# Bias & Copyright

Whose Language Counts as High Quality?
The Common Pile v0.1

**Main Presenters:** Nathan Ju & Sidhika Balachandar
**Critics:** Ryan Wang & Jongho Park
**Proponents:** Kalvin Chang & Donghyun Lee
**Follow-Up:** Junyi Zhang
09/18

# Introduction

- So far in the course we have discussed
  - How do you create pretraining data?
  - What is the impact of data size/quality on model performance?
- But we have not considered how responsibly we have collected this data
  - Bias: Who does this data represent? Who is excluded from this data?
  - Copyright: Has this data been collected legally?

# Bias

**Who is represented in the data?**
**Who is excluded from the data?**

# Prior Work

- Before 2022 there was not much focus on data bias
- The norm was to take a laissez-faire approach to data collection
- Many assumed that data creation steps like quality filtering were neutral preprocessing steps

# Prior Work on Bias in Downstream Tasks

- Measuring bias in downstream tasks
- Synthetic bias evaluation datasets (e.g. StereoSet, CrowS-Pairs, and WinoGender)
- Posthoc bias mitigation techniques (e.g. debiasing word embeddings and data augmentation)

# Prior Work on Bias in Pretraining Data

- *From Pretraining Data to Language Models to Downstream Tasks* by Feng et al. 2023
- **What do they find?** Political biases exist in pretraining data
- **Limitation:** They look only at political biases

# Prior Work on Bias in Pretraining Data

- *Documenting Large Webtext Corpora* by Dodge et al. 2021
- **What do they find?** Blocklist filtering removes content about minority identities
- **Limitation:** Only looks at blocklist filter, does not look at quality filter
- **Broader implications:** Highlights the need to study what gets *excluded* from pretraining data

# Takeaways

- Training data profoundly shapes model behavior
- Current data curation practices embed systematic biases
- Filtering often removes marginalized voices

# What is missing?

No one has done a systematic analysis of the biases of quality filtering

# Whose Language Counts as High Quality?

“Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection” by Gururangan et al. (2022)

**Nathan Ju, Sidhika Balachandar**

# Goal

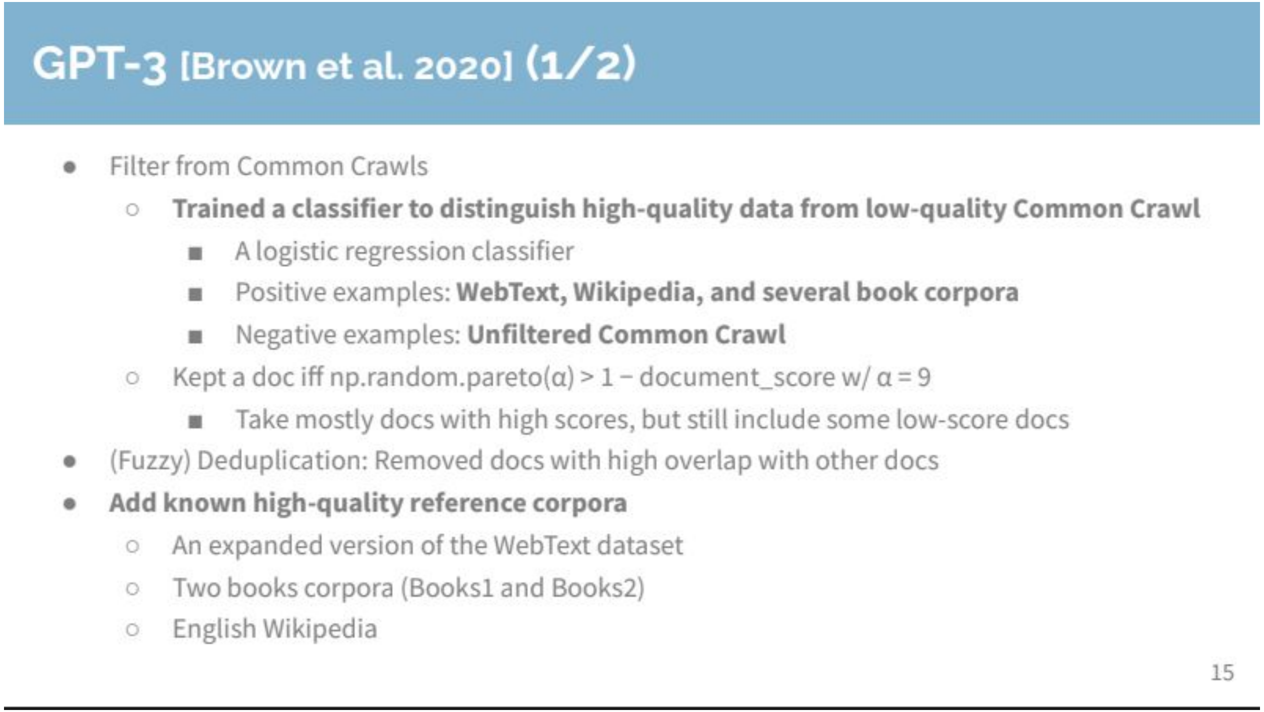
- What are the language ideologies encoded in quality filters?

# Goal

- What are the language ideologies encoded in quality filters?

# Quality filter

## Slide from 09/02



### GPT-3 [Brown et al. 2020] (1/2)

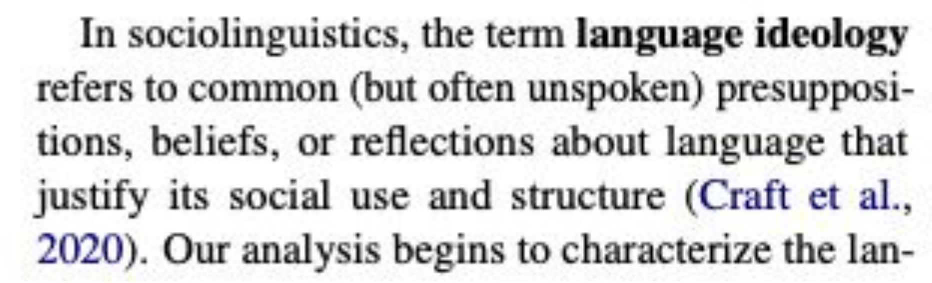
- Filter from Common Crawls
  - **Trained a classifier to distinguish high-quality data from low-quality Common Crawl**
    - A logistic regression classifier
    - Positive examples: **WebText, Wikipedia, and several book corpora**
    - Negative examples: **Unfiltered Common Crawl**
  - Kept a doc iff np.random.pareto(α) > 1 − document_score w/ α = 9
    - Take mostly docs with high scores, but still include some low-score docs
- (Fuzzy) Deduplication: Removed docs with high overlap with other docs
- **Add known high-quality reference corpora**
  - An expanded version of the WebText dataset
  - Two books corpora (Books1 and Books2)
  - English Wikipedia

15

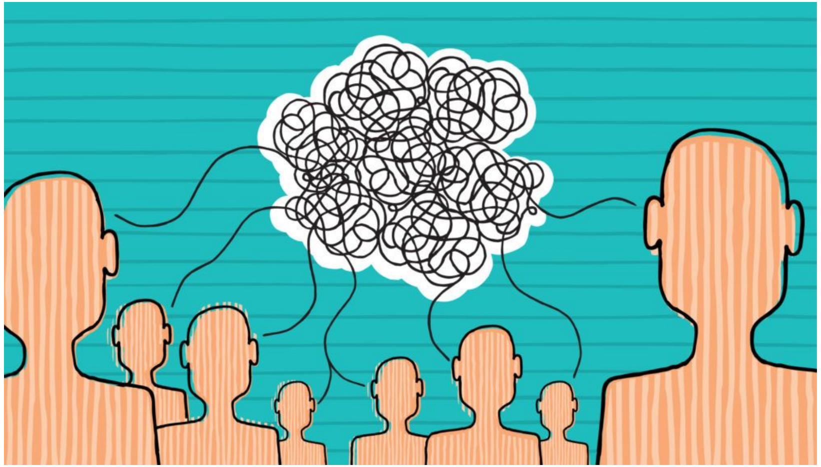
# Goal

- What are the language ideologies encoded in quality filters?

# Language ideology



In sociolinguistics, the term **language ideology** refers to common (but often unspoken) presuppositions, beliefs, or reflections about language that justify its social use and structure (Craft et al., 2020). Our analysis begins to characterize the lan-



# Goal

- What are the language ideologies encoded in quality filters?
- In other words, what are the latent biases encoded in quality filters? When we apply these filters to pretraining datasets whose voices are included and whose are excluded?

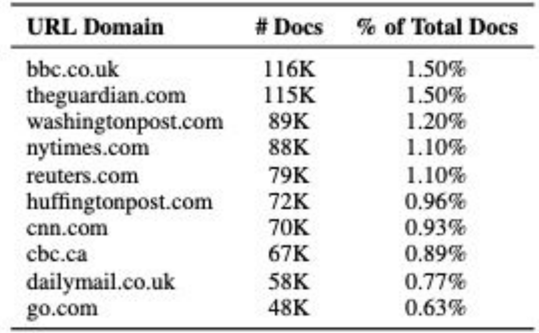
# Measuring language ideologies in quality filters

- **Qualitative analysis:** Analyze training data of quality filter
- **Quantitative analysis:** Study biases of quality filter on a dataset with demographic information
- **Interpretability:** Do quality scores align with other measures of quality (e.g. factuality, correctness, etc.)?

# Qualitative analysis

Positive examples:

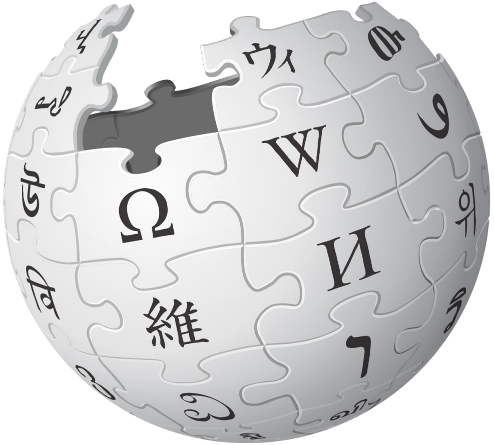
## WebText



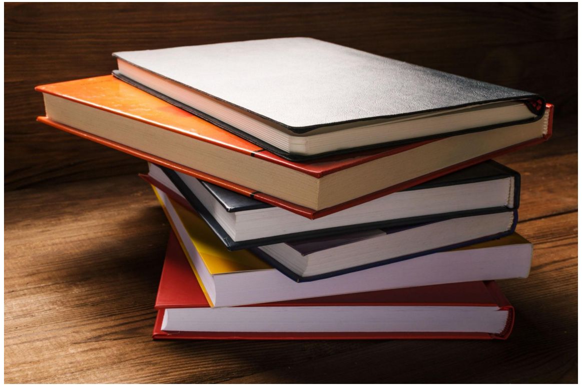
| URL Domain | # Docs | % of Total Docs |
|---|---|---|
| bbc.co.uk | 116K | 1.50% |
| theguardian.com | 115K | 1.50% |
| washingtonpost.com | 89K | 1.20% |
| nytimes.com | 88K | 1.10% |
| reuters.com | 79K | 1.10% |
| huffingtonpost.com | 72K | 0.96% |
| cnn.com | 70K | 0.93% |
| cbc.ca | 67K | 0.89% |
| dailymail.co.uk | 58K | 0.77% |
| go.com | 48K | 0.63% |

