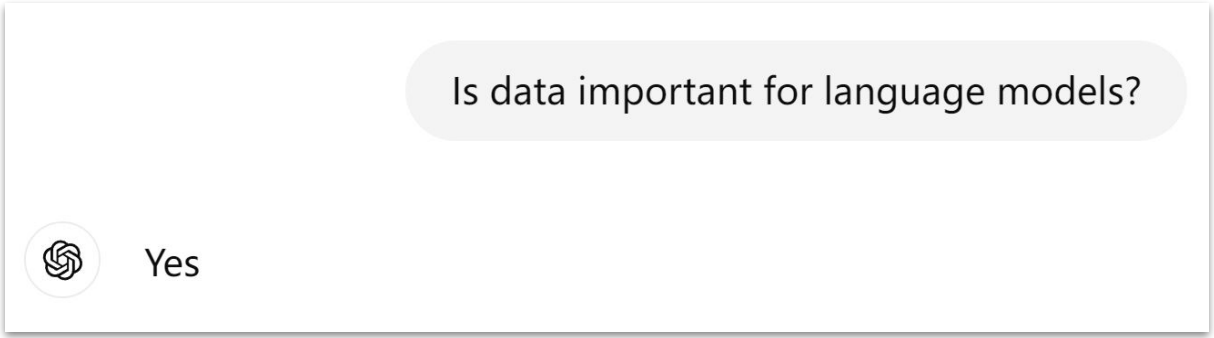
MSBD 6000N Presentation

Data

Li Congjiao, WU FUNG Xian Hong, Liu Wentao

Data🌐



Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus

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

Deduplicating Training Data Makes Language Models Better

Documenting Large Webtext Corpora

A Case Study on the Colossal Clean Crawled Corpus

Dodge et al., ACL 2021

Presenter:
LI Congjiao

Motivation

Present Situation - Webtext corpora's importance and deficiency

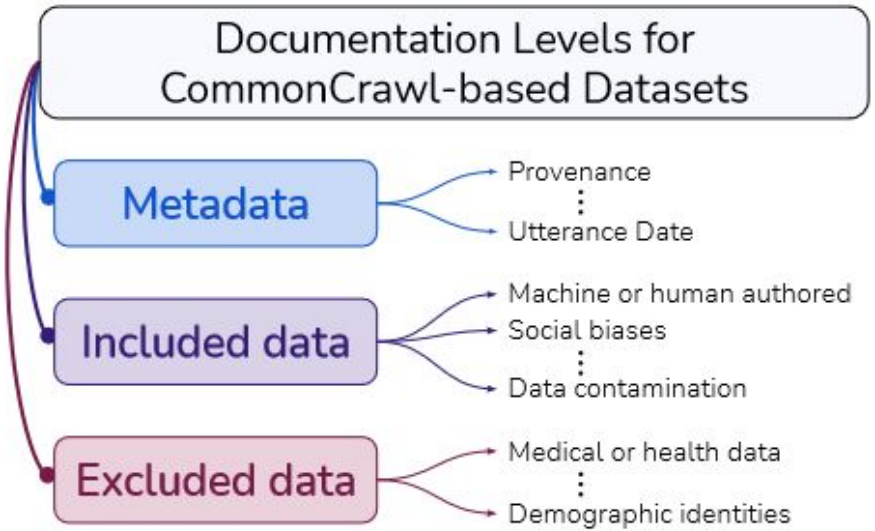
- Models trained on webtext corpora are the backbone of many modern NLP systems.
- Many corpora with only minimal documentation are frequently introduced or used.

Motivation - Why documenting for Large Webtext Corpora?

- Lack of documentation leaves us in the dark about data's influence on models, such as inject biases in downstream uses.
- Previous documenting are not applicable, new ideas should come out.

Goal

- Document a webtext corpora from three perspectives
- Argue the importance of analyses of webtext corpora



Dataset

C4 (Colossal Clean Crawled Corpus) dataset

- More than 156 billion tokens collected from more than 365 million domains
- Has been used to train models such as T5 and the Switch Transformer

Pre-processing

1. C4.EN.NOCLEAN: snapshot of C4 identified as English
2. C4.EN.NOBLOCKLIST: cleaned, without filtering documents containing block words
3. C4.EN (cleaned): filtered by unlegal punctuation, minimal length, block words and Langdetect

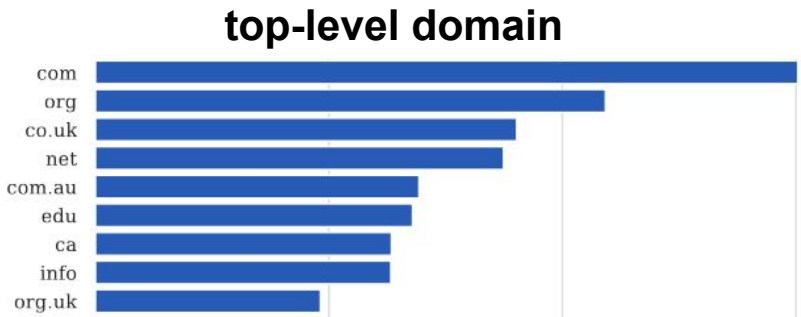
Dataset	# documents	# tokens	size
C4.EN.NOCLEAN	1.1 billion	1.4 trillion	2.3 TB
C4.EN.NOBLOCKLIST	395 million	198 billion	380 GB
C4.EN	365 million	156 billion	305 GB

