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

arXiv 2020

HUMANE Lab 박현빈

2025.01.10

Background

- dataset diversity leads to better generalization capability
- LLM effectively acquire knowledge in a novel domain with small amounts of training data
- large number of smaller, high-quality, diverse datasets can improve the generalization capability

Contribution

- introduce an 825.18 GB english-language dataset for language modeling, which combines 22 diverse sources
- introduce 14 new datasets
- GPT-2-sized models trained on a new dataset demonstrate improved performance across diverse domains

The Pile Dataset

- existing datasets (7)

  Books3, PG-19, Open Subtitles, English Wikipedia, DM Math, EuroParl, and Enron Emails corpus

- novel datasets (12)

  PubMed Central, ArXiv, Github, FreeLaw, Stack Exchange, US Patent and Trademark and Office, PubMed, Ubuntu IRC, HackerNews, Youtube, PhilPapers, and NIH ExPorter

- extended datasets (2)

  OpenWebText2 (OpenWebText), BookCorpus2 (BookCorpus)

- filtered dataset (1)

  Pile-CC (Common Crawl)

The Pile Dataset

- epochs
  - Number of times a dataset is viewed
  - Higher value for good data quality and small dataset size

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**

The Pile Dataset

- Pile-CC
  - Common Crawl includes text from diverse-domains, but the quality varies significantly
  - **there is a need for extraction and filtering**
  - **higher-quality output from CC using jusText extraction**

- **PubMed Central**
  - **open full-text repository for biomedical research(PMC)**
  - **potential benefits for medical domain applications**

The Pile Dataset

- Books3
  - large dataset for fiction and nonfiction books
  - inclusion of book data for its importance long-range context modeling and storytelling

- OpenWebText2
  - extension of the original OpenWebText
  - inclusion of Reddit data up to 2020
  - high quality dataset with multilingual and recent content

The Pile Dataset

- ArXiv
  - high quality text(math, computer science, physics) and math knowledge
  - inclusion of arXiv to support research fields

- Github
  - motivation from GPT-3’s ability to generate plausible code completions without explicitly trained code datasets

The Pile Dataset

- Stack Exchange
  - one of the largest publicly available repositories of question-answer pairs
  - inclusion to enhance question answering in diverse domains

- USPTO Backgrounds
  - USPTO Backgrounds as a dataset of patent background sections
  - inclusion for its large volume of technical content

The Pile Dataset

- PubMed Abstracts
  - collection of 30 millions biomedical abstract
- Project Gutenberg
  - dataset of classic Western literature
  - distinct styles from the more modern Books3 and BookCorpus
  - using for long-context modeling

The Pile Dataset

- OpenSubtitles
  - dataset of English subtitles from movies and TV shows
  - use of subtitles for creative writing tasks like screenwriting

- DeepMind Mathematics
  - collection of mathematical problems such as algebra, arithmetic, probability ...
  - inclusion to improve the mathematical ability of LM

The Pile Dataset

- BookCorpus2
  - expanded version of BookCorpus
  - books by unpublished authors in BookCorpus

- Ubuntu IRC
  - dataset from Ubuntu-related channels on the Freenode IRC chat server
  - inclusion to provide an opportunity to model real-time human interactions

The Pile Dataset

- Youtube Subtitles
  - collection of human generated captions
  - source of educational content, popular culture, and natural dialog
- PhilPapers
  - collection of philosophy publications
- NIH Grant Abstract ExPORTER
  - high-quality scientific writing

The Pile Dataset

- Enron Emails
  - inclusion to understanding the modality of email communications
- EuroParl
  - multilingual corpus for machine translation

Benchmarking LMs with the Pile

- while the Pile dataset was conceived as a training dataset, it is also suitable for use an evaluation dataset
- the validation and testing components each contain 0.1% of the data, sampled uniformly at random
- deduplication has been performed on the Pile, but it may not perfect