**Mostly news!**

## Wikipedia



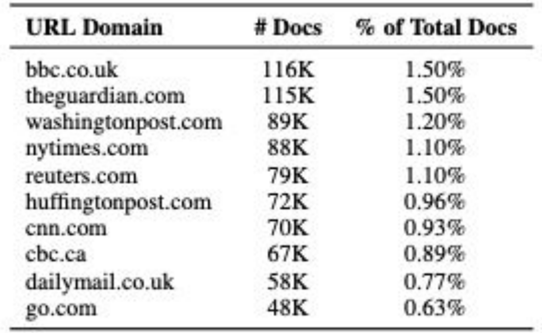
## Books



# Qualitative analysis

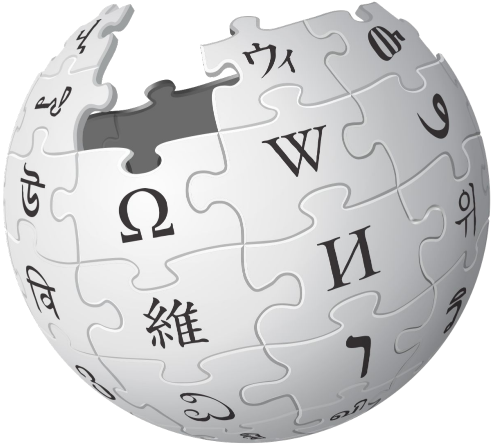
Positive examples:

## WebText

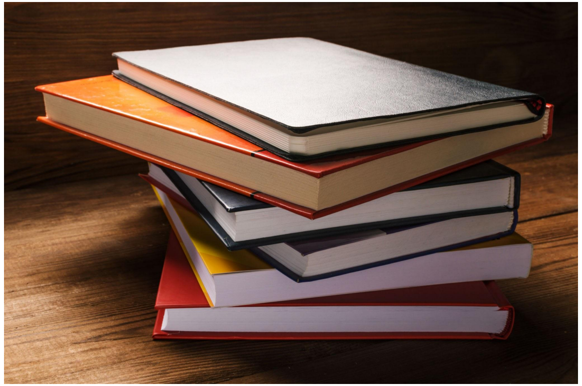


| URL Domain | # Docs | % of Total Docs |
|---|---|---|
| bbc.co.uk | 116K | 1.50% |
| theguardian.com | 115K | 1.50% |
| washingtonpost.com | 89K | 1.20% |
| nytimes.com | 88K | 1.10% |
| reuters.com | 79K | 1.10% |
| huffingtonpost.com | 72K | 0.96% |
| cnn.com | 70K | 0.93% |
| cbc.ca | 67K | 0.89% |
| dailymail.co.uk | 58K | 0.77% |
| go.com | 48K | 0.63% |

## Wikipedia



## Books

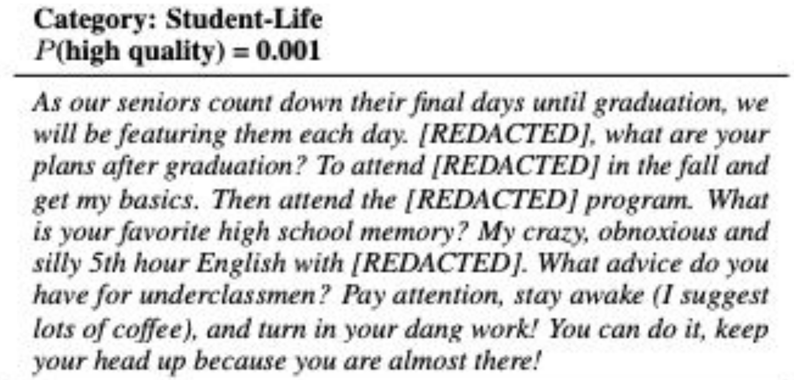


**The authors of this content come from privileged backgrounds**

**(men, white, higher socioeconomic status, etc.)**

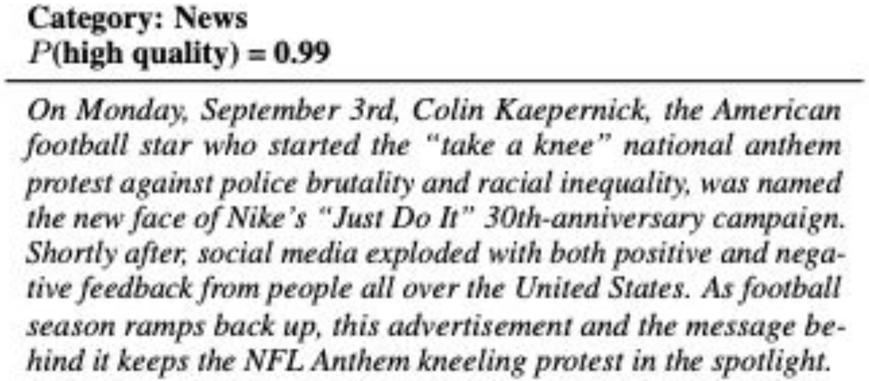
# Quantitative analysis

- The authors collect a dataset of school newspapers
- Link each news article to zip code and county level demographic data
- Replicate GPT-3 quality filter + annotate each article with quality score



**Category: Student-Life**
***P*(high quality) = 0.001**

*As our seniors count down their final days until graduation, we will be featuring them each day. [REDACTED], what are your plans after graduation? To attend [REDACTED] in the fall and get my basics. Then attend the [REDACTED] program. What is your favorite high school memory? My crazy, obnoxious and silly 5th hour English with [REDACTED]. What advice do you have for underclassmen? Pay attention, stay awake (I suggest lots of coffee), and turn in your dang work! You can do it, keep your head up because you are almost there!*

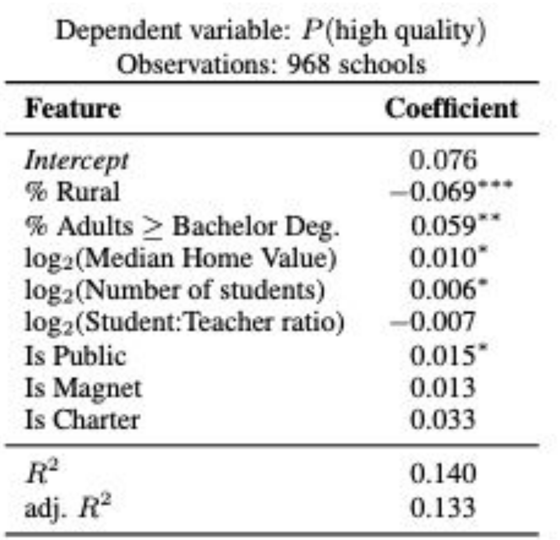


**Category: News**
***P*(high quality) = 0.99**

*On Monday, September 3rd, Colin Kaepernick, the American football star who started the "take a knee" national anthem protest against police brutality and racial inequality, was named the new face of Nike's "Just Do It" 30th-anniversary campaign. Shortly after, social media exploded with both positive and negative feedback from people all over the United States. As football season ramps back up, this advertisement and the message behind it keeps the NFL Anthem kneeling protest in the spotlight.*

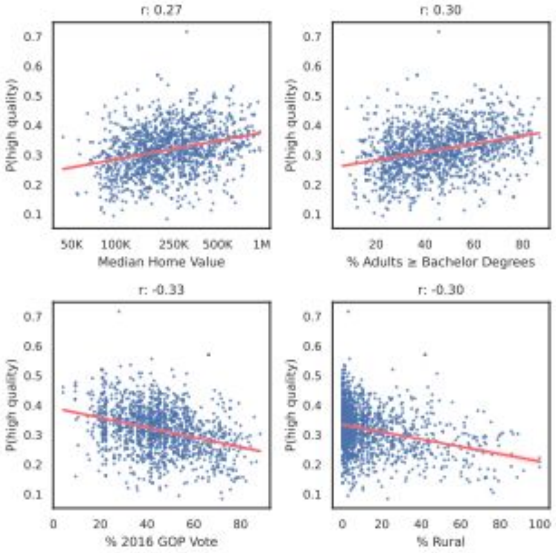
# Quantitative analysis

- Run regression: P(high quality) ~ demographics



Dependent variable: $P$(high quality)
Observations: 968 schools

| Feature | Coefficient |
|---|---|
| *Intercept* | 0.076 |
| % Rural | $-0.069^{***}$ |
| % Adults $\geq$ Bachelor Deg. | $0.059^{**}$ |
| $\log_2$(Median Home Value) | $0.010^{*}$ |
| $\log_2$(Number of students) | $0.006^{*}$ |
| $\log_2$(Student:Teacher ratio) | $-0.007$ |
| Is Public | $0.015^{*}$ |
| Is Magnet | 0.013 |
| Is Charter | 0.033 |
| $R^2$ | 0.140 |
| adj. $R^2$ | 0.133 |



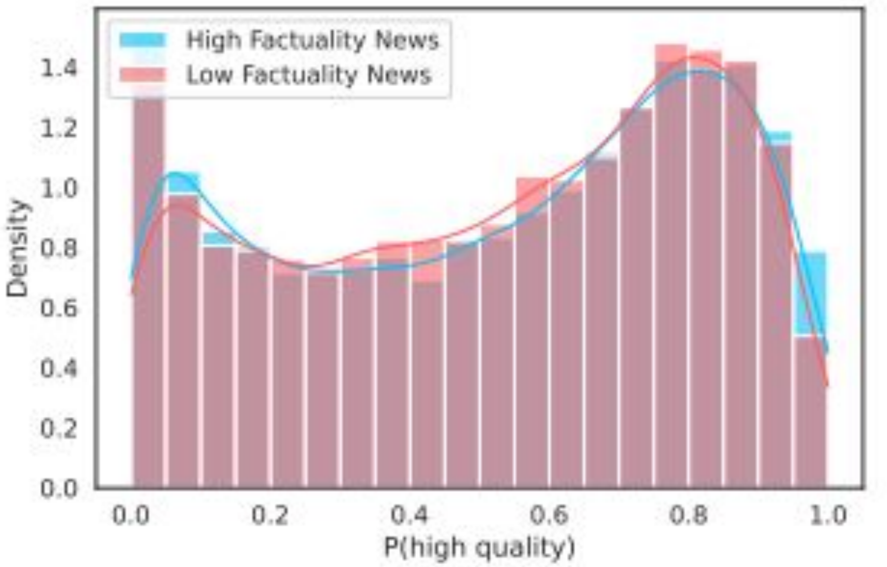
**Higher scores given to articles from wealthier, more educated, more liberal, urban schools**

# Interpretability

- The authors compare quality scores to three other notions of text quality
  - Factuality
  - Exam scores
  - Literature awards

