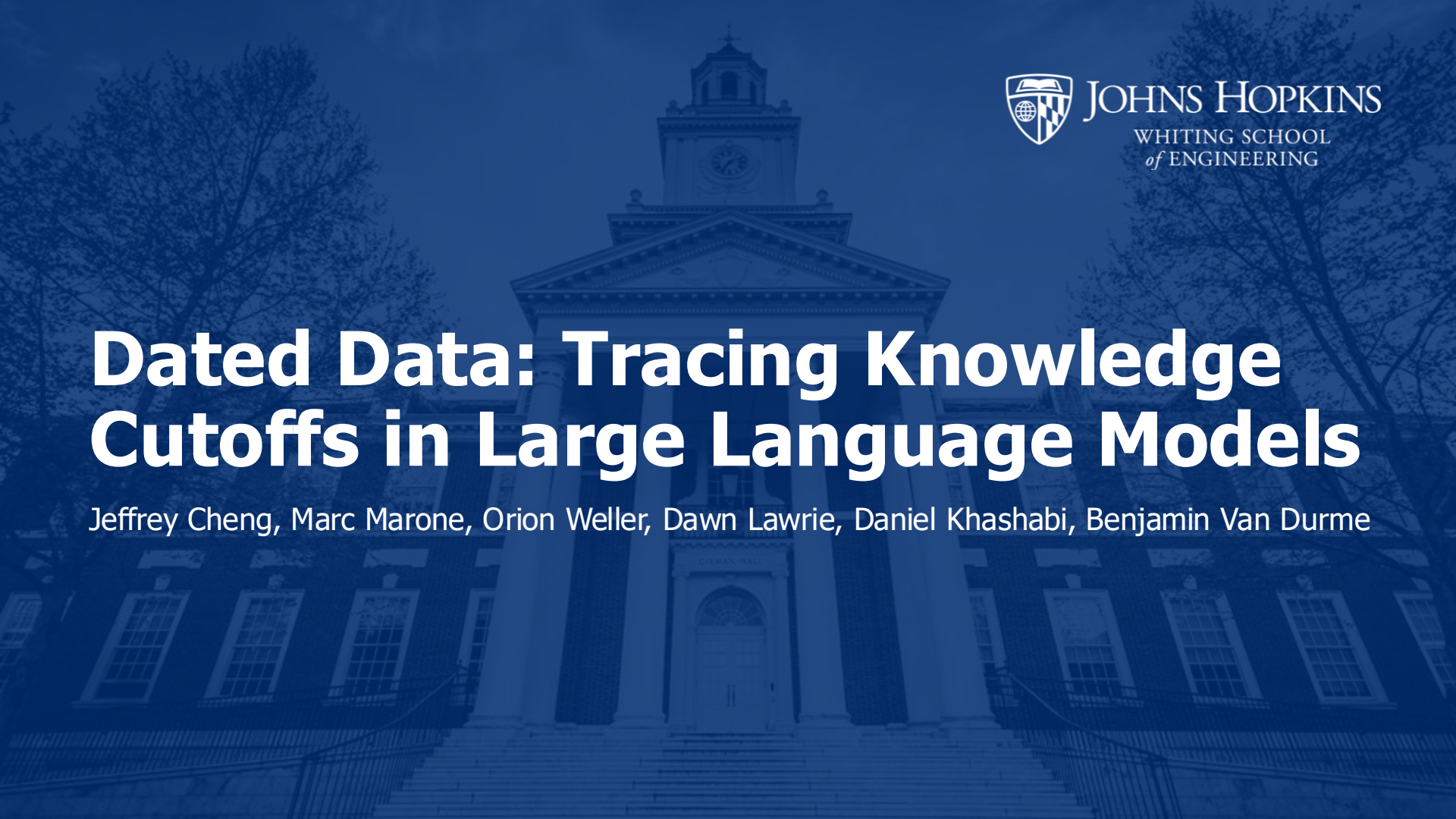


Dated Data: Tracing Knowledge Cutoffs in Large Language Models

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, Benjamin Van Durme

Knowledge Cutoffs

My knowledge was last updated in April 2024. When discussing events or developments, I approach it as a well-informed individual from April 2024 would if speaking to someone in October 2024. If you have questions about more recent events, I'll do my best to help based on trends and information available up to my knowledge cutoff, but I may not be aware of specific developments that occurred after April 2024.