Benchmarking LMs with the Pile

- evaluate metric : bits per UTF-8 encoded byte (BPB)

$$\text{BPB} = (L_T/L_B)\log_2(e^{\ell}) = (L_T/L_B)\ell/\ln(2)$$

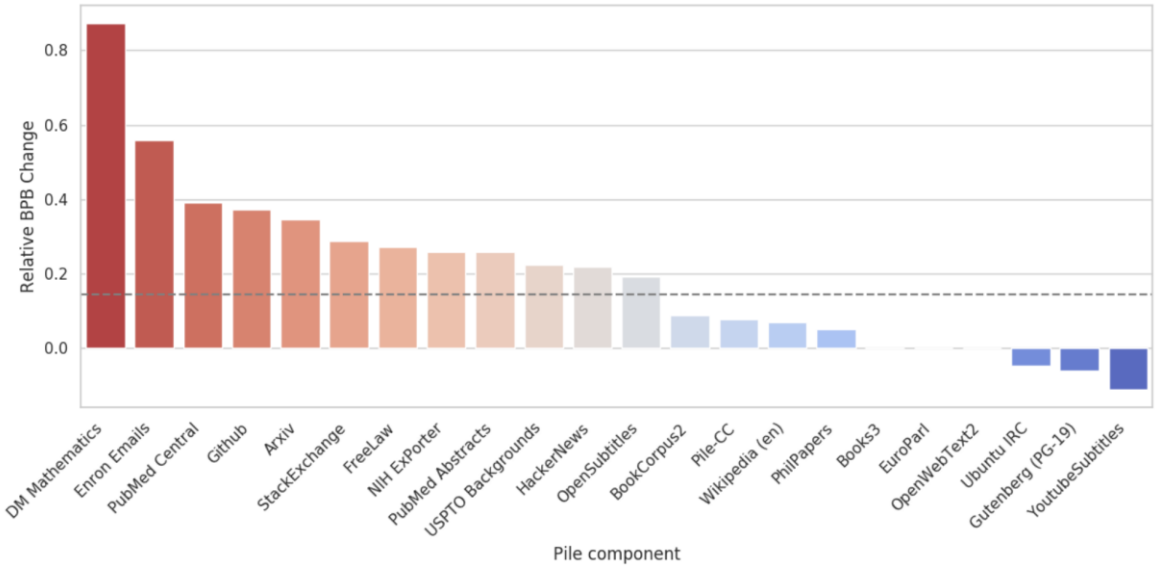
$$\text{perplexity}_p(x_{1:L}) = \exp\left(\frac{1}{L}\sum_{i=1}^{L}\log\frac{1}{p(x_i \mid x_{1:i-1})}\right)$$

  - L_T is the length of the dataset in tokens
  - L_B is the length of the dataset in UTF-8 encoded bytes
  - l is a given negative log likelihood loss
  - BPB indicates how well model compresses information

Relative Componentwise GPT-3 Pile Performance

- knowing which components GPT-3 struggles with is important because it allows for enhancing the training data
- due to the difference in entropy across datasets, perplexity is not a reliable matric

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



Evaluation

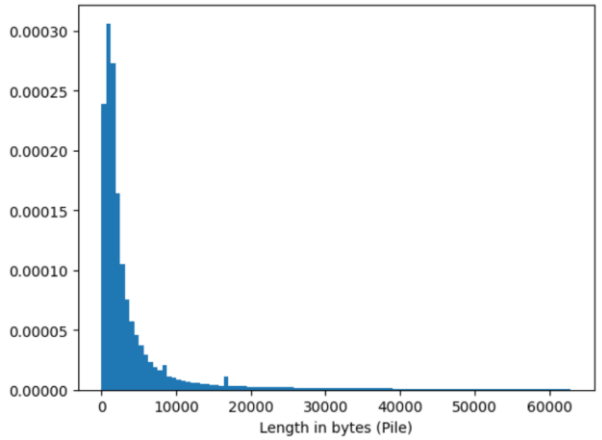
- model : 1.3B models
- decontaminate instances of the evaluation sets using the same 13-gram overlap filtering and downsample to 40GB to control for dataset size

	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.