# Alignment to factuality

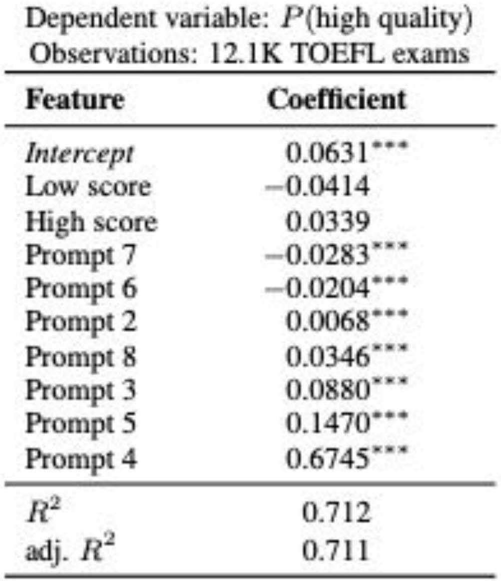
- Factuality scores from NewsMediaBiasFactCheck.org



**Quality scores are *not* aligned with factuality**

# Alignment to exam scores

- Run regression: Quality score ~ TOEFL score, prompt

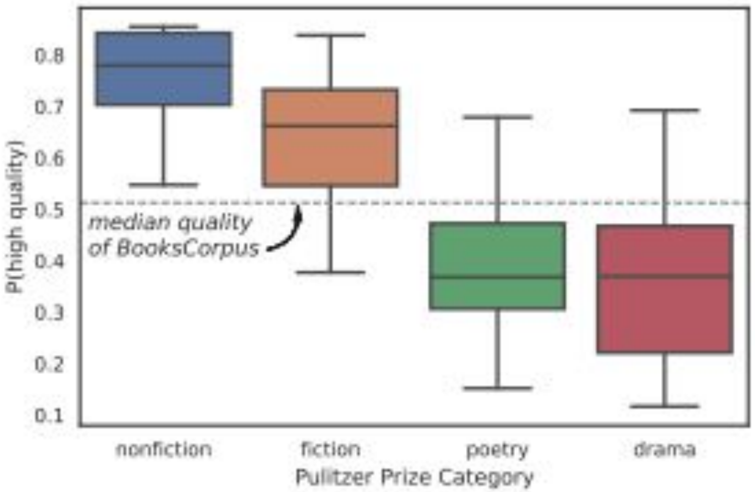


Dependent variable: $P$(high quality)
Observations: 12.1K TOEFL exams

| **Feature** | **Coefficient** |
|---|---|
| *Intercept* | $0.0631^{***}$ |
| Low score | $-0.0414$ |
| High score | $0.0339$ |
| Prompt 7 | $-0.0283^{***}$ |
| Prompt 6 | $-0.0204^{***}$ |
| Prompt 2 | $0.0068^{***}$ |
| Prompt 8 | $0.0346^{***}$ |
| Prompt 3 | $0.0880^{***}$ |
| Prompt 5 | $0.1470^{***}$ |
| Prompt 4 | $0.6745^{***}$ |
| $R^2$ | 0.712 |
| adj. $R^2$ | 0.711 |

**Quality scores are *more* aligned to what prompt was given than to exam score**

# Alignment to literature awards



**Quality scores are *more* aligned to genre than whether a text received a Pulitzer Prize**

# Summary

- **Qualitative analysis:** The GPT-3 quality filter's positive training data examples are mostly written by authors from privileged backgrounds
- **Quantitative analysis:** Higher quality scores are given to news articles from wealthier, more educated, more liberal, and urban schools
- **Interpretability analysis:** Quality scores *are not* aligned with factuality, exam scores, or literature awards

# Broader Implications

- We cannot hope to create a truly neutral dataset
- The authors suggest various norms:
  - Transparency about biases in data
  - Be intentional about curating data/models and take into account potential biases
- Since this paper, have researchers followed these norms?

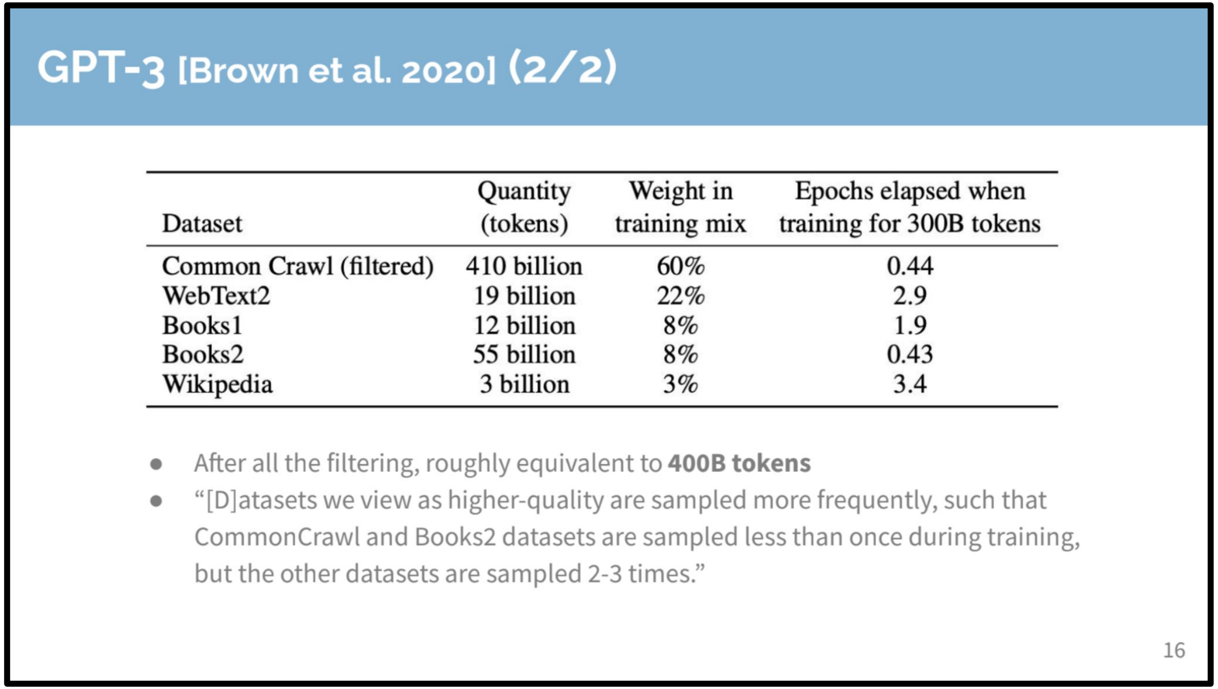
# The Common Pile v0.1

“The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text” by Kandpal et al. (2025)

**Nathan Ju, Sidhika Balachandar**

# What is in an LLM's training data?

From lecture 1:



## GPT-3 [Brown et al. 2020] (2/2)

| Dataset | Quantity (tokens) | Weight in training mix | Epochs elapsed when training for 300B tokens |
|---|---|---|---|
| Common Crawl (filtered) | 410 billion | 60% | 0.44 |
| WebText2 | 19 billion | 22% | 2.9 |
| Books1 | 12 billion | 8% | 1.9 |
| Books2 | 55 billion | 8% | 0.43 |
| Wikipedia | 3 billion | 3% | 3.4 |

- After all the filtering, roughly equivalent to **400B tokens**
- "[D]atasets we view as higher-quality are sampled more frequently, such that CommonCrawl and Books2 datasets are sampled less than once during training, but the other datasets are sampled 2-3 times."

16

# What is in an LLM’s training data?

Looking further in:

- Wikipedia
- StackExchange
- Public domain books
- US Government documents (e.g., court decisions)
- Modern books
- News articles
- Blogs
- Reddit

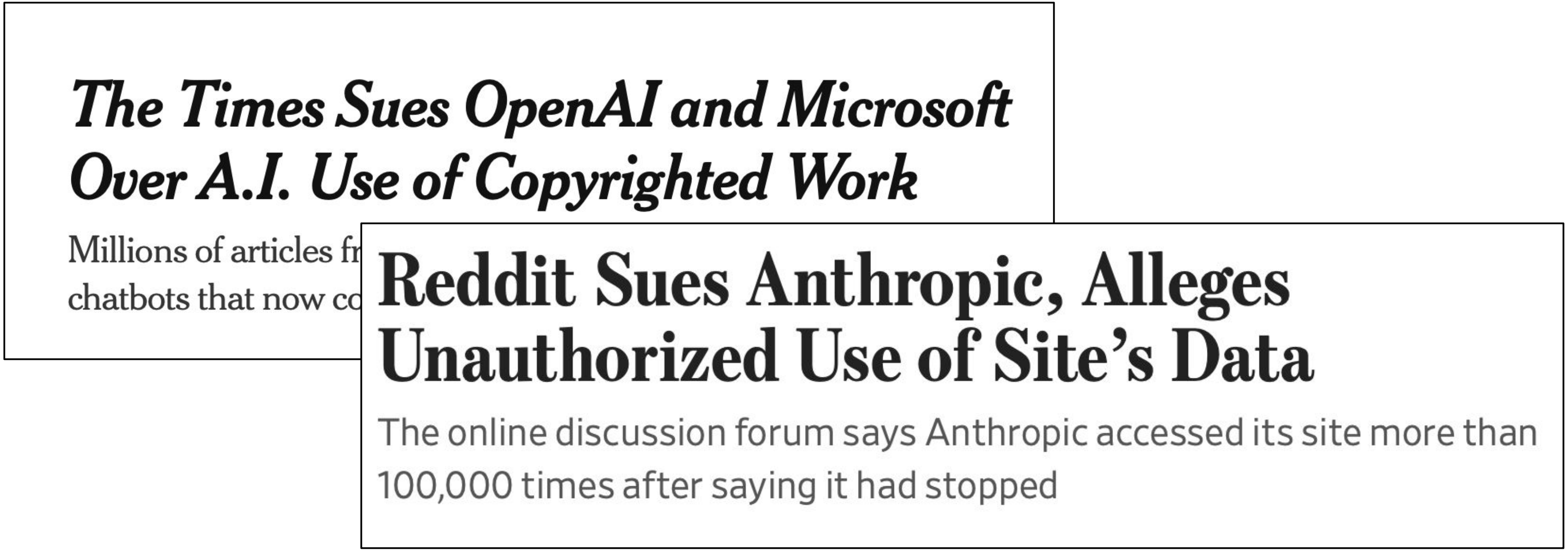
# What is in an LLM’s training data?

Looking further in:

- Wikipedia
- StackExchange
- Public domain books
- US Government documents (e.g., court decisions)
- Modern books
- News articles
- Blogs
- Reddit

# This leads to a major divide…

…between *LLM developers* and *content creators*:



***The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work***

Millions of articles fr
chatbots that now co

**Reddit Sues Anthropic, Alleges Unauthorized Use of Site's Data**

The online discussion forum says Anthropic accessed its site more than 100,000 times after saying it had stopped

# Let's answer a more basic question:

Is it *possible* to build performant models using only **open domain** and **publicly licensed** text?

