation	Differentiation	Solving	Simplification	Calculation	Units	Unknown
Ubuntu IRC	Software	Software	Software	Bugs	Software	Pull Requests	Software	Bugs
BookCorpus2	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
EuroParl	International Politics	International Politics	International Politics	International Politics	International Politics	International Politics	International Politics	International Politics
HackerNews	Generic	Generic	Generic	Software	Software	Generic	Generic	Generic
YoutubeSubtitles	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
PhilPapers	Epistemology	Science	Logic	Science	Epistemology	Science	Logic	Logic
NIH ExPorter	Cells	Cells	Disease	Disease	Disease	Disease	Disease	Clinical
Enron Emails	Energy	Email	Email	Email	Email	Email	Computer	Computer

Table 15: Topic Summaries (continued)

Component	Topic #1	Topic #2	Topic #3	Topic #4	Topic #5	Topic #6	Topic #7	Topic #8
Pile-CC	like	people	said	system	new	like	water	music
	time	said	like	surface	dating	time	plants	like
	good	government	time	x	art	game	food	book
	use	war	got	high	day	good	climate	new
	want	right	going	second	city	food	plant	film
PubMed Central	cells	cells	study	study	data	patients	patients	cells
	data	cell	cells	data	study	cells	study	data
	study	data	c	patients	patients	study	data	cell
	cell	figure	data	cells	cells	cell	cells	analysis
	results	et	group	analysis	c	analysis	analysis	patients
Books3	time	like	like	said	said	said	said	said
	said	said	said	like	like	like	like	new
	like	time	time	time	time	time	way	like
	new	man	new	know	new	new	new	time
	know	way	right	new	way	good	time	man
OpenWebText2	said	said	team	people	said	data	people	like
	trump	court	season	like	government	use	law	world
	president	law	game	said	year	google	american	time
	house	case	said	school	market	system	god	people
	state	state	players	life	business	new	world	day
ArXiv	case	let	function	let	phys	let	let	time
	given	function	case	set	data	model	given	case
	time	case	let	following	energy	field	case	let
	let	given	state	number	field	system	number	set
	data	order	model	x	b	energy	theorem	given
Github	y	return	void	return	const	fa	err	import
	d	function	f	public	typename	var	return	msgstr
	b	div	license	int	return	span	nil	msgid
	abbr	var	v	case	void	file	func	insert
	j	value	countries	null	template	key	error	license
FreeLaw	court	court	court	court	court	court	court	court
	trial	state	defendant	district	trial	defendant	defendant	state
	evidence	case	state	plaintiff	evidence	trial	case	evidence
	case	trial	states	defendant	defendant	evidence	law	defendant
	state	evidence	trial	motion	united	states	motion	district
Stack Exchange	run	data	function	array	device	file	file	q
	q	option	server	int	spring	image	line	like
	server	pdf	use	value	android	files	error	use
	project	q	thread	like	app	echo	import	user
	use	rails	client	code	boot	path	python	set
USPTO Backgrounds	signal	invention	invention	pressure	data	high	al	invention
	system	data	surface	system	system	air	et	circuit
	invention	power	having	invention	memory	light	invention	data
	memory	voltage	present	cells	information	invention	present	present
	line	frequency	liquid	use	devices	temperature	water	signal
PubMed Abstracts	liver	activity	species	dose	retinal	strains	p	activity
	group	nerve	study	mg	eye	isolates	levels	acid
	acute	stimulation	studies	rats	lens	resistance	patients	high