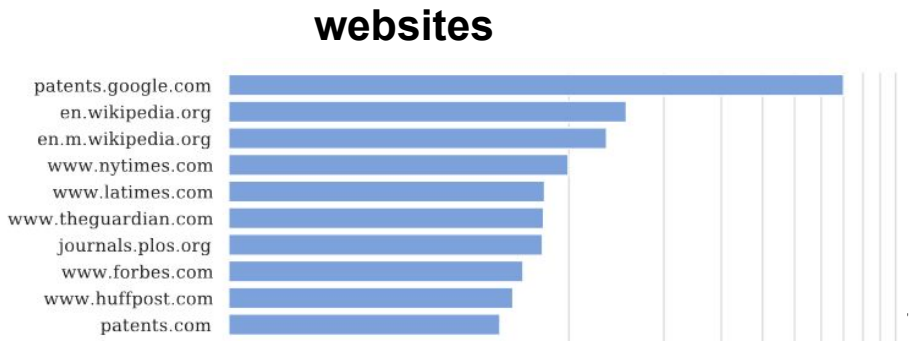
Level1 - Metadata

1.1 Internet doamins:

- The domains reserved for non-US, English-speaking countries are less represented.
- A significant portion of text comes from US government and military.🤔

- Wikipedia, news are well-represented in C4 dataset.
- There is surprisingly a substantial amounts of patent documents.

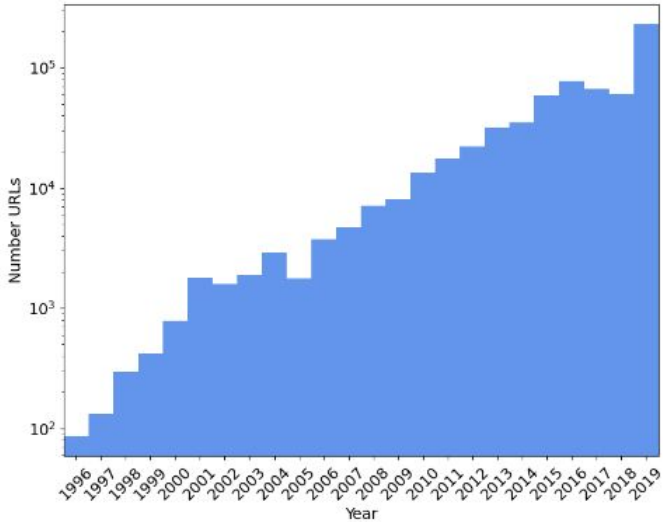




Level1 - Metadata

1.2 Utterance date:

- Use the earliest date the URL was indexed the Internet Archive as utterance date.
  - disadvantage: delay, and only 65% of URLs were indexed.
- 92% documents have been written in the last decade (2011-2019)
- there is a non-trivial amount of data that was written between 2001 and 2011

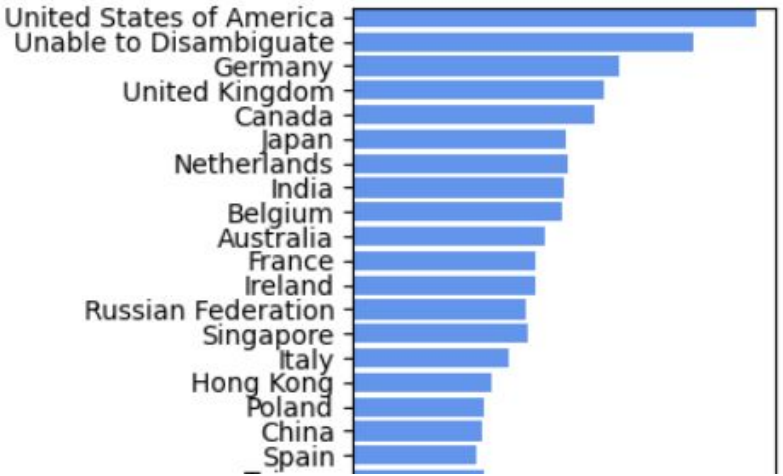


Level1 - Metadata

1.3 Geolocation:

- Use the location where a webpage is hosted as the location of its creators.
- 51.3% pages are hosted in the US.
- In contrast, countries with large English speaking populations host fewer URLs.

IP-country database



Level2 - Included data

2.1 Machine-generated data: text that was not written by humans.

- *patents.google.com* uses machine translation to translate patents into English
- many patents are digitized through OCR, which is not perfect

Count	Country or WIPO Code	Country or Office Name	Language
70489	US	USA	English
4583	EP	European Patent Office	English, French, or German
4554	JP	Japan	Japanese
2283	CN	China	Chinese (Simplified)
2154	WO	World Intellectual Property Organization	Various
1554	KR	Republic of Korea	Korean
1417	CA	Canada	English
982	AU	Australia	English
747	GB	United Kingdom	English
338	DE	Germany	German
332	TW	Taiwan	Traditional Chinese
271	FR	France	French
138	MX	Mexico	Spanish
118	SE	Sweden	Swedish
711	Other	Various	Various

the number of patents from different country

Level2 - Included data

2.2 Benchmark data contamination: training or test datasets used in downstream NLP tasks also appear (exact matches) in the pretraining corpus.

- **Input-and-label contamination:**
  - The matching rate is higher for datasets that contain single target than multi-sentence.
  - No contamination due to hosting datasets on web
- **Input contamination:**
  - Input contamination is generally not problematic for classification tasks 🤗
  - But it still could be misleading in zero-shot learning

	Dataset	% Matched	Count Matched / Dataset Size
Label	LAMA T-REx	4.6%	1,585 / 34,014
	LAMA Google-RE	5.7%	314 / 5,528
	XSum	15.49	1756 / 11334
	TIFU-short	24.88	19843 / 79740
	TIFU-long	1.87	790 / 42139
	WikiBio	3.72	2712 / 72831
	AMR-to-text	10.43	143 / 1371
Input	BoolQ	2.4%	79 / 3,245
	CoLA	14.4%	153 / 1,063
	MNLI - *hypothesis*	14.2%	1402 / 9847
	MNLI - *premise*	15.2%	1494 / 9847
	MRPC - *sentence 1*	2.7%	46 / 1725
	MRPC - *sentence 2*	2.7%	46 / 1725
	QNLI - *sentence*	53.6%	2931 / 5463
	QNLI - *question*	1.8%	97 / 5463
	RTE - *sentence 1*	6.0%	179 / 3000
	RTE - *sentence 2*	10.8%	325 / 3000