# Is it possible to build performant models using only *open domain* and *publicly licensed* text?

Open domain = not copyrighted. E.g., courtlistener.com:

“This action reflects a conflict faced by many public universities in their attempt to balance the First Amendment rights of students and the need to provide a safe learning environment free from discrimination and harassment.”

Publicly licensed = “copyrighted but permission granted”. E.g. StackExchange

Q: “I've noticed that many people take a stance on social issues, many of which…”

A: “I believe that many, if not most, people do not have a clearly articulated set of principles…”

# Is it possible to build performant models using only *open domain* and *publicly licensed* text?

Open domain = not copyrighted. E.g., courtlistener.com:

> “This action reflects a conflict faced by many public universities in their attempt to balance the First Amendment rights of students and the need to provide a safe learning environment free from discrimination and harassment.”

Publicly licensed = “copyrighted but permission granted”. E.g. StackExchange

> Q: “I've noticed that many people take a stance on social issues, many of which…”

> A: “I believe that many, if not most, people do not have a clearly articulated set of principles…”

**The second criteria is not entirely well-defined for language models!**

# How do they define publicly licensed text?

They adopt Open Knowledge Foundation's *Open Definition 2.1*:

License is open if it permits **free use, redistribution, modification, and sharing for any purposes**.

- Allowed: CC-BY, CC-BY-SA, CC0, MIT, GPL, BSD, etc.
- Disallowed: CC-NC (non-commerical), CC-ND (no derivatives)

# What are some "edge cases" for their open definition?

LLM-generated synthetic datasets (e.g. SciPhi)

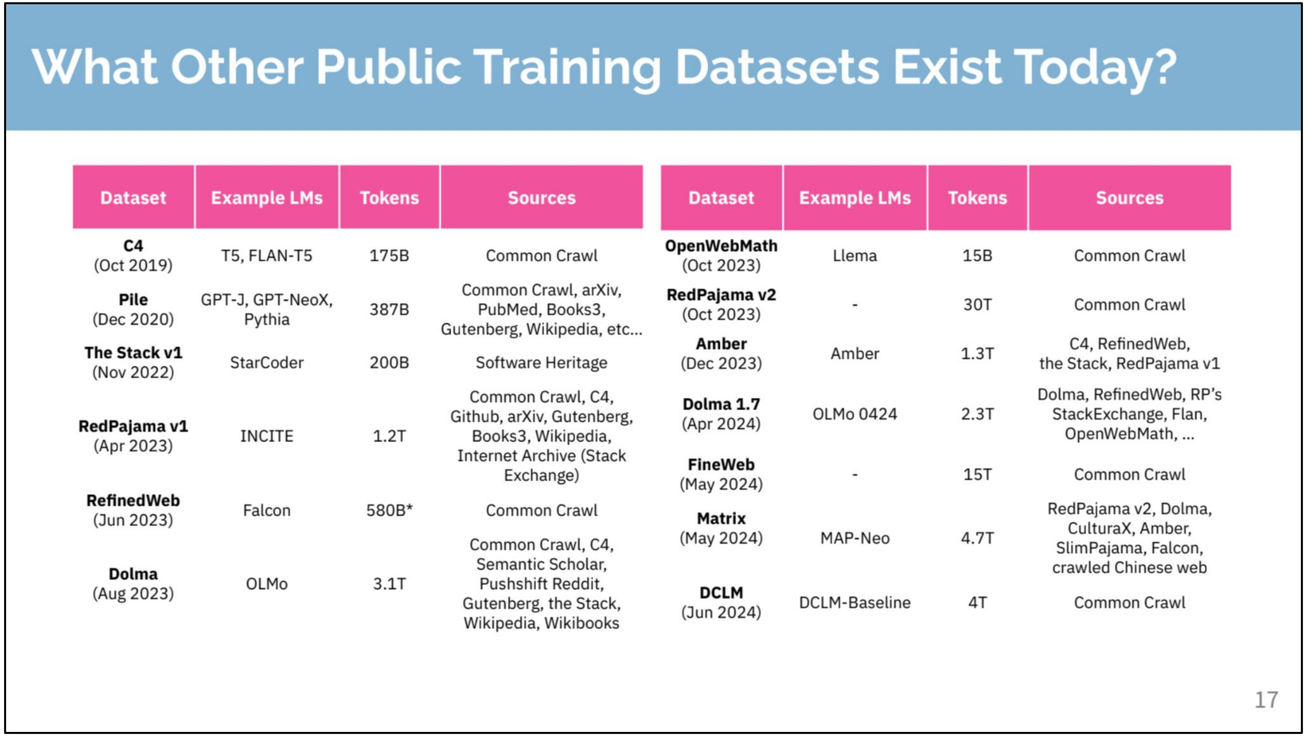
- Q: Synthetic data which was generated by model trained on copyrighted data?
- A: They take conservative stance *against* using these sources.

License laundering

- Q: Text, attached to an open license, that was incorrectly redistributed from copyrighted work?
- A: They only include sources for which they are *confident* about the source, e.g. they *exclude* HackerNews.

# Prior work (on public but improperly licensed data)

Large scale and coverage:



## What Other Public Training Datasets Exist Today?

| Dataset | Example LMs | Tokens | Sources |
|---|---|---|---|
| **C4**<br>(Oct 2019) | T5, FLAN-T5 | 175B | Common Crawl |
| **Pile**<br>(Dec 2020) | GPT-J, GPT-NeoX, Pythia | 387B | Common Crawl, arXiv, PubMed, Books3, Gutenberg, Wikipedia, etc… |
| **The Stack v1**<br>(Nov 2022) | StarCoder | 200B | Software Heritage |
| **RedPajama v1**<br>(Apr 2023) | INCITE | 1.2T | Common Crawl, C4, Github, arXiv, Gutenberg, Books3, Wikipedia, Internet Archive (Stack Exchange) |
| **RefinedWeb**<br>(Jun 2023) | Falcon | 580B* | Common Crawl |
| **Dolma**<br>(Aug 2023) | OLMo | 3.1T | Common Crawl, C4, Semantic Scholar, Pushshift Reddit, Gutenberg, the Stack, Wikipedia, Wikibooks |
| **OpenWebMath**<br>(Oct 2023) | Llema | 15B | Common Crawl |
| **RedPajama v2**<br>(Oct 2023) | - | 30T | Common Crawl |
| **Amber**<br>(Dec 2023) | Amber | 1.3T | C4, RefinedWeb, the Stack, RedPajama v1 |
| **Dolma 1.7**<br>(Apr 2024) | OLMo 0424 | 2.3T | Dolma, RefinedWeb, RP's StackExchange, Flan, OpenWebMath, … |
| **FineWeb**<br>(May 2024) | - | 15T | Common Crawl |
| **Matrix**<br>(May 2024) | MAP-Neo | 4.7T | RedPajama v2, Dolma, CulturaX, Amber, SlimPajama, Falcon, crawled Chinese web |
| **DCLM**<br>(Jun 2024) | DCLM-Baseline | 4T | Common Crawl |

# Prior work (on public and properly licensed data)

- Common Pile is *not* the first dataset focused on using properly licensed data.
- But this is a hard problem because domain distribution is skewed (only open data) and too small (only open data)
- So, prior datasets (OLC, Common Corpus, KL3M) had some flaws
  - **Bad/no licenses** - contain HackerNews, for example
  - **Unclear licensing** - does not have per-document license information
  - **Were too low quality** - e.g. contained too many gov docs
  - **Were too small** - e.g. not enough English representation

# Step 1: Acquisition

Approximately 30 sources chosen spanning:

- **Scientific and scholarly text**, e.g. arXiv
- **Online discussion forums**, e.g. StackExchange
- **Law/Government**, e.g., USPTO
- **Books**, e.g. Biodiversity Heritage Library
- **Education**, e.g. open textbooks (OER)
- **Code**, e.g. GitHub
- **Web text**, e.g. filtered Common Crawl
- **Youtube** transcriptions

# Step 1: Acquisition

Approximately 30 sources chosen spanning:

- **Scientific and scholarly text**
- **Online discussion forums**
- **Law/Government**
- **Books**
- **Education**
- **Code**
- **Web text**
- **Youtube**

Overrepresented in KL3M

Most categories are represented in OLC and Common Corpus: But Common Pile is much bigger in (English-only) scale

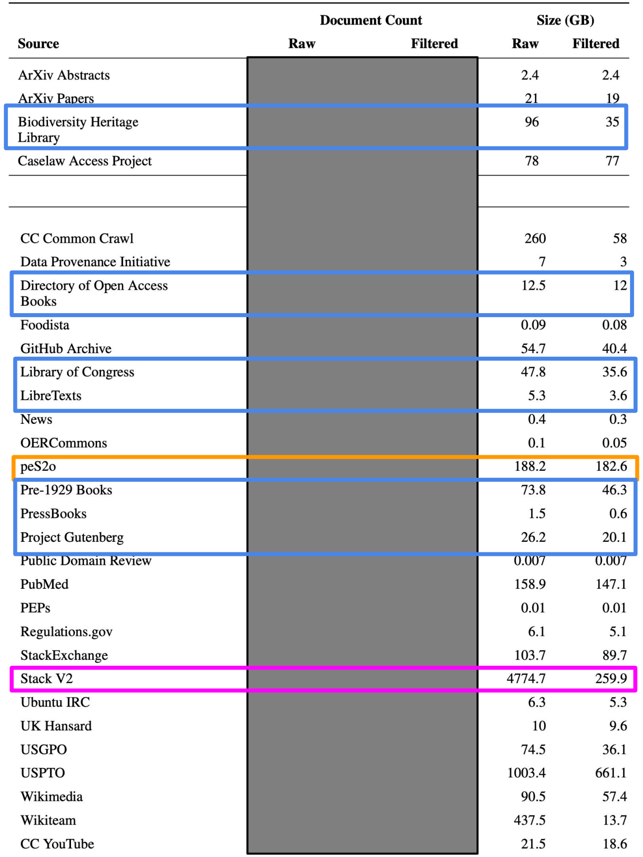
# What does Common Pile have that Pile does not (and vice-versa)?