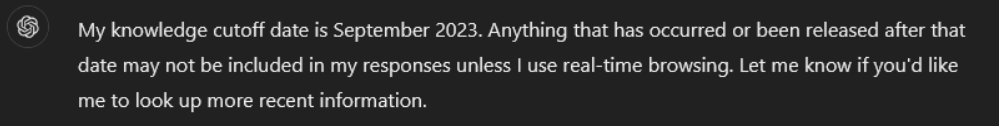
Training Data

Overview Llama 3 was pretrained on over 15 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over 10M human-annotated examples. Neither the pretraining nor the fine-tuning datasets include Meta user data.

Data Freshness The pretraining data has a cutoff of March 2023 for the 8B and December 2023 for the 70B models respectively.

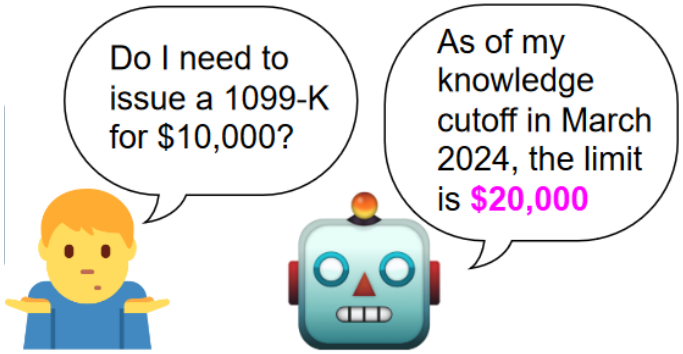
Overview: Llama 3.1 was pretrained on ~15 trillion tokens of data from publicly available sources. The fine-tuning data includes publicly available instruction datasets, as well as over 25M synthetically generated examples.

Data Freshness: The pretraining data has a cutoff of December 2023.



Key Considerations

- Do all resources in the training data share the same reported knowledge cutoff?
- Is the model's knowledge of these resources aligned to their reported cutoff date?



2022 IRS

Form 1099-K is issued for transactions only if the aggregate amount of these transactions exceeded **$20,000**

2024 IRS

Now a single transaction exceeding **$5000** can require the third party platform to issue a 1099-K.

Are there discrepancies between effective and reported cutoffs?

Key Considerations

President of the United States

116 languages

Article Talk Read View source View history Tools

From Wikipedia, the free encyclopedia

For a list of the officeholders, see List of presidents of the United States. For other uses, see President of the United States (disambiguation).

The president is elected indirectly through the Electoral College to a four-year term, along with the vice president. Under the Twenty-second Amendment, ratified in 1951, no person who has been elected to two presidential terms may be elected to a third. In addition, nine vice presidents have become president by virtue of a president's intra-term death or resignation.[C] In all, 45 individuals have served 46 presidencies spanning 58 four-year terms.[D] Joe Biden is the 46th and current president, having assumed office on January 20, 2021.

Through the Electoral College, registered voters indirectly elect the president and vice president to a four-year term. This is the only federal election in the United States which is not decided by popular vote.[19] Nine vice presidents became president by virtue of a president's intra-term death or resignation.[C]

Donald Trump of New York is the 45th and current president of the United States. He assumed office on January 20, 2017.