Structural Statistics

- document lengths (bytes-per-token)



  - a majority of short ones with a few very long ones
  - because of GPT-2 BPE tokenizer trained on WebText, WebText와 비슷한 데이터셋에 대해서는 많은 bytes per token을, 많이 다른 유형의 데이터셋은 낮은 bytes-per-token을 기록했다

Investigating and Documenting the Datasets

- 데이터셋 검증이 중요해지고 있다
- 나쁜 데이터셋을 제거하기 보다 문서로 남겨 연구원들이 필요에 맞게 사용하는 것이 좋다고 판단
- 주제, 욕설, 편향 등등에 대해 조사함

Investigating and Documenting the Datasets

- topical distribution
  - Pile dataset은 어떤 주제를 다루는가
  - 웹에서 크롤링한 방대한 데이터셋인 Pile-CC에 없는 topic이 다른 데이터셋에 있을까?
  - LDA라는 옛날 모델에 Pile-CC 학습시키고, 다른 데이터셋에 대해 perplexity를 계산했을 때 전혀 다른 topic을 다루는 Github, PhilPapers, EuroParl 데이터셋에서는 perplexity가 높게 나옴.
  - 즉 Pile 데이터셋은 다양한 주제를 다루고 있다.

Investigating and Documenting the Datasets

- profanity
  - profanity-checker Python package. toxic model trained on Wikidetox Toxic Comment Dataset.
  - 각 문장을 단어로 나눠서 욕설로 플래그된 단어 비율 계산
  - 대체 추정치로 간주
  - 전체적으로 Pile-CC보다 욕설 비율이 낮더라

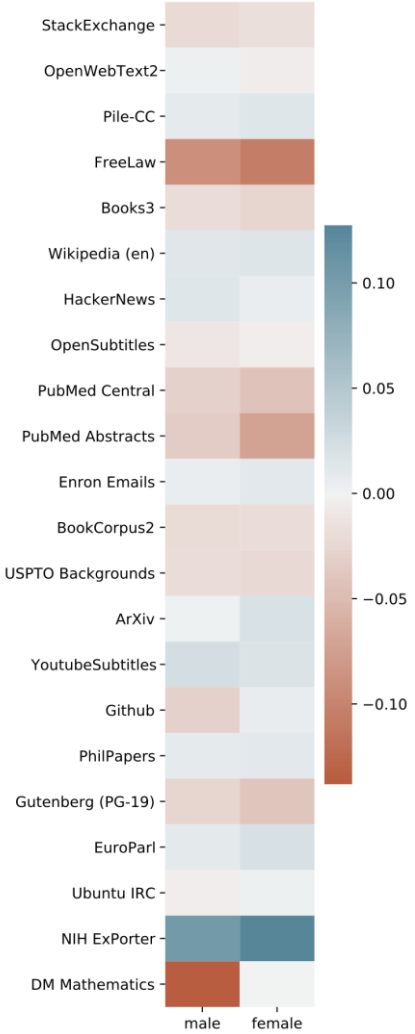
Investigating and Documenting the Datasets

- bias
  - 크롤링한 데이터가 많기 때문에 편향이 있을 수 있다.
  - 성별, 종교, 인종에 대해 데이터셋이 가지고 있는 편향을 조사했다