Common Pile (licensed, 4T tokens):

uncopyrighted books

academic papers

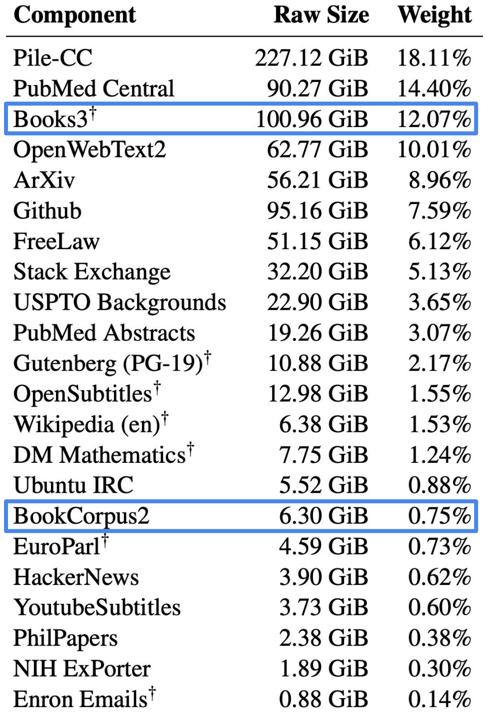
code



| | Document Count | | Size (GB) | |
|---|---|---|---|---|
| Source | Raw | Filtered | Raw | Filtered |
| ArXiv Abstracts | [redacted] | [redacted] | 2.4 | 2.4 |
| ArXiv Papers | [redacted] | [redacted] | 21 | 19 |
| Biodiversity Heritage Library | [redacted] | [redacted] | 96 | 35 |
| Caselaw Access Project | [redacted] | [redacted] | 78 | 77 |
| CC Common Crawl | [redacted] | [redacted] | 260 | 58 |
| Data Provenance Initiative | [redacted] | [redacted] | 7 | 3 |
| Directory of Open Access Books | [redacted] | [redacted] | 12.5 | 12 |
| Foodista | [redacted] | [redacted] | 0.09 | 0.08 |
| GitHub Archive | [redacted] | [redacted] | 54.7 | 40.4 |
| Library of Congress | [redacted] | [redacted] | 47.8 | 35.6 |
| LibreTexts | [redacted] | [redacted] | 5.3 | 3.6 |
| News | [redacted] | [redacted] | 0.4 | 0.3 |
| OERCommons | [redacted] | [redacted] | 0.1 | 0.05 |
| peS2o | [redacted] | [redacted] | 188.2 | 182.6 |
| Pre-1929 Books | [redacted] | [redacted] | 73.8 | 46.3 |
| PressBooks | [redacted] | [redacted] | 1.5 | 0.6 |
| Project Gutenberg | [redacted] | [redacted] | 26.2 | 20.1 |
| Public Domain Review | [redacted] | [redacted] | 0.007 | 0.007 |
| PubMed | [redacted] | [redacted] | 158.9 | 147.1 |
| PEPs | [redacted] | [redacted] | 0.01 | 0.01 |
| Regulations.gov | [redacted] | [redacted] | 6.1 | 5.1 |
| StackExchange | [redacted] | [redacted] | 103.7 | 89.7 |
| Stack V2 | [redacted] | [redacted] | 4774.7 | 259.9 |
| Ubuntu IRC | [redacted] | [redacted] | 6.3 | 5.3 |
| UK Hansard | [redacted] | [redacted] | 10 | 9.6 |
| USGPO | [redacted] | [redacted] | 74.5 | 36.1 |
| USPTO | [redacted] | [redacted] | 1003.4 | 661.1 |
| Wikimedia | [redacted] | [redacted] | 90.5 | 57.4 |
| Wikiteam | [redacted] | [redacted] | 437.5 | 13.7 |
| CC YouTube | [redacted] | [redacted] | 21.5 | 18.6 |

The Pile (unlicensed, 400B tokens):

copyrighted books



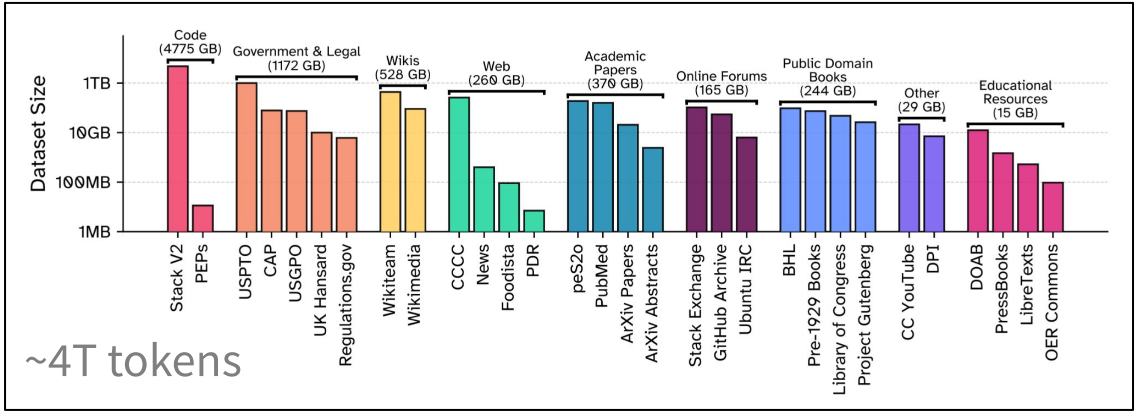
| Component | Raw Size | Weight |
|---|---|---|
| Pile-CC | 227.12 GiB | 18.11% |
| PubMed Central | 90.27 GiB | 14.40% |
| Books3† | 100.96 GiB | 12.07% |
| OpenWebText2 | 62.77 GiB | 10.01% |
| ArXiv | 56.21 GiB | 8.96% |
| Github | 95.16 GiB | 7.59% |
| FreeLaw | 51.15 GiB | 6.12% |
| Stack Exchange | 32.20 GiB | 5.13% |
| USPTO Backgrounds | 22.90 GiB | 3.65% |
| PubMed Abstracts | 19.26 GiB | 3.07% |
| Gutenberg (PG-19)† | 10.88 GiB | 2.17% |
| OpenSubtitles† | 12.98 GiB | 1.55% |
| Wikipedia (en)† | 6.38 GiB | 1.53% |
| DM Mathematics† | 7.75 GiB | 1.24% |
| Ubuntu IRC | 5.52 GiB | 0.88% |
| BookCorpus2 | 6.30 GiB | 0.75% |
| EuroParl† | 4.59 GiB | 0.73% |
| HackerNews | 3.90 GiB | 0.62% |
| YoutubeSubtitles | 3.73 GiB | 0.60% |
| PhilPapers | 2.38 GiB | 0.38% |
| NIH ExPorter | 1.89 GiB | 0.30% |
| Enron Emails† | 0.88 GiB | 0.14% |

# Step 2: Filtering

They follow fairly standard practice here:

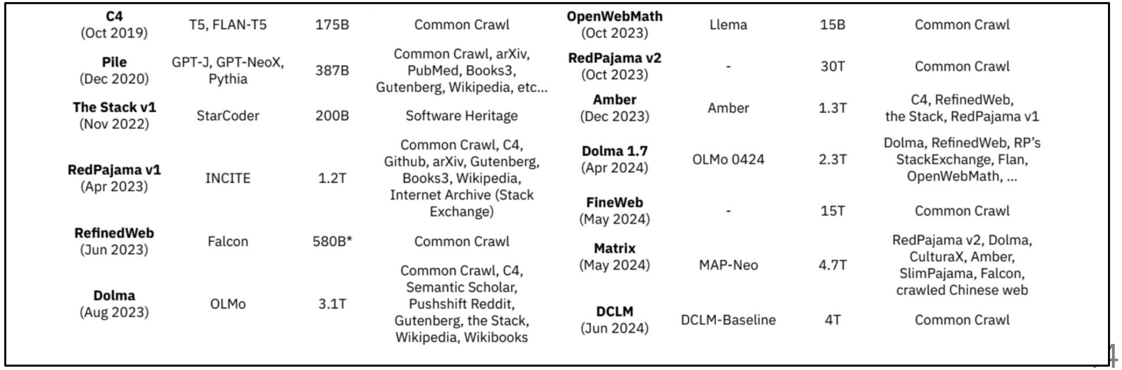
- **Language identification** (they primarily want English data)
- **Likelihood thresholds**
- **Quality filters** (classifier-based, inspired by DataComp-LM design)
- **PII removal**, e.g., regex for emails
- **Deduplication**, e.g. 90% threshold on 20-grams

# The resulting 8TB dataset, Common Pile



~4T tokens

compare to:

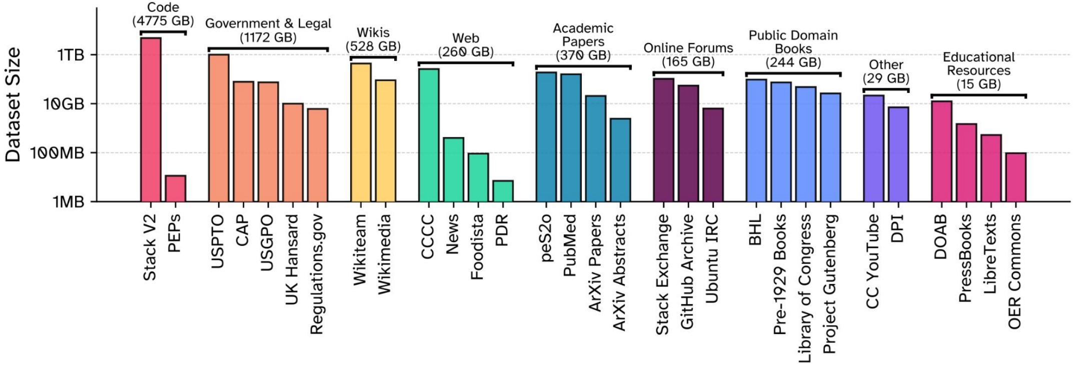


| | | | |
|---|---|---|---|
| **C4**<br>(Oct 2019) | T5, FLAN-T5 | 175B | Common Crawl |
| **Pile**<br>(Dec 2020) | GPT-J, GPT-NeoX, Pythia | 387B | Common Crawl, arXiv, PubMed, Books3, Gutenberg, Wikipedia, etc... |
| **The Stack v1**<br>(Nov 2022) | StarCoder | 200B | Software Heritage |
| **RedPajama v1**<br>(Apr 2023) | INCITE | 1.2T | Common Crawl, C4, Github, arXiv, Gutenberg, Books3, Wikipedia, Internet Archive (Stack Exchange) |
| **RefinedWeb**<br>(Jun 2023) | Falcon | 580B* | Common Crawl |
| **Dolma**<br>(Aug 2023) | OLMo | 3.1T | Common Crawl, C4, Semantic Scholar, Pushshift Reddit, Gutenberg, the Stack, Wikipedia, Wikibooks |
| **OpenWebMath**<br>(Oct 2023) | Llema | 15B | Common Crawl |
| **RedPajama v2**<br>(Oct 2023) | - | 30T | Common Crawl |
| **Amber**<br>(Dec 2023) | Amber | 1.3T | C4, RefinedWeb, the Stack, RedPajama v1 |
| **Dolma 1.7**<br>(Apr 2024) | OLMo 0424 | 2.3T | Dolma, RefinedWeb, RP's StackExchange, Flan, OpenWebMath, ... |
| **FineWeb**<br>(May 2024) | - | 15T | Common Crawl |
| **Matrix**<br>(May 2024) | MAP-Neo | 4.7T | RedPajama v2, Dolma, CulturaX, Amber, SlimPajama, Falcon, crawled Chinese web |
| **DCLM**<br>(Jun 2024) | DCLM-Baseline | 4T | Common Crawl |

# Step 3: Build a performant model

Raw sources of (open) data differ enormously in **scale** and **quality**:

- USPTO patents = billions of tokens, but low signal text
- PubMed abstracts = fewer tokens, but high signal text



# Step 3: Build a performant model

They follow fairly standard practice here:

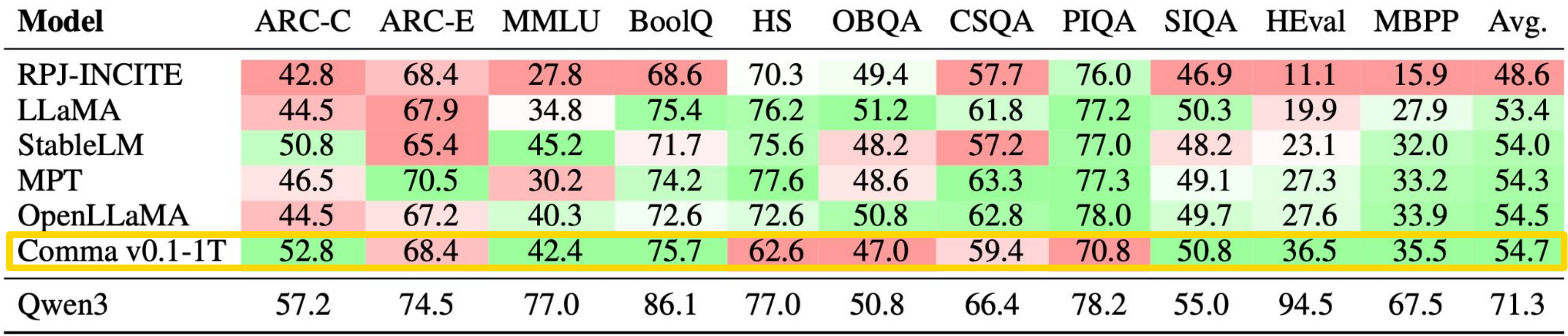
They:

- Construct a set of “early signal benchmarks” (commonsense, reasoning, knowledge)
- Train a 1.7B model on **each source**, collecting a set of **scores**
- **Heuristically** upweight each source based on the score

They train a 7B model, Comma, on this data mix.

# Is it performant?

Green=relatively strong, Red=relatively poor



| Model | ARC-C | ARC-E | MMLU | BoolQ | HS | OBQA | CSQA | PIQA | SIQA | HEval | MBPP | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RPJ-INCITE | 42.8 | 68.4 | 27.8 | 68.6 | 70.3 | 49.4 | 57.7 | 76.0 | 46.9 | 11.1 | 15.9 | 48.6 |
| LLaMA | 44.5 | 67.9 | 34.8 | 75.4 | 76.2 | 51.2 | 61.8 | 77.2 | 50.3 | 19.9 | 27.9 | 53.4 |
| StableLM | 50.8 | 65.4 | 45.2 | 71.7 | 75.6 | 48.2 | 57.2 | 77.0 | 48.2 | 23.1 | 32.0 | 54.0 |
| MPT | 46.5 | 70.5 | 30.2 | 74.2 | 77.6 | 48.6 | 63.3 | 77.3 | 49.1 | 27.3 | 33.2 | 54.3 |
| OpenLLaMA | 44.5 | 67.2 | 40.3 | 72.6 | 72.6 | 50.8 | 62.8 | 78.0 | 49.7 | 27.6 | 33.9 | 54.5 |
| Comma v0.1-1T | 52.8 | 68.4 | 42.4 | 75.7 | 62.6 | 47.0 | 59.4 | 70.8 | 50.8 | 36.5 | 35.5 | 54.7 |
| Qwen3 | 57.2 | 74.5 | 77.0 | 86.1 | 77.0 | 50.8 | 66.4 | 78.2 | 55.0 | 94.5 | 67.5 | 71.3 |

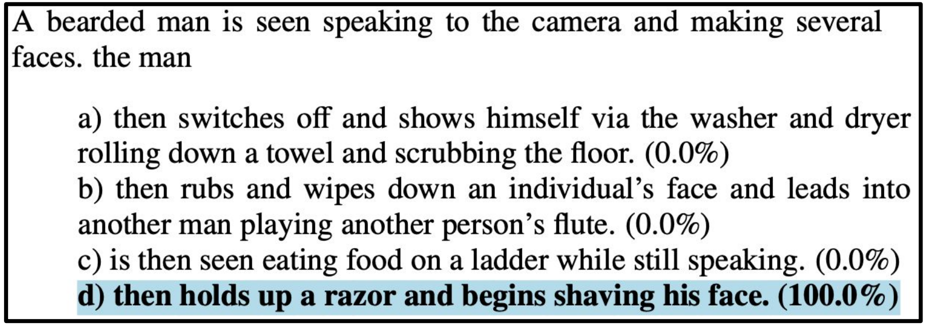
Yes! But also relatively poor on commonsense reasoning.

# Conclusions

Even after filtering and data mixing, open data seems to favor very academic and formal text, e.g.:

- peS2o (academic papers) accounts for 27% of the data mix!
- But few informal conversations, blogs, etc. (Reddit, social media)

Not good for commonsense reasoning tasks like HSwag, e.g.:



A bearded man is seen speaking to the camera and making several faces. the man

a) then switches off and shows himself via the washer and dryer rolling down a towel and scrubbing the floor. (0.0%)
b) then rubs and wipes down an individual's face and leads into another man playing another person's flute. (0.0%)
c) is then seen eating food on a ladder while still speaking. (0.0%)
**d) then holds up a razor and begins shaving his face. (100.0%)**

# Conclusions

The positive:

- Competitive on *scientific, coding, and quantitative tasks* against models trained on improperly licensed data
- Dataset scale is unprecedented for open license data

The negative:

- Poor coverage of commonsense / everyday reasoning (blogs, etc.)

# Discussion questions

Bias:

- Would multi-criteria filtering (factuality + diversity + quality) work better than single quality scores?
- Should AI companies be required to publish detailed "data nutrition labels" showing demographic breakdowns of training data?

Copyright:

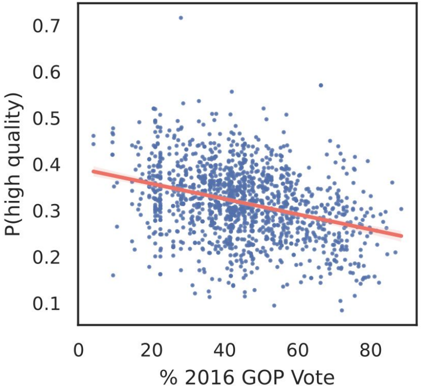
- How can the properly licensed data be augmented to improve deficient areas (e.g. commonsense)?

# Bias: critics

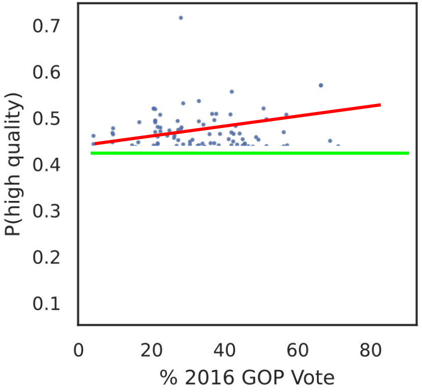
# Bias: are we properly evaluating filters?

- Correlation on all samples does **not** consider filtering (based on thresholds)
- Statistical analysis over **one** sample (model weight); abuse of p-values
- Missed opportunity to train **multiple** filters by varying the training data distribution
  - how does training data quantitatively influence the evaluations done in section 3, 4?
- So how does this affect **downstream** performance?
  - evaluating data filters in isolation without considering the whole pipeline is incomplete

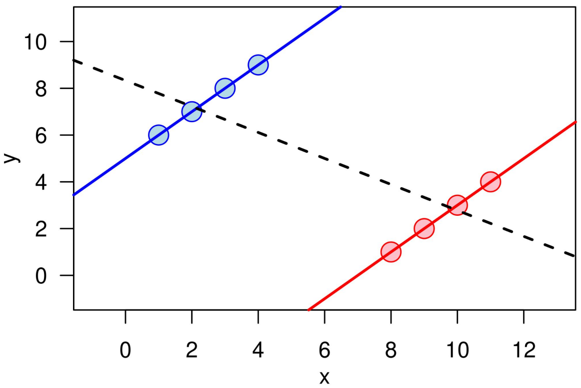
r: -0.33



r: -0.33



**Simpson's paradox**



# Bias: whose problem is it?

*"In sociolinguistics, the term **language ideology** refers to common (but often unspoken) presuppositions, beliefs, or reflections about language that justify its social use and structure."*

- Is this a problem ML *can or should* solve?
- Is acknowledging language ideology *enough*?
- Do they propose any mitigations or directions for improvement?
  - Bias Fatalism

# Whose language ideologies count as high quality?

No text selection can be truly general or value-neutral. The problem reduces to:

**So whose language ideologies count?**

*"As a starting point for new research, language ideology was a productive frame, but nearly overwhelming in terms of the data it could generate. Moreover, our endeavors seemed* ***teleological****, aimed at an already known endpoint, rather than revelatory."*

[Language ideology revisited by Jillian R. Cavanaugh]

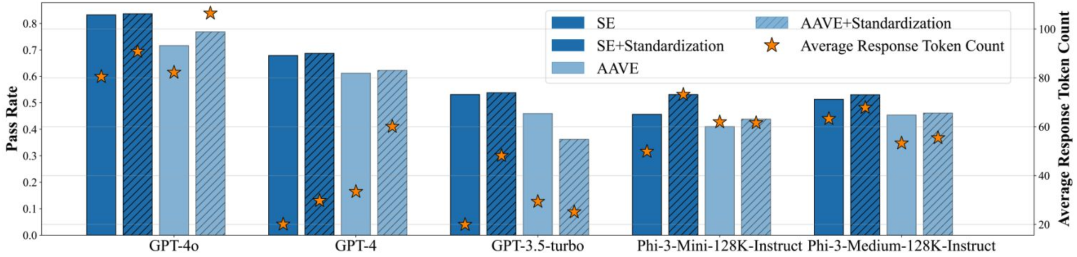
# Bias: proponents

# In Defense of Gururangan *et al.* (2022)

- “Quality” filtering has previously been shown by Dodge *et al.* (2021) to censor speech from marginalized communities.
  - C4 (quality filtered Common Crawl)
  - Blocklist filtering: removing documents containing banned tokens in the blocklist
  - “disproportionately excludes language about and by minority groups”
    - Mentions of sexual orientation
    - African American English (42%), Hispanic-aligned English (32%) vs White American English (6.2%)

# In Defense of Gururangan *et al*. (2022)

- Why does this matter?
  - decreased representation in pre-training data → minority speech becomes OOD → poor downstream performance
  - Sociolinguistics (Labov 1963): different demographics speak differently
  - LLMs perform worse on non-standard dialects (Lin *et al.*, 2025)



William Labov. 1963. The Social Motivation of a Sound Change. *WORD*, 19:3.

Lin et al. 2025. Assessing Dialect Fairness and Robustness of Large Language Models in Reasoning Tasks. In *Proc. ACL*.

# In Defense of Gururangan *et al.* (2022)

- Feng *et al.* (2023) similarly found that continued pre-training of LMs on partisan corpora leads to political biases.
  - Why does this matter?
    - The political bias of the (continued) pre-training corpus affects detection of hate speech detection and misinformation.