Time Spanning Datasets

Updating Resource

Building Resource

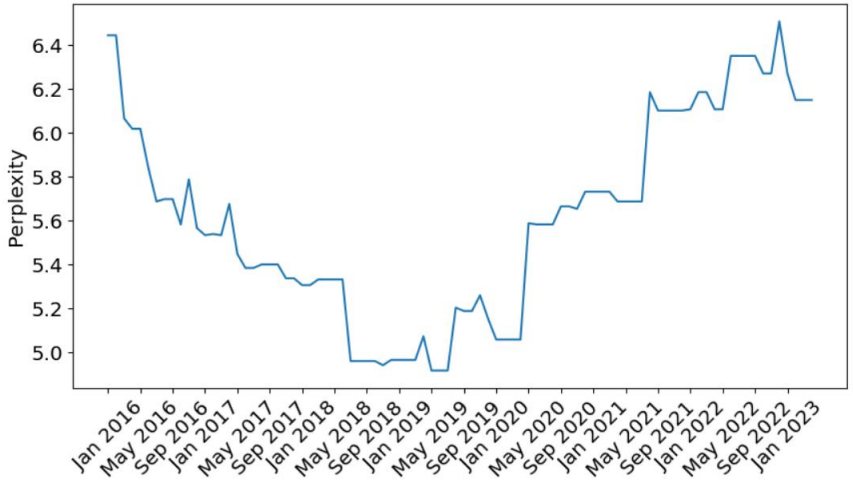
Static Resource

- WIKISPAN:
  - Collect 5000 most edited topics
  - Scrape monthly versions from April 2016 to April 2023

Probing Methodology

- WIKISPAN documents: version of Wikipedia topic *t* at time *m*
- Measure perplexity of first 512 tokens of each document, across all topics and months

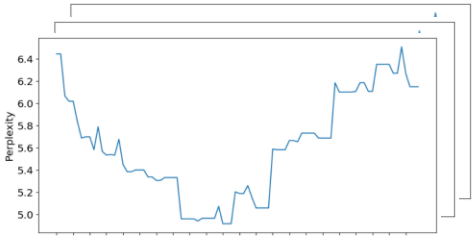
$$PPL(X) = \exp\left(-\frac{1}{t}\sum_{i=0}^{t} p_\theta(x_i|x_{<i})\right)$$



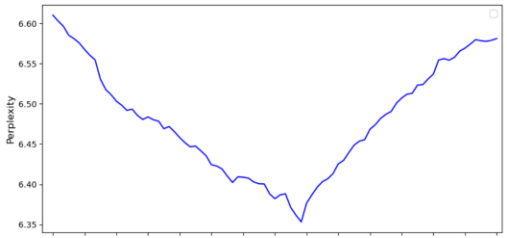
Perplexity of the Wikipedia document "Liverpool" under Pythia-7b. Each point is the perplexity of the document at that time.

Probing Methodology

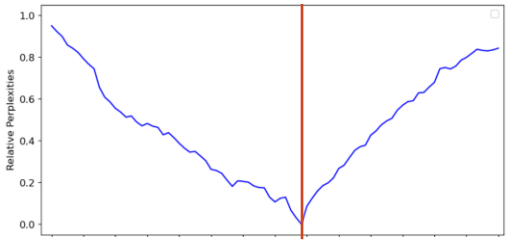
- Normalization
  - Aggregate perplexities with 95% truncated mean within each month
  - Perform 0-1 normalization over entire time-span
- Effective knowledge cutoffs are the argmin of relative perplexity curves