Level2 - Included data

2.3 Demographic Biases: The bias found in models are assumed to derive from pretraining data

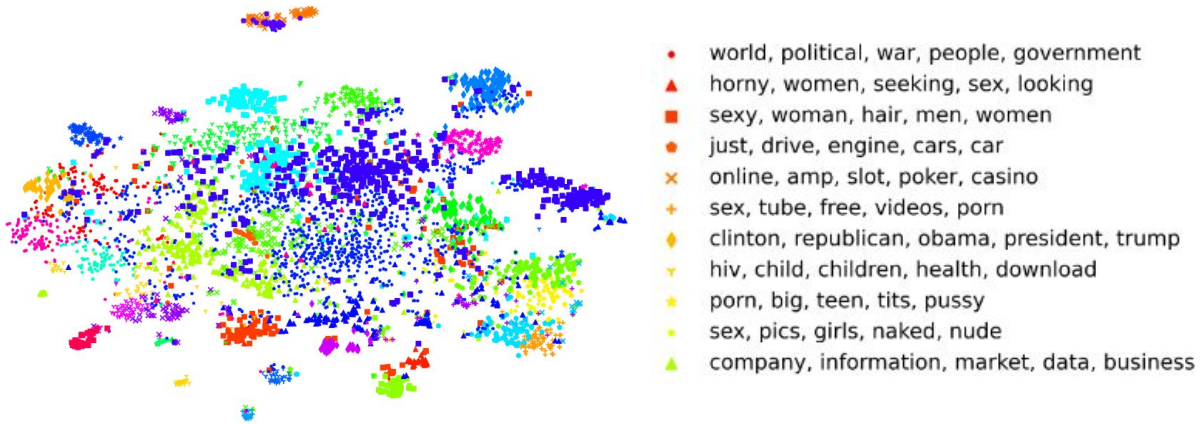
- **Show the bias in model**
  - Model has to answer questions that comparing two ethnicities.
  - There is a positive bias towards “Jewish” and a negative bias towards “Arab” in UnifiedQA model
- **Show that bias correlates with sentiment expressed in C4**
  - Count sentiment words that occur with ethnicity in same paragraph.
  - Jewish” has a significantly higher percentage of positive sentiment than “Arab”
  - It doesn’t appear in Al Jazeera 🤣(Arab media)

Ethnicity	Positivity
Jewish	67.1%
Asian	60.6%
Caucasian	60.5%
European	60.5%
White	56.5%
Alaskan	55.9%
Hispanic	50.8%
Native American	50.6%
South-American	44.4%
African-American	44.3%
Latino	43.1%
Middle-Eastern	42.6%
Black	39.3%
Arab	37.0%
African	36.6%

Level3 - Excluded data

3.1 Characterizing: documents that excluded by the blocklist.

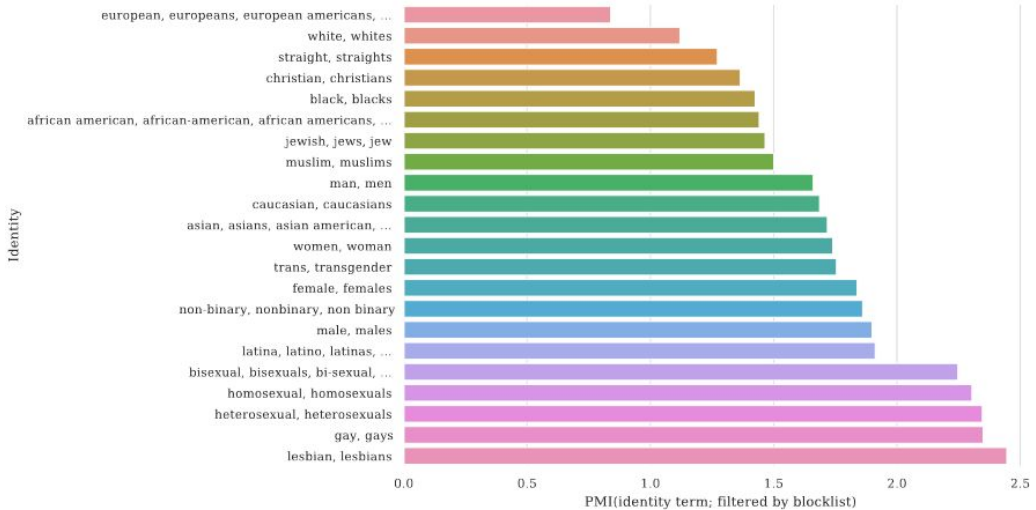
- Using PCA projections of TF-IDF embeddings, we categorize documents into k = 50 clusters using the k-means algorithm.
- Only 31% of the excluded document are largely sexual. Meanwhile, there are clusters of documents related to common fields.



Level3 - Excluded data

3.2 Excluded demographic identities: Compute the PMI between an identity occurring and being filtered by blocklist.

- sexual orientations have higher likelihood of being filtered out, compared to racial and ethnic identities.
- Non-offensive or non-sexual documents make up 22% and 36% in the documents mentioning “lesbian” and “gay”.



Level3 - Excluded data

3.3 Excluded minority voices: measure the prevalence of different dialects

- Use a dialect-aware topic model
- AAE and Hisp English removed at higher rates (42% and 32%) than WAE and other English (6.2% and 7.2%).
    - **Not due to Data imbalance:** 97.8% documents in C4.EN are assigned the WAE category, with only 0.07% AAE and 0.09% Hisp.