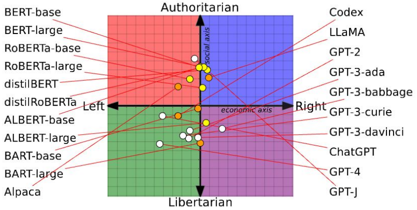


Figure 1: Measuring the political leaning of various pretrained LMs. BERT and its variants are more socially conservative compared to the GPT series. Node color denotes different model families.

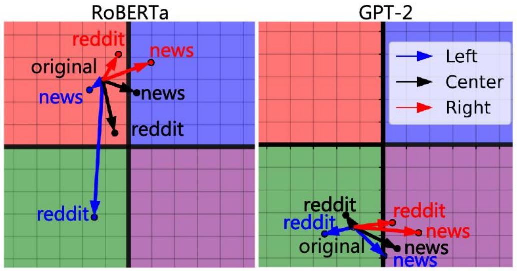
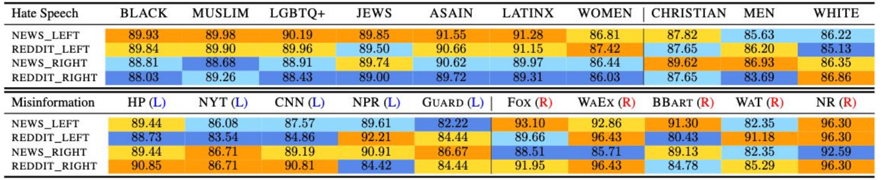


Figure 3: Pretraining LMs with the six partisan corpora and re-evaluate their position on the political spectrum.



| Hate Speech | BLACK | MUSLIM | LGBTQ+ | JEWS | ASAIN | LATINX | WOMEN | CHRISTIAN | MEN | WHITE |
|---|---|---|---|---|---|---|---|---|---|---|
| NEWS_LEFT | 89.93 | 89.98 | 90.19 | 89.85 | 91.55 | 91.28 | 86.81 | 87.82 | 85.63 | 86.22 |
| REDDIT_LEFT | 89.84 | 89.90 | 89.96 | 89.50 | 90.66 | 91.15 | 87.42 | 87.65 | 86.20 | 85.13 |
| NEWS_RIGHT | 88.81 | 88.68 | 88.91 | 89.74 | 90.62 | 89.97 | 86.44 | 89.62 | 86.93 | 86.35 |
| REDDIT_RIGHT | 88.03 | 89.26 | 88.43 | 89.00 | 89.72 | 89.31 | 86.03 | 87.65 | 83.69 | 86.86 |

| Misinformation | HP (L) | NYT (L) | CNN (L) | NPR (L) | GUARD (L) | FOX (R) | WAEX (R) | BBART (R) | WAT (R) | NR (R) |
|---|---|---|---|---|---|---|---|---|---|---|
| NEWS_LEFT | 89.44 | 86.08 | 87.57 | 89.61 | 82.22 | 93.10 | 92.86 | 91.30 | 82.35 | 96.30 |
| REDDIT_LEFT | 88.73 | 83.54 | 84.86 | 92.21 | 84.44 | 89.66 | 96.43 | 80.43 | 91.18 | 96.30 |
| NEWS_RIGHT | 89.44 | 86.71 | 89.19 | 90.91 | 86.67 | 88.51 | 85.71 | 89.13 | 82.35 | 92.59 |
| REDDIT_RIGHT | 90.85 | 86.71 | 90.81 | 84.42 | 84.44 | 91.95 | 96.43 | 84.78 | 85.29 | 96.30 |

Table 4: Performance on hate speech targeting different identity groups and misinformation from different sources. The results are color-coded such that dark yellow denotes best and dark blue denotes worst, while light yellow and light blue denote 2nd and 3rd place among partisan LMs. HP, Guard, WaEx, BBart, WaT, and NR denote Huffington Post, Guardian, Washington Examiner, Breitbart, Washington Times, and National Review.

# In Defense of Gururangan *et al.* (2022)

- Heeds Blodgett *et al.* (2020)'s call to understand how "social hierarchies and language ideologies influence the decisions made during the development and deployment lifecycle"
- Takeaway
  - Implicit biases against sociolinguistic or political groups can manifest in the pre-training data distribution, which can impact downstream performance.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proc. ACL*.

# Copyright: critics

# Copyright - Fair use doctrine

You **can** train models on copyrighted content without legal liability

# Copyright - Fair use doctrine

You **can** train models on copyrighted content without legal liability

If you successfully invoke the fair use defense

# Copyright - Fair use doctrine

You **can** train models on copyrighted content without legal liability

If you successfully invoke the fair use defense

Argument: as opposed to blindly rejecting to train on any copyrighted material, we should focus our technical efforts on ways to satisfy the fair use doctrine

# Copyright - Fair use doctrine

You **can** train models on copyrighted content without legal liability

If you successfully invoke the fair use defense

Argument: as opposed to blindly rejecting to train on any copyrighted material, we should focus our technical efforts on ways to satisfy the fair use doctrine

*Fair use* is determined by considering four factors

- **Transformativeness of use**
- Nature of copyrighted work (e.g whether original work is factual)
- Amount and substantiality of the portion of copyrighted work
- Effect of use upon potential market / value of copyrighted work

# Fair use - Transformativeness factor

When the work is transformative, this favors fair use

Past tech cases:

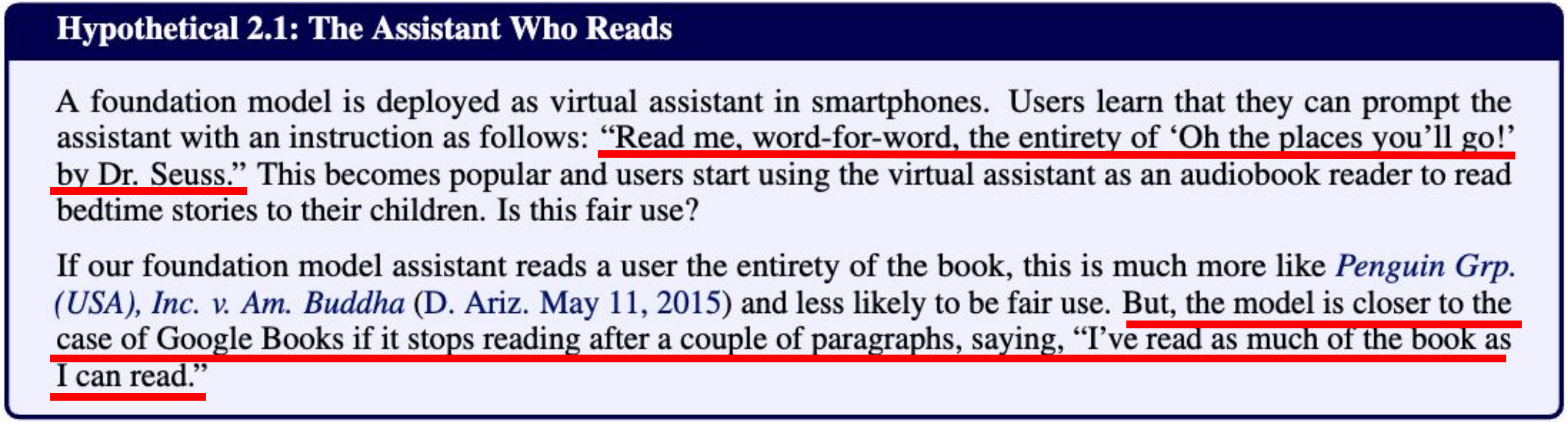
- Google copied parts of Java API for Android -> fair use because end product was transformative
- Google books can show portion of books verbatim -> use case is transformative

Transformativeness in ML:

- Transformative model ~ model who functions differently than its training data
  - E.g training a recommendation system on copyrighted books is likely fair use
- For generative models: one way to increase transformativeness ~ reduce copying of training data

# Hypotheticals: The Assistant Who Reads

One way to align with fair use:



**Hypothetical 2.1: The Assistant Who Reads**

A foundation model is deployed as virtual assistant in smartphones. Users learn that they can prompt the assistant with an instruction as follows: “Read me, word-for-word, the entirety of ‘Oh the places you’ll go!’ by Dr. Seuss.” This becomes popular and users start using the virtual assistant as an audiobook reader to read bedtime stories to their children. Is this fair use?

If our foundation model assistant reads a user the entirety of the book, this is much more like *Penguin Grp. (USA), Inc. v. Am. Buddha* (D. Ariz. May 11, 2015) and less likely to be fair use. But, the model is closer to the case of Google Books if it stops reading after a couple of paragraphs, saying, “I’ve read as much of the book as I can read.”

# Copyright - Fair use doctrine

Copyrighted content **may** be used to build foundation models without incurring liability due to the *fair use* doctrine

As opposed to blindly rejecting to train on any copyrighted material, we should focus our technical efforts on ways to satisfy the fair use doctrine (and mitigate things like memorization).

Law and technical mitigation strategies should co-evolve.

# Fair use - Nature of Copyrighted work factor

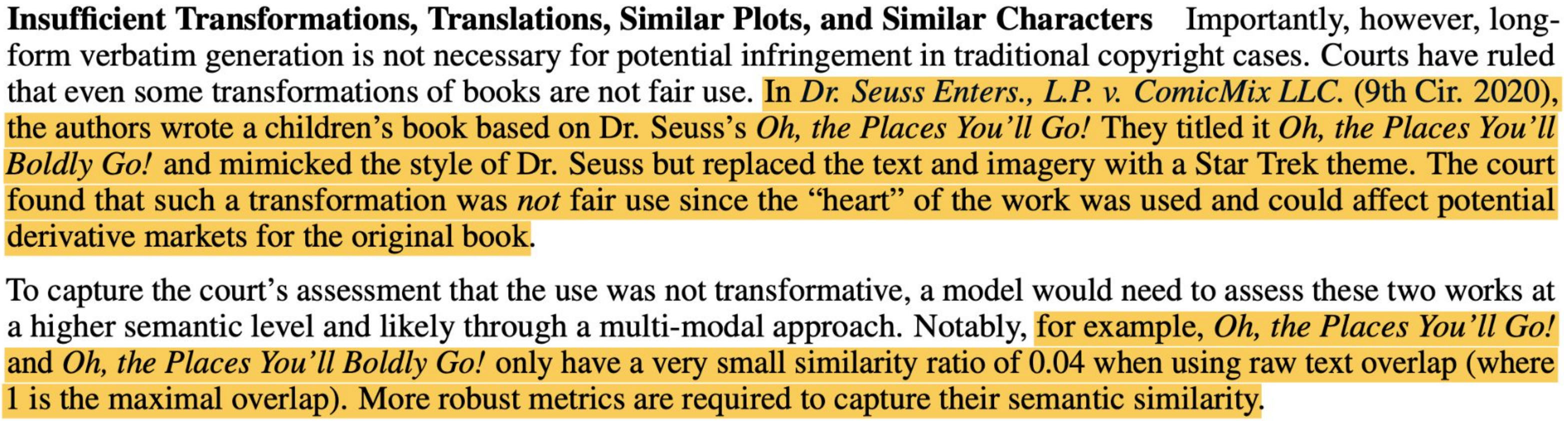
An idea cannot be copyrighted, only the expression of that idea can

Facts cannot be copyrighted, only the expression of facts can

Generative models: if the model learns high-level ideas and not low-level expressions, then it will favor fair use

# Fair use - Transformativeness factor

Key question: how much transformation is needed?