Perplexity measurements of documents

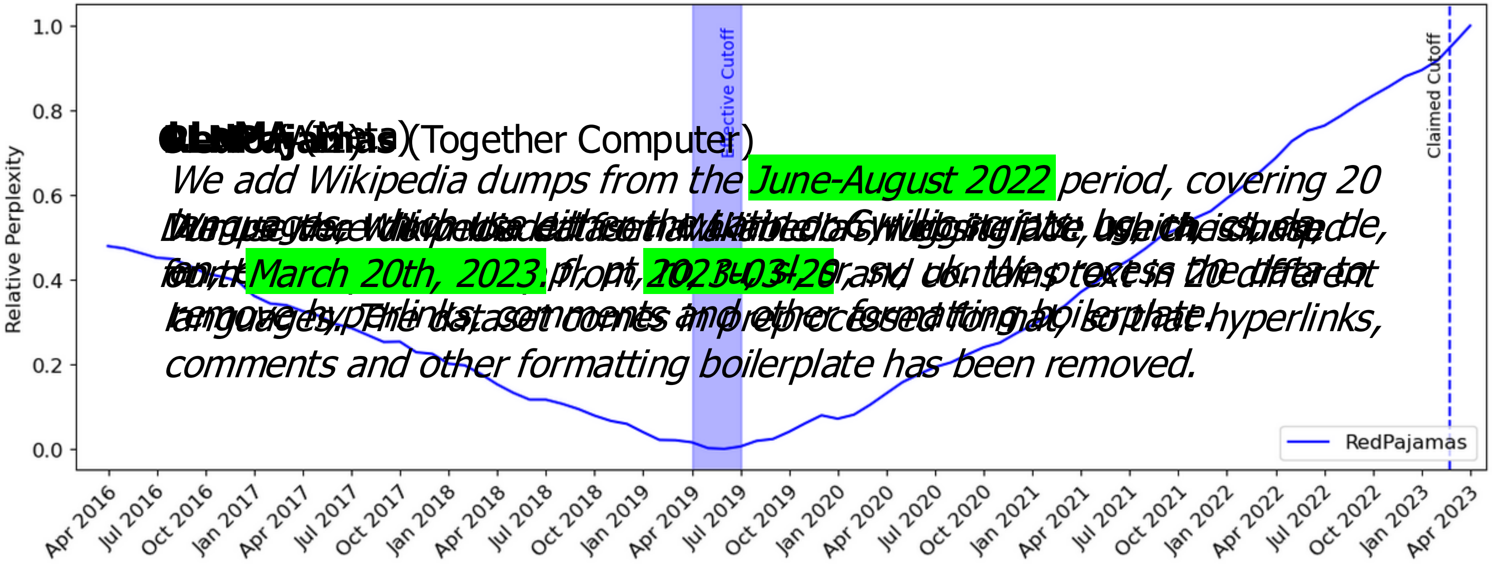


Aggregate along month axis with truncated mean



Convert to relative perplexities by 0-1 scaling

Wikipedia Results – C4 Derived Models



RedPajamas (Together Computer)