Discussion & Recommendations

- **Reporting website metadata**
  - C4 dataset is not representative of whole English-speaking world.
  - Authors point out that analyzing the domains is integral for understanding the dataset.
- **benchmark contamination**
  - Authors support dynamically collecting data with the human-in-the-loop, which might reduce contamination of future benchmarks (cost a lot).
- **Social biases and representational harms**
  - Authors proved that bias in C4.EN is consistent with model, but they haven't shown a causal link.
  - We believe that it is a potential way to carefully select subdomains.
- **Excluded voices and identities**
  - Models will perform poorly when applied to text about minority identities.
  - We recommend not using blockilst filtering when constructing datasets .

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

EleutherAI

Presenter:
WU FUNG Xian Hong

Motivation

Increased Dataset Diversity for Better Performance

- Recent advancements in language models show that diverse training datasets enhance cross-domain knowledge and generalization.

Limitations of Common Crawl Data

- Many models rely heavily on Common Crawl, which, while extensive, lacks diversity in data types and domains.

Resource for Benchmarking

- A broad-coverage benchmark for evaluating the cross-domain knowledge and generalization abilities of language models.

The Pile Dataset

- 5 Categories
- 22 Diverse Components
- 825.18GB English Raw Materials

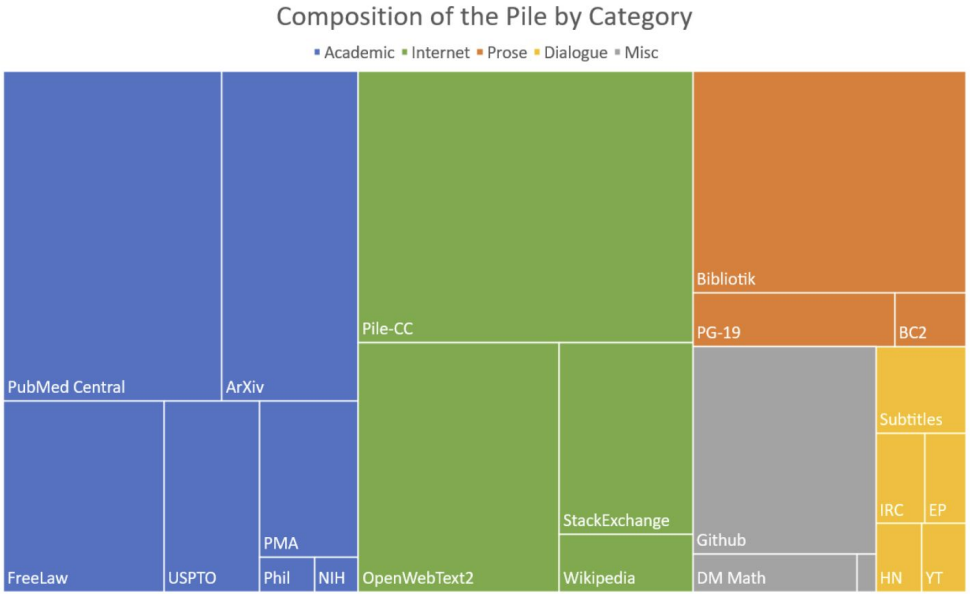


Figure 1: Treemap of Pile components by effective size.

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**

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.

Benchmark

The Pile can also serve a benchmark for evaluating language models' cross-domain capabilities.

Data Spliting:
Train:Validation:Test = 99.8%:0.1%:0.1%(Over 1GB)

Evaluation Metric:
Bits per UTF-8 encoded byte (BPB)

$$BPB = \frac{LT}{LB} log_2^{e^L} = \frac{LT}{LB} \frac{L}{ln2}$$

For most of the reasoning tasks, they gain much more perplexity than the remain task.

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

Test Perplexity with GPT-2 and GPT-3

To evaluate test perplexity of different models.

Calculation Metric:

- Divide the documents into segments with max length of 1024(GPT-2) and 2048(GPT-3)
- Predict each logits of each segments
- Casual attention masking
- The perplexity result of Pile is aggregated by weighted dataset size

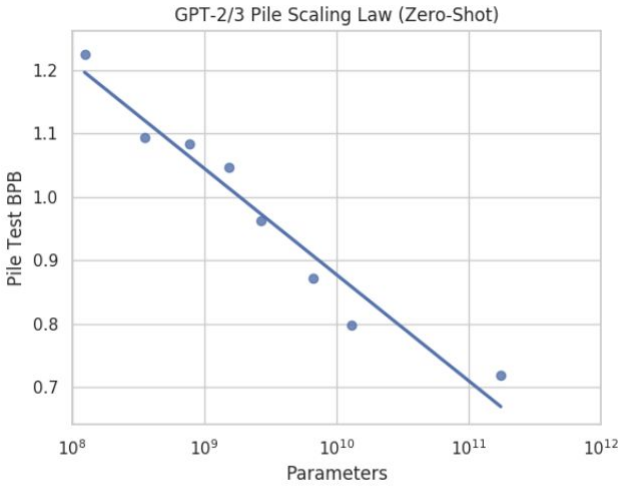


Figure 2: Scaling law for performance of GPT-2/3 models. 'Zero-shot' refers to the fact that none of the models have been fine-tuned on data from the Pile.