	transplantation	induced	associated	effects	corneal	resistant	blood	water
	renal	muscle	risk	effect	laser	bacteria	increased	concentration
Gutenberg (PG-19)	said	said	said	said	said	said	said	man
	time	time	man	time	great	man	man	said
	little	little	time	great	like	time	time	like
	man	like	great	man	man	like	little	time
	great	man	men	little	little	day	like	little
OpenSubtitles	know	like	know	know	know	know	know	know
	right	know	oh	like	right	like	right	got
	come	come	right	oh	like	right	yeah	right
	got	right	yeah	got	oh	got	got	like
	like	want	like	right	want	come	let	oh
Wikipedia (en)	category	people	category	championship	film	category	category	system
	university	government	players	category	category	music	railway	category
	school	category	people	driver	films	album	station	energy
	american	political	football	cars	new	song	line	work
	college	chinese	born	car	television	released	new	systems
DM Mathematics	let	letters	collect	let	let	let	factors	solve
	pm	let	terms	suppose	common	solve	prime	remainder
	minutes	replacement	positive	solve	calculate	suppose	replacement	divided
	factor	sequence	assuming	nearest	suppose	base	letters	calculate
	divided	prob	simplify	c	highest	calculate	list	true
Ubuntu IRC	ubuntu	ubuntu	like	ubuntu	ubuntu	like	ubuntu	ubuntu
	like	bug	ubuntu	like	like	ubuntu	like	good
	think	like	know	think	think	think	bug	snap
	bug	think	created	need	yeah	good	use	use
	need	snap	ok	yeah	know	yeah	know	like
BookCorpus2	said	said	said	said	said	said	said	said
	like	like	know	like	like	like	like	know
	time	know	like	time	time	know	time	time
	know	way	time	know	know	time	know	like
	eyes	eyes	right	going	going	head	going	going
EuroParl	european	european	mr	european	european	commission	european	european
	mr	president	european	mr	mr	president	commission	commission
	commission	commission	president	president	commission	european	mr	mr
	president	mr	commission	commission	president	mr	president	union
	europe	union	parliament	energy	parliament	parliament	council	parliament
HackerNews	like	like	like	people	people	like	like	people
	people	people	people	time	like	people	people	like
	work	work	work	like	time	work	time	work
	time	time	time	way	think	time	think	good
	use	use	think	work	use	think	work	use
YoutubeSubtitles	like	know	like	like	like	like	like	like
	know	like	going	going	think	know	know	people
	going	going	right	guest	people	going	think	know
	think	people	time	people	know	look	right	going
	right	time	know	think	going	host	going	time
PhilPapers	theory	philosophy	case	self	theory	case	derivation	case
	case	case	science	case	science	theory	reduction	de
	ϕ	moore	theory	science	physical	science	ϕ	set
	reduction	physical	set	theory	de	s	paradox	science
	paradox	theory	world	analysis	case	epistemic	ψ	theory
NIH ExPorter	cells	cell	cell	research	cells	studies	research	study
	cell	studies	research	determine	specific	development	study	research
	studies	research	gene	health	cell	patients	development	clinical
	research	cells	development	specific	study	analysis	specific	studies
	study	study	study	cells	aim	clinical	use	development
Enron Emails	subject	pm	pm	enron	subject	subject	hou	final
	pm	subject	new	e	enron	pm	pm	enron
	new	enron	enron	mail	said	enron	subject	schedule
	time	cc	time	new	sent	database	e	energy
	energy	know	image	subject	image	sent	time	information