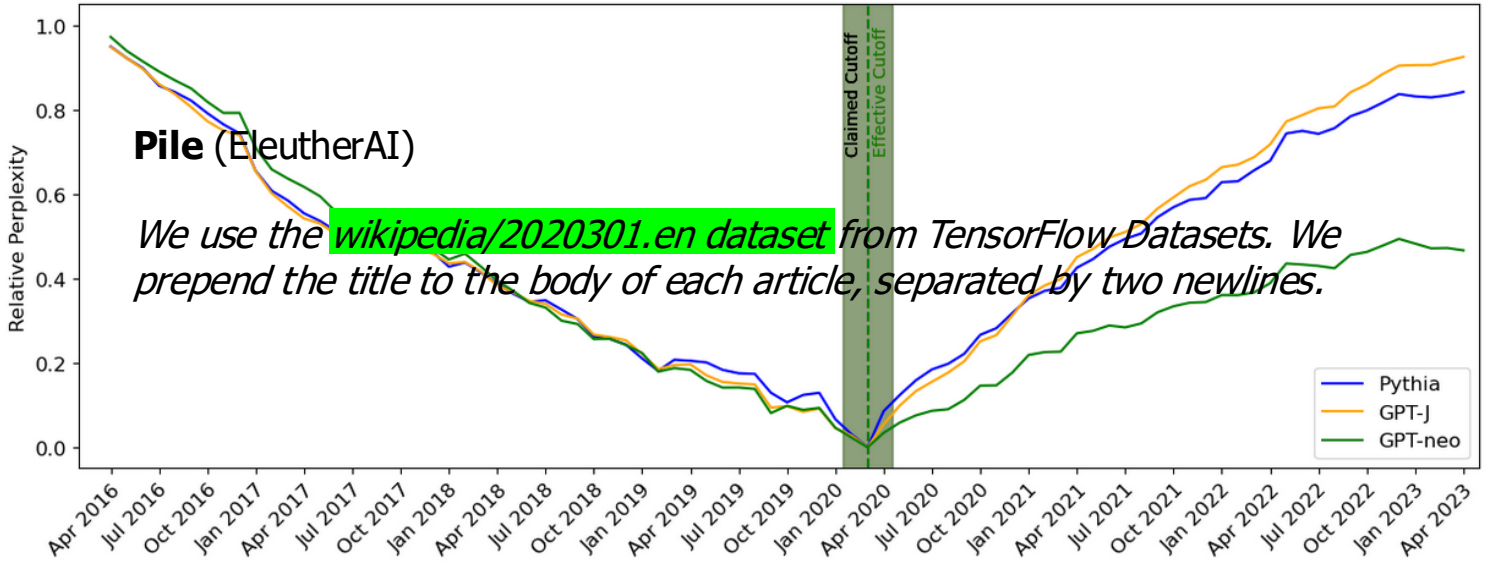
We add Wikipedia dumps from the June-August 2022 period, covering 20

March 20th, 2023

... from 2023-03-20 and contains text in 20 different languages. The dataset comes in preprocessed format, so that hyperlinks, comments and other formatting boilerplate has been removed.

JOHNS HOPKINS WHITING SCHOOL of ENGINEERING

Wikipedia Results – Pile Derived Models



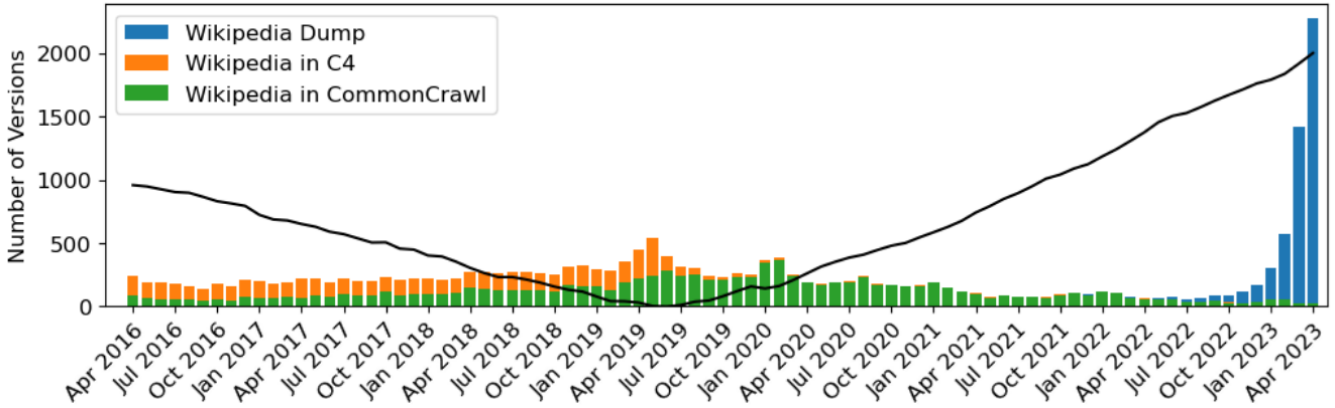
Why do there exist discrepancies between effective and reported cutoffs?