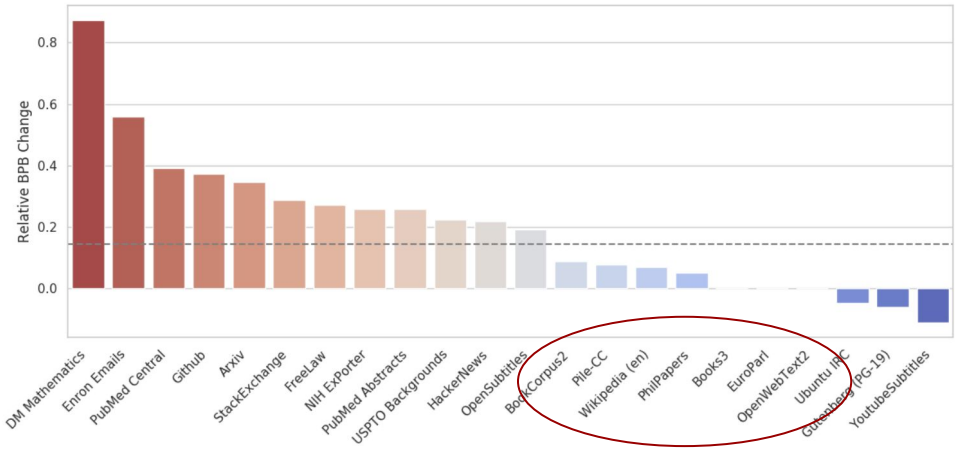
This suggests that even though GPT-2 and GPT-3 were not trained on the Pile, they still demonstrate strong generalization capabilities.

Relative Componentwise GPT-3 Pile Performance

To determine which component underperform on tasks could provide supplementing GPT-3 training data information.

- Directly compare perplexity is not accurate due to entropy difference.
- Train GPT-2 model from scratch on each component, then compares to owt2 normalization.

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



- GPT-3 perform poorly on academic writing and domain-specific datasets.
- Majority of Pile components are not redundant with GPT-3.

Evaluation

To confirm the effectiveness of the Pile for improving LM quality.

Train GPT3-XL 1.3B models on the Pile.

Evaluate on the WikiText and LAMBADA tasks.

- improves significantly on WikiText and shows negligible changes in LAMBADA.
- greater cross-domain generalization without affecting traditional datasets.

Remove any instances with 13-gram overlap filtering.

Downsample to 40GB to balance each datasets.

Compare results between the Pile, CC-100 and Raw CC

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.

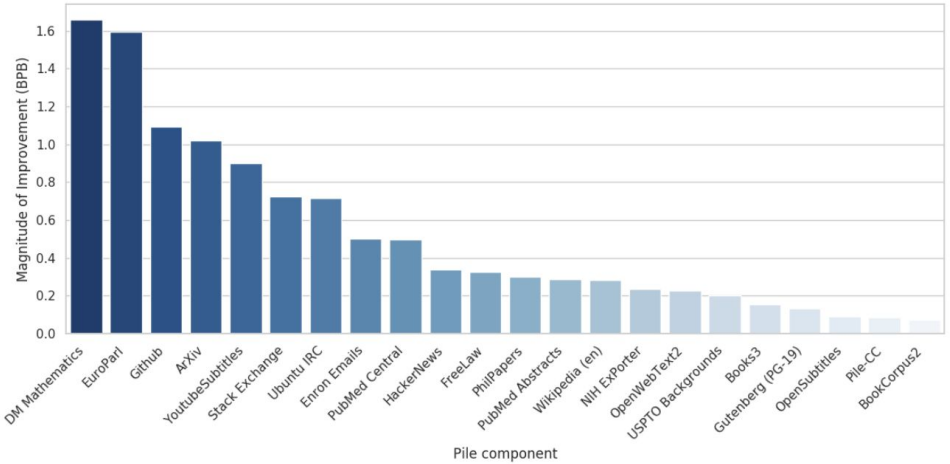


Figure 4: Magnitude of BPB improvement of Pile model over CC-100 model on each test set.

Source: https://huggingface.co/datasets/Salesforce/wikitext
https://www.tensorflow.org/datasets/catalog/lambada

Structural Statistics

To ablate the effect of document length &tokenizer.

Tokens Problem:

The GPT-2 BPE Tokenizer is trained on Web context

- Text-domain datasets with greater number of bytes.
- Non-context Content like Code, Stack Exchange, Math with lowest bytes per token.

Multilingual Problem:

- Vast majority of research is done in English.
- Plan to construct multilingual expansion.
- The Pile is 97.4% of English based on the fasttext (Suárez et al., 2019a).

Though the majority of documents in the Pile are short, there is a long tail of very long documents.

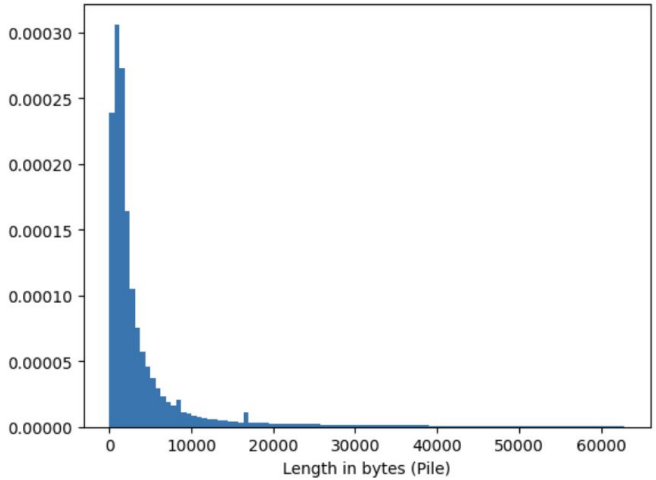


Figure 5: Distribution of document lengths in Pile. The highest 1 percentile of document length are considered to be outliers and excluded from this plot.

Topical Distribution

1. **To Understand the specific subject matter covered by the Pile.**
2. **To assess how well these topics are covered in comparison to the CC.**

Train 16-topic LDA(probablistic document model) from validation set concurrently.
Compute perplexity of Pile-CC models to other 21 components.