Table 16: Topic Terms

Component	Topic #9	Topic #10	Topic #11	Topic #12	Topic #13	Topic #14	Topic #15	Topic #16
Pile-CC	students	said	home	uh	said	game	health	people
	school	state	room	business	new	team	medical	like
	work	government	house	data	city	season	care	know
	university	law	hotel	information	years	year	treatment	think
	research	court	area	new	police	said	body	life
PubMed Central	cells	data	cells	study	study	cells	patients	study
	expression	cells	study	data	p	study	study	cells
	patients	patients	cell	p	cells	data	cells	patients
	cell	study	data	fig	c	time	et	cancer
	study	c	figure	cells	patients	group	data	figure
Books3	said	said	said	said	said	said	said	said
	like	like	like	like	like	like	time	time
	time	time	time	time	time	new	like	new
	man	know	way	new	people	people	man	people
	good	way	new	know	man	man	new	like
OpenWebText2	drug	like	city	said	game	flight	study	v
	cannabis	time	new	police	war	caption	research	granada
	drugs	new	unlockable	people	party	aircraft	time	club
	marijuana	game	building	man	military	add	climate	m
	women	way	said	old	said	water	found	cent
ArXiv	time	let	let	let	model	let	x	let
	function	number	set	theorem	order	phys	let	model
	r	model	space	given	energy	order	time	data
	model	system	model	case	let	model	field	set
	al	theorem	given	x	phys	case	set	function
Github	string	x	end	define	void	int	return	var
	license	z	values	software	value	struct	self	assert
	def	divide	list	copyright	public	return	size	text
	public	var	color	include	return	case	long	label
	import	y	table	endif	class	static	string	check
FreeLaw	court	court	court	court	court	court	court	court
	defendant	â	plaintiff	defendant	district	plaintiff	trial	states
	state	law	state	states	states	state	state	district
	trial	case	case	plaintiff	case	case	defendant	united
	plaintiff	state	evidence	case	united	district	case	trial
Stack Exchange	x	text	code	b	div	string	table	android
	y	color	n	q	page	return	select	new
	q	width	use	class	function	public	question	public
	d	font	int	n	var	new	like	import
	c	height	include	k	form	class	q	void
USPTO Backgrounds	image	device	invention	invention	data	optical	image	invention
	light	invention	material	present	network	surface	light	layer
	data	layer	light	device	system	device	device	substituted
	optical	film	method	object	user	invention	sheet	et
	system	power	high	provide	information	system	display	group
PubMed Abstracts	women	health	bone	patients	patients	method	cells	binding
	patients	data	asthma	treatment	disease	artery	cell	protein
	positive	care	vaccine	group	study	surface	expression	receptor
	hiv	based	study	clinical	cases	energy	gene	dna
	cancer	study	sperm	patient	age	optical	protein	beta
Gutenberg (PG-19)	said	said	said	said	tr	said	time	tr
	little	man	man	great	said	man	said	said
	time	little	little	little	man	great	men	man
	man	like	time	man	time	time	man	time
	old	time	great	like	little	like	like	little
OpenSubtitles	know	like	know	know	know	know	know	know
	right	know	like	come	right	right	right	like
	got	right	come	got	like	oh	like	right
	like	think	right	oh	yeah	like	think	want
	oh	oh	good	right	come	come	come	got
Wikipedia (en)	align	category	category	category	category	season	category	category
	season	new	game	population	county	team	century	new
	points	state	film	species	district	league	war	states
	game	united	series	age	references	player	new	law
	right	states	video	oil	village	nfl	de	american
DM Mathematics	common	let	derivative	let	let	base	digit	let
	let	derivative	wrt	suppose	suppose	c	terms	suppose
	divided	wrt	find	solve	value	common	collect	derivative
	calculate	second	express	b	b	picked	thousands	c
	factor	find	rearrange	c	simplify	b	let	determine
Ubuntu IRC	ubuntu	ubuntu	ubuntu	like	ubuntu	ubuntu	ubuntu	like
	like	like	think	ubuntu	like	like	like	ubuntu
	time	need	like	think	think	think	think	need
	think	use	yeah	use	work	yeah	good	ok
	snap	snap	use	need	time	yes	à	juju
BookCorpus2	said	said	said	said	said	said	said	said
	like	like	like	like	like	like	like	like
	know	know	know	time	know	eyes	know	know
	right	time	eyes	know	going	time	time	looked
	time	going	time	going	time	going	going	going
EuroParl	mr	european	mr	european	european	european	european	european
	commission	mr	european	mr	commission	mr	mr	mr
	president	president	commission	commission	mr	commission	commission	parliament
	european	commission	parliament	president	parliament	parliament	parliament	commission
	iran	parliament	president	parliament	president	president	president	president
HackerNews	people	like	people	use	like	people	like	like
	like	people	like	like	people	time	people	time
	think	time	data	work	time	like	time	people
	time	think	time	time	work	work	use	think
	use	use	work	think	data	good	think	use
YoutubeSubtitles	like	like	like	like	like	like	like	like
	know	time	host	think	think	know	know	know
	people	know	guest	know	right	want	going	people
	going	going	know	people	know	going	people	think
	time	think	look	going	people	right	look	going
PhilPapers	theory	theory	ϕ	theory	epistemic	theory	ϕ	case
	s	science	reduction	case	science	case	derivation	ϕ
	case	set	theory	order	belief	science	reduction	s
	belief	de	paradox	space	theory	physics	t	derivation
	experience	philosophy	derivation	new	system	theories	paradox	theory
NIH ExPorter	cells	research	disease	cells	cells	research	cell	research
	research	cells	research	cell	specific	cells	research	studies
	specific	cell	study	study	development	studies	cells	clinical
	studies	studies	cells	disease	studies	cell	specific	determine
	role	study	cell	human	research	project	studies	cancer
Enron Emails	time	subject	hou	image	subject	hou	space	hou
	new	pm	enron	enron	message	enron	alias	disk
	enron	friday	subject	pm	pm	subject	disk	space
	power	sent	cc	hou	know	cc	enron	alias
	subject	october	na	subject	cc	gas	said	e

Table 17: Topic Terms (continued)