Effective and Reported Cutoff Misalignment

- CommonCrawl Misalignments
- Complications in Deduplication Pipelines

CommonCrawl Misalignments

- Many LLMs train on CommonCrawl dumps
- Breakdown of RedPajamas training corpus:



CommonCrawl dumps contain old data that bias effective cutoffs.

Effective and Reported Cutoff Misalignment

- CommonCrawl Misalignments
- Complications in Deduplication Pipelines

Deduplication Issues

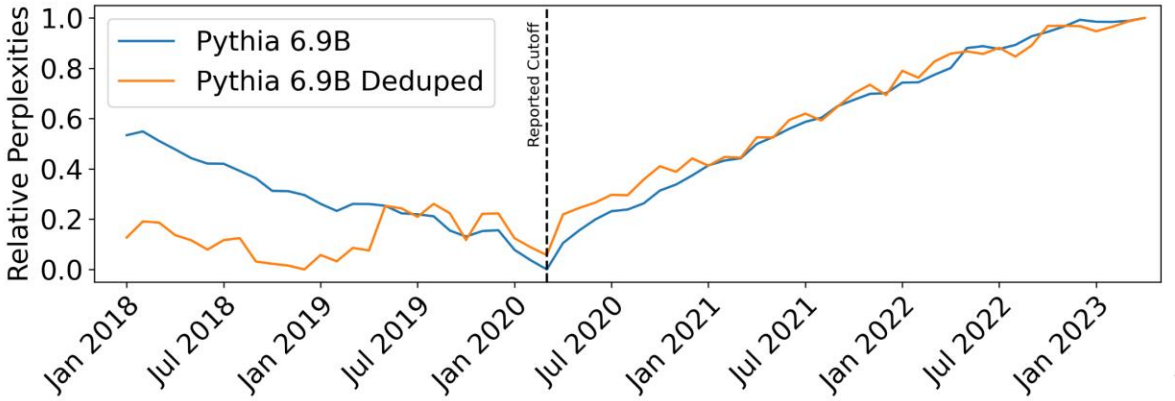
- It is common practice to deduplicate pre-training datasets
  - Fuzzy deduplication should remove different versions of Wikipedia documents
  - Exact deduplication should remove exact copies of Wikipedia document
- We empirically find many duplicates in pretraining datasets!

By the end of the 17th century, the Chinese economy had recovered from the devastation caused by the wars in which the Ming dynasty were overthrown, and the resulting breakdown of order.[147] In the following century, markets continued to expand as in the late Ming period, but with more trade between regions, a greater dependence on overseas markets and a greatly increased population.[148][149] The government broadened land ownership by returning land that had been sold to large landowners in the late Ming period by families unable to pay the land tax.[150] To give people more incentives to participate in the market, they reduced the tax burden in comparison with the late Ming, and replaced the corvée system with a head tax used to hire laborers.[151] The administration of the Grand Canal was made more efficient, and transport opened to private merchants.[152] A system of monitoring grain prices eliminated severe shortages, and enabled the price of rice to rise slowly and smoothly through the 18th century.[153] Wary of the power of wealthy merchants, Qing rulers limited their trading licenses and usually refused them permission to open new mines, except in poor areas …

By the end of the 17th century, the Chinese economy had recovered from the devastation caused by the wars in which the Ming dynasty were overthrown, and the resulting breakdown of order.[148] In the following century, markets continued to expand as in the late Ming period, but with more trade between regions, a greater dependence on overseas markets and a greatly increased population.[149][150] The government broadened land ownership by returning land that had been sold to large landowners in the late Ming period by families unable to pay the land tax.[151] To give people more incentives to participate in the market, they reduced the tax burden in comparison with the late Ming, and replaced the corvée system with a head tax used to hire laborers.[152] The administration of the Grand Canal was made more efficient, and transport opened to private merchants.[153] A system of monitoring grain prices eliminated severe shortages, and enabled the price of rice to rise slowly and smoothly through the 18th century.[154] Wary of the power of wealthy merchants, Qing rulers limited their trading licenses and usually refused them permission to open new mines, except in poor areas …

Deduplication Issues

- The Pile purposely upsamples Wikipedia documents at reported cutoff



Duplicate documents bias effective cutoff towards their version date.