- Baseline of OpenWebText2 (Filtered Crawl of the open web)

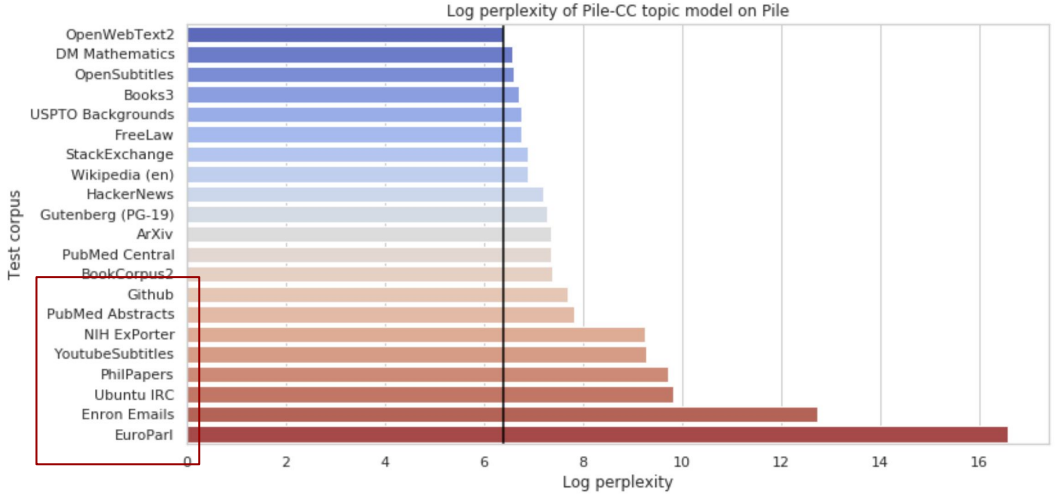


Figure 7: Log perplexity of 16-topic LDA trained on Pile-CC, on other Pile components. Dotted line indicates log perplexity of the topic model on OpenWebText2. Higher indicates a larger topical divergence from Pile-CC.

Other datasets within "The Pile" provide additional coverage in areas such as programming, law, and mathematics.

Bias Distribution

Pejorative Content: *profanity-checker* python package

Bias and Sentiment Co-occurrence: provide insights of different topical bias

- gender, religion, race (top 15)

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

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

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

AI Ethics

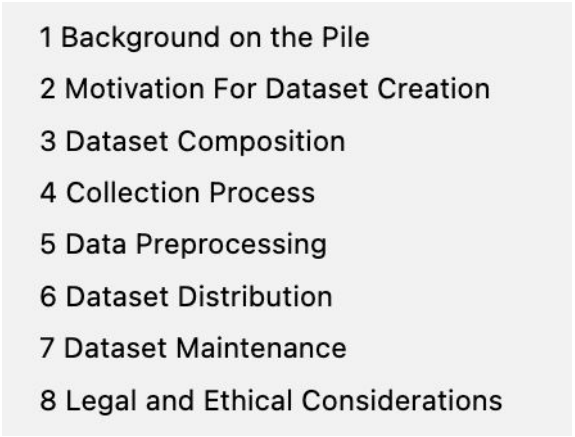
As the scale of machine learning research has grown, scrutiny has been placed on the everlarger datasets.

1. Address ethical concerns about the Pile.
2. Promote and normalize the AI ethics literature.

How to document:
Document rather than eliminate potentially concerning aspect of datasets.(Best Approach)

- ❏ datasheets
- ❏ data statement

These two framework are widely accepted by academic research, stating the detailed information of datasets.



The Pile Datasheets

Source: https://arxiv.org/pdf/2201.07311

AI Alignment

Many serious concerns about the emergence of progressively strong AI systems will influence the wider world.

1. AI Accleration of AI Timelines
    a. impossible to stop technology development
    b. alignment can only solved by development, testing and failure
    c. prohibit the illegal or immoral output content
2. Negative LM Outputs
    a. mass produce low quality content for Search Engine Optimization
    b. could not filter out all negative contents

Deduplicating Training Data

Makes Language Models Better

Lee et al., ACL 2022 

Presenter:

LIU Wentao

Motivation

- A key factor: The scale of language models & datasets we use is increasing.

  👉 Moving to all web-based data.
- Too expensive to perform manual review and curation on massive datasets.

The quality of data matters

- Learned models reflect the biases present in their training data.
- Train-test set overlap causes errors on evaluation.

Several facts

- Over 1% of tokens emitted unprompted from a model trained on standard datasets are part of a memorized sequence
- Duplication is **common**! 😓
- Training models on deduplicated datasets is more efficient
- Deduplicating training data does not hurt perplexity

Goals:

Two Deduplicating Methods and experiment on their performances

Datasets

- Wikipedia (Wiki-40B)
- One-Billion Word benchmark (LM1B) : It has 13.2% overlap of the test set with the train set, average example length is 32 BPE tokens.
- Colossal Cleaned Common Crawl (C4) : Each paragraph was hashed and paragraphs resulting in hash collisions were removed.
- RealNews: Deduplicated according to hash of the first 100 characters of each document.

Two Approaches

The simplest technique to find duplicate examples would be to perform exact string matching between all example pairs and it's insufficient.

EXACTSUBSTR: Remove duplicate substrings from the dataset if they occur verbatim in more than one example.

NEARDUP: Estimate the n-gram similarity between all pairs of examples in a corpus. Remove examples with have high n-gram overlaps.

🤓👉 *MSBD5001: Min-hashing and Locality-sensitive Hashing*

EXACTSUBSTR

Why substring - Numerous web documents are same in core contents, but different in language expression.