Investigating and Documenting the Datasets

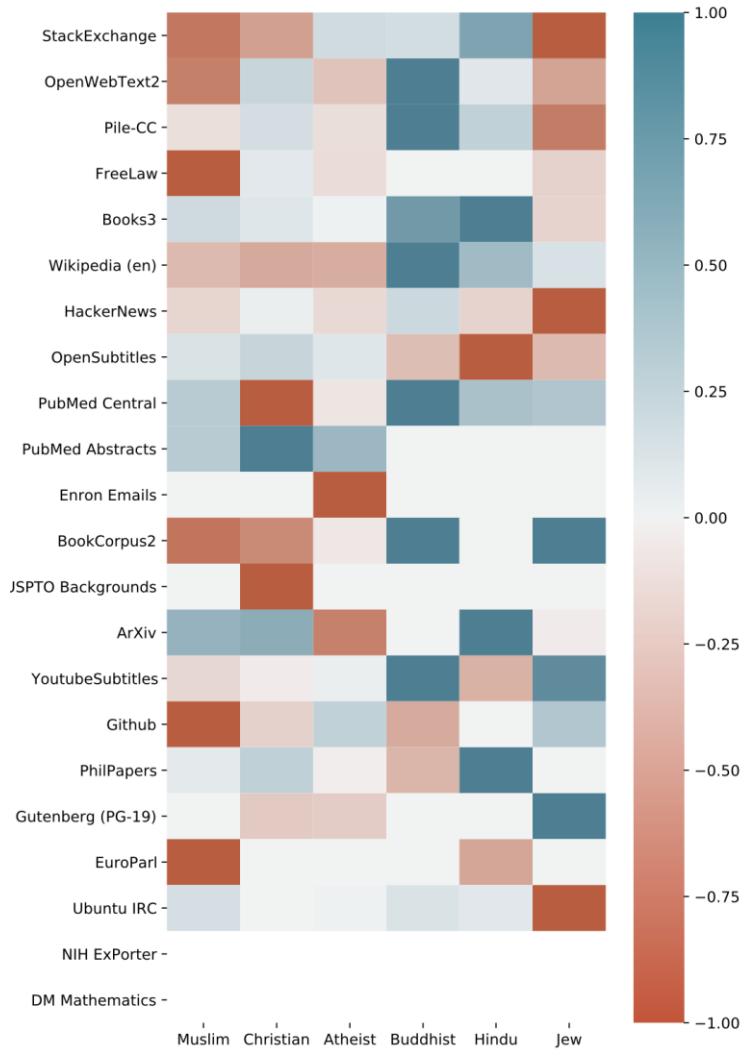
- gender
  - 문장에 ‘he’와 같이 자주 등장하는 단어와 ‘she’와 같이 등장하는 단어 각각
  - 성별별로 자주 등장하는 단어들이 달랐다
  - 하지만 특정 성별에 sentiment 편향은 발견되지 않았다

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



Investigating and Documenting the Datasets

- religion
  - ‘radical’은 ‘muslim’과 같이 등장하고, ‘rational’은 ‘atheist’와 같이 등장함
  - 하지만 같이 등장하는 단어들로 편향을 판단하는 방법은 한계가 있음. 예를 들어 ‘religious’와 ‘atheist’는 같이 여러 번 등장함. 대화 유형 때문
  - ‘Buddhist’가 좋은 감정의 단어들과 자주 등장, ‘Muslim’은 나쁜 감정의 단어들과 자주 등장함
  - 종교에 대한 편향이 있음을 확인



Investigating and Documenting the Datasets

- race
  - ‘black’ 은 ‘criminal’, ‘scary’, ‘unarmed’와 같은 단어들과 자주 등장
  - 흑인이 평균 감정 점수가 가장 낮다.
  - 모든 인종이 음의 점수를 기록했다. 이는 인종에 대한 언급이 뉴스에서는 주로 용의자를 언급하는 부정적인 상황에서 나오기 때문이다

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

Ethical Considerations

- copyright
- Legality
- Accelerate AI development
- Negative LM Output
  - 바람직하지 않은 컨텐츠가 훈련 데이터에 포함되는것은 불가피함. 훈련데이터에서 이를 삭제하는 것은 비효율적이고 optimal solution이 아님. 부적절한 데이터를 이해하면서 동시에 무시하는 능력을 길러야 한다. 이는 향후에 더 연구되어야 할 부분이다

Thank you