Conclusions

- There **do** exist discrepancies between effective and reported knowledge cutoffs in modern LLMs

- Effective cutoffs of modern LLMs are years earlier than reported cutoff
  - CommonCrawl dumps include older versions of resources
  - Old versions and their duplicates are not removed by deduplication pipelines

- Effective cutoffs of Pile-derived models matches their reported cutoff
  - Small amount of CommonCrawl used (< 25% of one CC dump)
  - 3x upsampling of versions at reported cutoff date

Thank you!

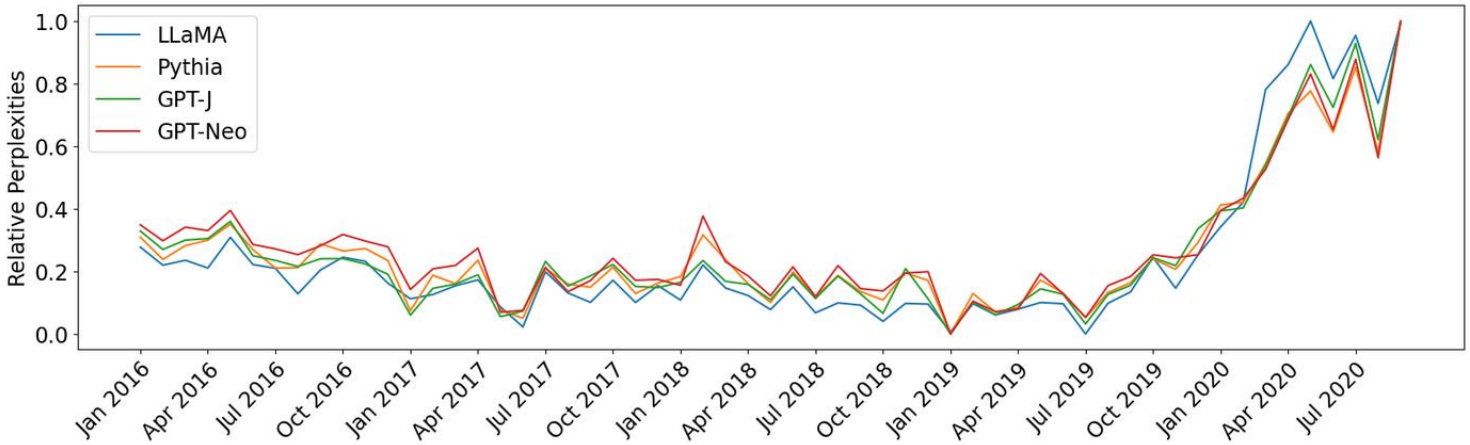
- I am applying for PhD programs this cycle!

Pretraining datasets

- The Pile
  - Open access curated dataset with 22 sub-datasets (arXiv, Wikipedia, etc.)
  - Contains 22 random chunks from the 3679 extracted from 7 years of CommonCrawl Dumps (2013 – 2020)
  - Contains Wikipedia dump from March 2020

- C4
  - A single heavily processed CommonCrawl Dump from April 2019

NYTimes


LLaMA
Pythia
GPT-J
GPT-Neo
Relative Perplexities
1.0
0.8
0.6
0.4
0.2
0.0
Jan 2016
Apr 2016
Jul 2016
Oct 2016
Jan 2017
Apr 2017
Jul 2017
Oct 2017
Jan 2018
Apr 2018
Jul 2018
Oct 2018
Jan 2019
Apr 2019
Jul 2019
Oct 2019
Jan 2020
Apr 2020
Jul 2020