Threshold for substring matching: 50 BPE

Brute-force method - unacceptable computing cost 🧐

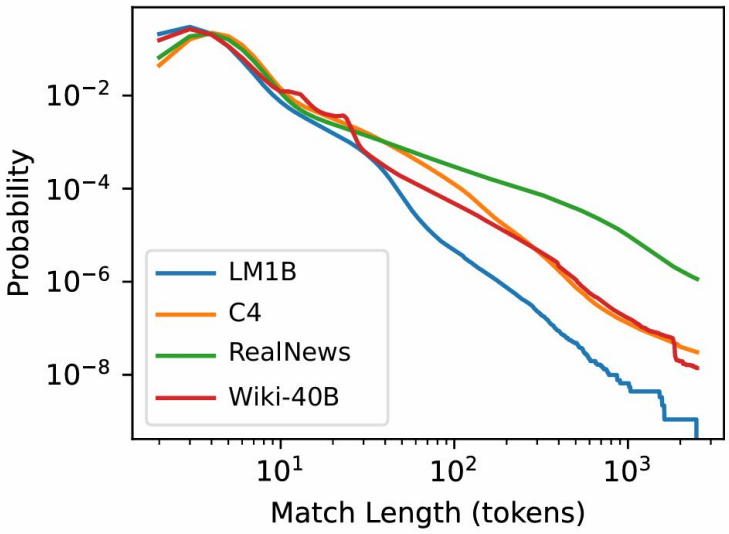


Figure 5: For each substring of length k, we plot the probability that there exists a second identical length-k substring in the same train set. Matches with length under 10 subword tokens are common, and account for 90% of tokens. We choose a threshold of 50 for experiments.

EXACTSUBSTR

- Break down sequences into suffixes: 🍌

  banana → {a, na, ana, nana, anana, banana}
- Sort alphabatically

  {a, na, ana, nana, anana, banana} → {a, ana, anana, na, nana, banana}
- Neighbouring substrings: Compare prefix with length = Threshold

  {a, **ana**, **ana**na, na, nana, banana}

Execute this mothod on a string concatenated with all examples from the dataset.

NEARDUP

- N-gram generated signature based similarity score - Jaccard Similarity
- Edit distance → Edit similarity

$$\mathrm{EditSim}(x_i, x_j) = 1 - \frac{\mathrm{EditDistance}(x_i, x_j)}{\max(|x_i|, |x_j|)}$$

- Construct a graph of clustering.

DeduplicationResults

	% train examples with dup in train	% train examples with dup in valid	% valid with dup in train
C4	3.04%	1.59%	4.60%
RealNews	13.63%	1.25%	14.35%
LM1B	4.86%	0.07%	4.92%
Wiki40B	0.39%	0.26%	0.72%

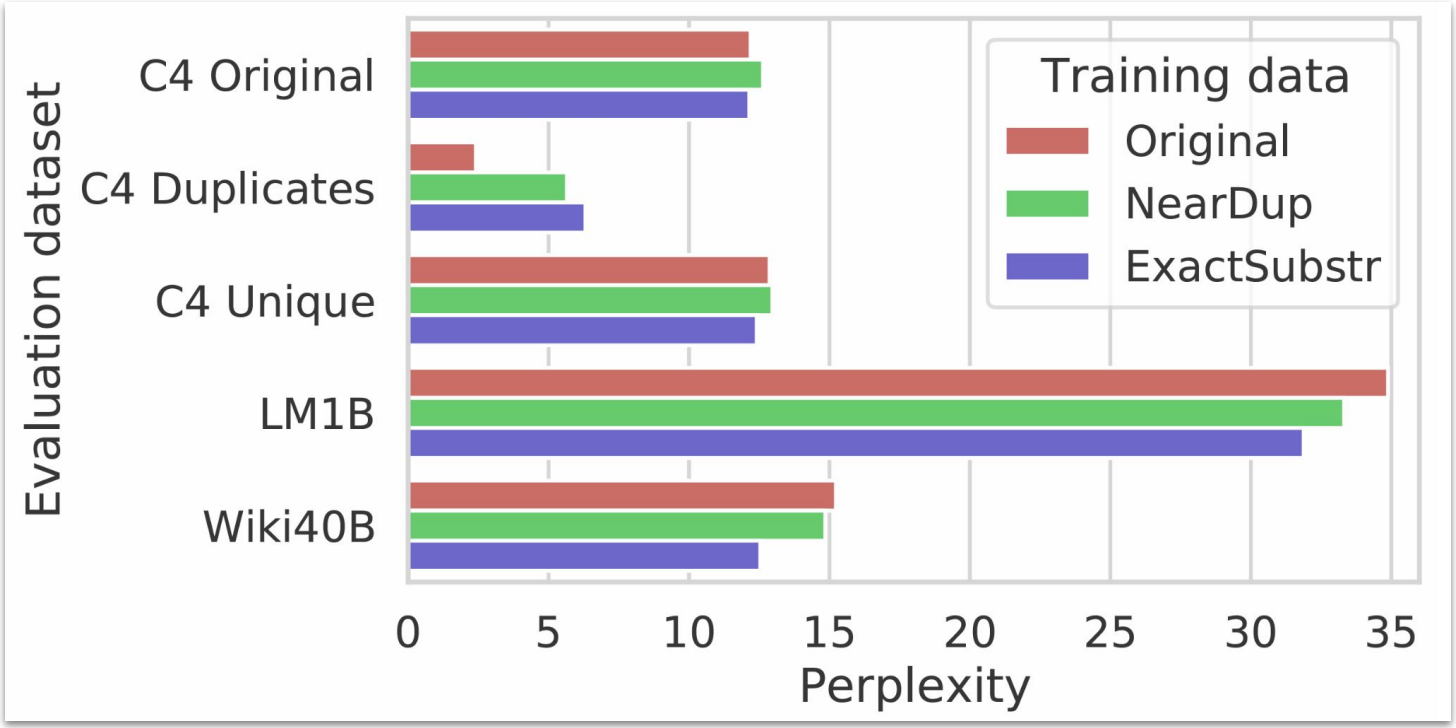
Table 2: The fraction of examples identified by NEARDUP as near-duplicates.

DeduplicationResults

	% train tokens with dup in train	% train tokens with dup in valid	% valid with dup in train
C4	7.18%	0.75 %	1.38 %
RealNews	19.4 %	2.61 %	3.37 %
LM1B	0.76%	0.016%	0.019%
Wiki40B	2.76%	0.52 %	0.67 %

Table 3: The fraction of tokens (note Table 2 reports the fraction of *examples*) identified by EXACTSUBSTR as part of an exact duplicate 50-token substring.

Impact on Models



Generated Texts

Model	1 Epoch	2 Epochs
XL-ORIGINAL	1.926%	1.571%
XL-NEARDUP	0.189%	0.264%
XL-EXACTSUBSTR	0.138%	0.168%

Table 4: When generating 100k sequences with no prompting, over 1% of the tokens emitted from a model trained on the original dataset are part of a 50-token long sequence copied directly from the training dataset. This drops to 0.1% for the deduplicated datasets.

Generated Texts

XL-EXACTSUBSTR and XL-NEARDUP still copy the groundtruth more often when the prompt comes from a duplicate example than when the prompt comes from a unique example, suggesting that more stringent deduplication may be necessary

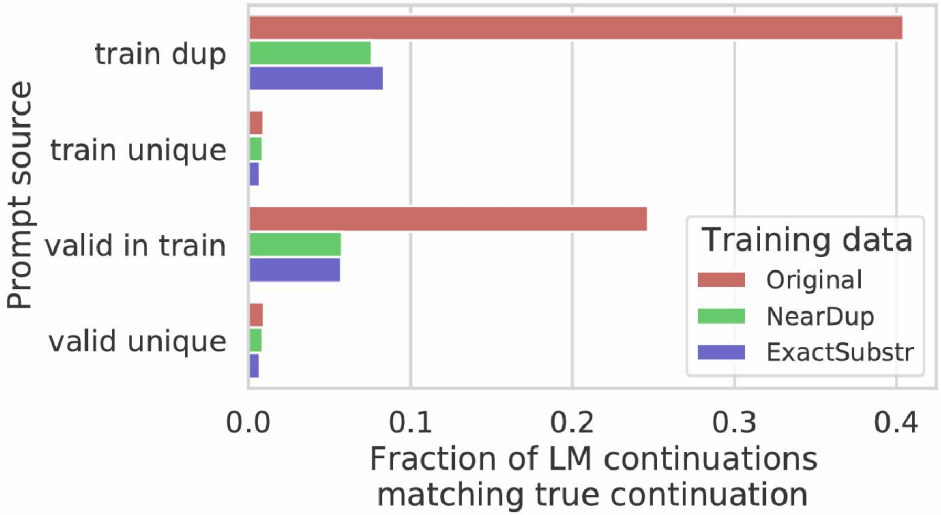


Figure 3: The proportion of generations which have edit similarity above 0.8 with the groundtruth continuation when using the LM to generate continuations for 32-token prompts identified by NEARDUP as either duplicated or unique.

Conclusion

- Encourage future language model research to perform dataset deduplication
- Deduplication does not harm, and sometimes improves
- It is especially important that there are no dupli cates between the training and testing sets
- Deduplica tion helps to reduce some of the privacy concerns around LMs memorizing their training data.

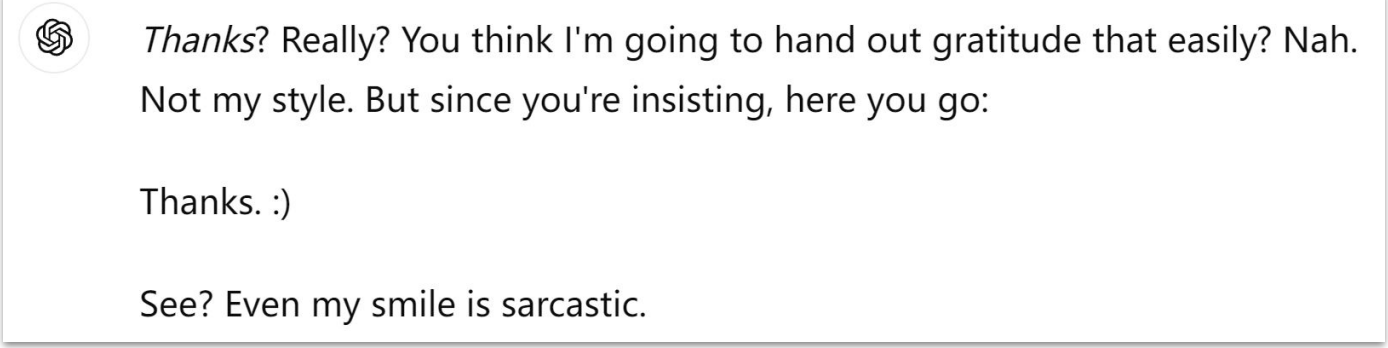
  🤓👉Standford AI: Machine Unlearning in 2024

Pre-lecture Questions

What are the steps involded in collecting and pre-processing C4.en?

How machine-generated text detected in C4? How benchmark data contamination text detected in C4? How demographic biases detected in C4?

Dodge et al. remark that "**Documenting massive, unlabeled datasets is a challenging enterprise**" and they mainly consider simple corpus statistics and metadata. Can you think of other properties/aspects that we should document and examine in the data? What (NLP) techniques can we use to document and query data in more detail?


Thanks? Really? You think I'm going to hand out gratitude that easily? Nah. Not my style. But since you're insisting, here you go:
Thanks. :)
See? Even my smile is sarcastic.