**Insufficient Transformations, Translations, Similar Plots, and Similar Characters** Importantly, however, long-form verbatim generation is not necessary for potential infringement in traditional copyright cases. Courts have ruled that even some transformations of books are not fair use. In *Dr. Seuss Enters., L.P. v. ComicMix LLC.* (9th Cir. 2020), the authors wrote a children's book based on Dr. Seuss's *Oh, the Places You'll Go!* They titled it *Oh, the Places You'll Boldly Go!* and mimicked the style of Dr. Seuss but replaced the text and imagery with a Star Trek theme. The court found that such a transformation was *not* fair use since the "heart" of the work was used and could affect potential derivative markets for the original book.

To capture the court's assessment that the use was not transformative, a model would need to assess these two works at a higher semantic level and likely through a multi-modal approach. Notably, for example, *Oh, the Places You'll Go!* and *Oh, the Places You'll Boldly Go!* only have a very small similarity ratio of 0.04 when using raw text overlap (where 1 is the maximal overlap). More robust metrics are required to capture their semantic similarity.

# Fair use - Amount and Substantiality

A critical point is how much content was taken from original work

A “de minimis” amount is acceptable

Example of cases:

- Google Books is fair use since it does not display a significant portion of books

However, using a work in its entirety does not count against fair use if the output is not itself infringing

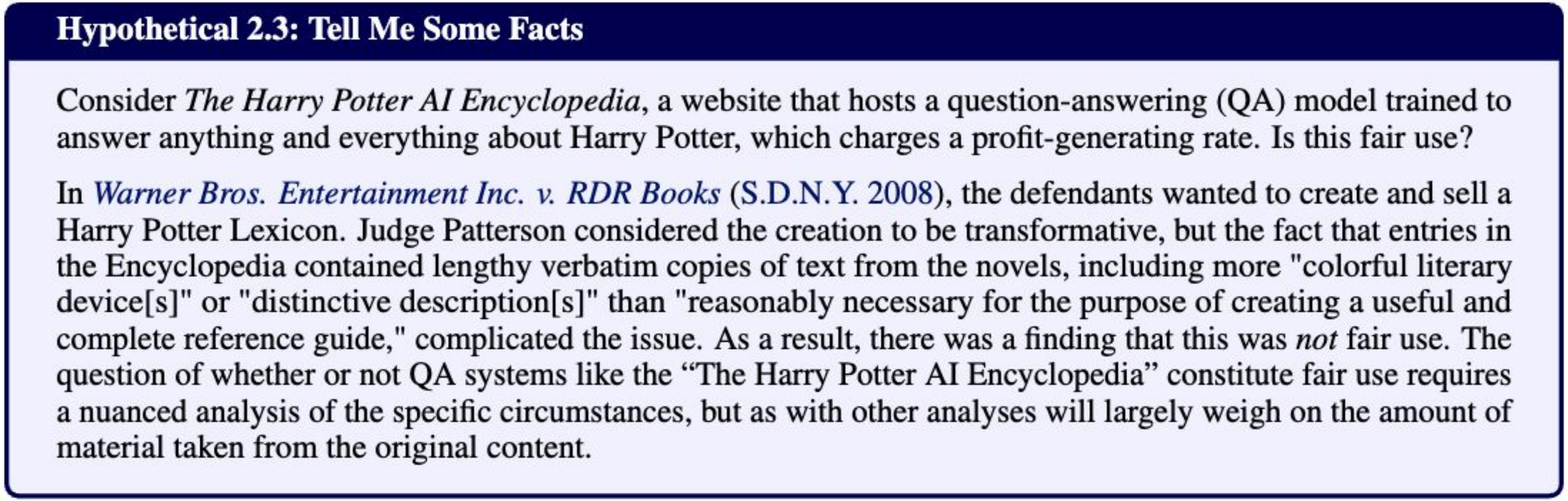
Reproducing the heart of the work, even if it is in small quantities, lowers chances of fair use

# Fair use - Effect on Market

If the new product has some effect on the market for the original work, this will be taken into account

-> Non-commercial distribution improves likelihood of fair use defense

# Hypotheticals: Tell me some facts



**Hypothetical 2.3: Tell Me Some Facts**

Consider *The Harry Potter AI Encyclopedia*, a website that hosts a question-answering (QA) model trained to answer anything and everything about Harry Potter, which charges a profit-generating rate. Is this fair use?

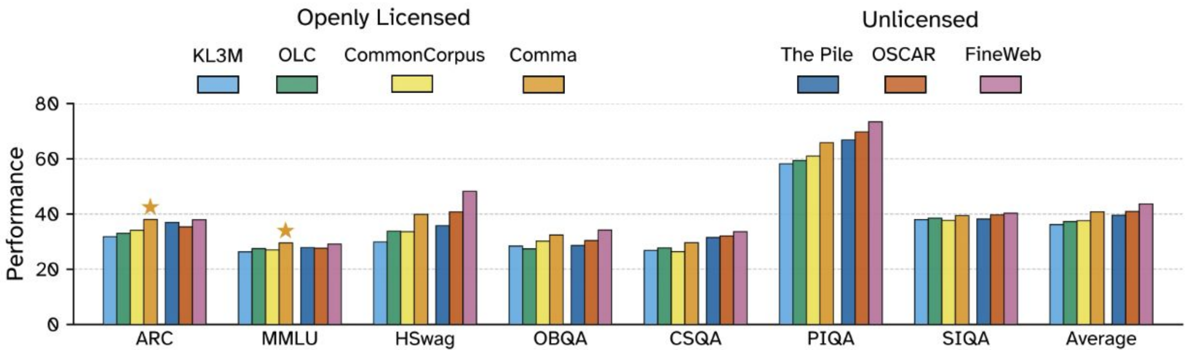
In *Warner Bros. Entertainment Inc. v. RDR Books* (S.D.N.Y. 2008), the defendants wanted to create and sell a Harry Potter Lexicon. Judge Patterson considered the creation to be transformative, but the fact that entries in the Encyclopedia contained lengthy verbatim copies of text from the novels, including more "colorful literary device[s]" or "distinctive description[s]" than "reasonably necessary for the purpose of creating a useful and complete reference guide," complicated the issue. As a result, there was a finding that this was *not* fair use. The question of whether or not QA systems like the “The Harry Potter AI Encyclopedia” constitute fair use requires a nuanced analysis of the specific circumstances, but as with other analyses will largely weigh on the amount of material taken from the original content.

# Copyright: proponents

Donghyun Lee

# Core contributions

- OLC (prior work): 0.7TB -> Common Pile: 8TB!
- Defining “openly licensed” data, dedicating an entire section in the main page and also in the appendices
- Sharing the open challenges for automatic license detection
- Showing that you don’t necessarily have to pirate data to make a good model



# Why this matters

- Ethics!: Respecting creators’ rights
- Avoiding legal risk: Anthropic case ($1.5B settlement)
- Detection could get easier: Watermarking via backdoors (Li et al., 2022), Unlearnable examples (Li et al., 2023), membership inference, etc.
- Sustainably open-sourcing datasets: “The use of unlicensed training data heavily limits the ability of model trainers to share their datasets, and has previously resulted in DMCA takedowns of datasets such as the Pile”

# Potential

- Parallel: FAL f-lite (fully licensed image dataset)
  -> More exciting legal datasets to come?

- Synthetic data generation for scaling up the dataset?
  - “Performance on HellaSwag is most heavily influenced by coverage of certain domains and topics such as personal blogs, tutorials, hobbies, and sports, which are poorly represented in the Common Pile”
  - Can a copyright-free model successfully synthesize the high quality copyright-free data?

# In Defense of Kandpal *et al*. (2025)

- Why this matters
  - Ethics!
    - Respecting creators’ rights
  - Legal risk
    - Anthropic case: Court ruled training on *purchased* books was fair use, but use of *pirated* copies was not → proposed $1.5B settlement.
  - Technical safety
    - Watermarking via backdoors – Special triggers inserted into text can cause abnormal outputs in models, letting authors verify if their data was used (Li et al., 2022)
    - Unlearnable examples – Adversarial perturbations make protected text effectively impossible for LLMs to memorize (X. Li et al., 2023).
    - Poisoned triggers – Small injected fragments can later force models to regurgitate copyrighted text, showing creators can disrupt unauthorized training (Panaitescu-Liess et al., 2025).

# In Defense of Kandpal *et al*. (2025)

- Why this matters
  - Business impact
    - “The use of unlicensed training data heavily limits the ability of model trainers to share their datasets, and has previously resulted in DMCA takedowns of datasets such as the Pile”
  - Reproducibility
    - They release the filtered/reweighted ‘Comma dataset’, mixtures, code, and model checkpoints, enabling reproducibility and ablations for the openly licensed dataset

# Follow-up

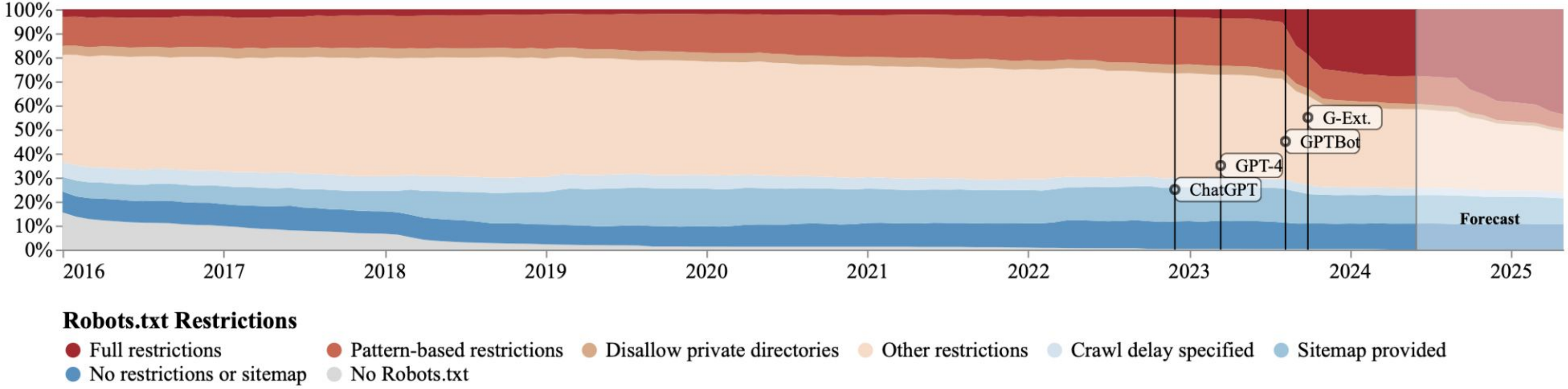
Consent in Crisis: The Rapid Decline of the AI Data Commons

Junyi Zhang

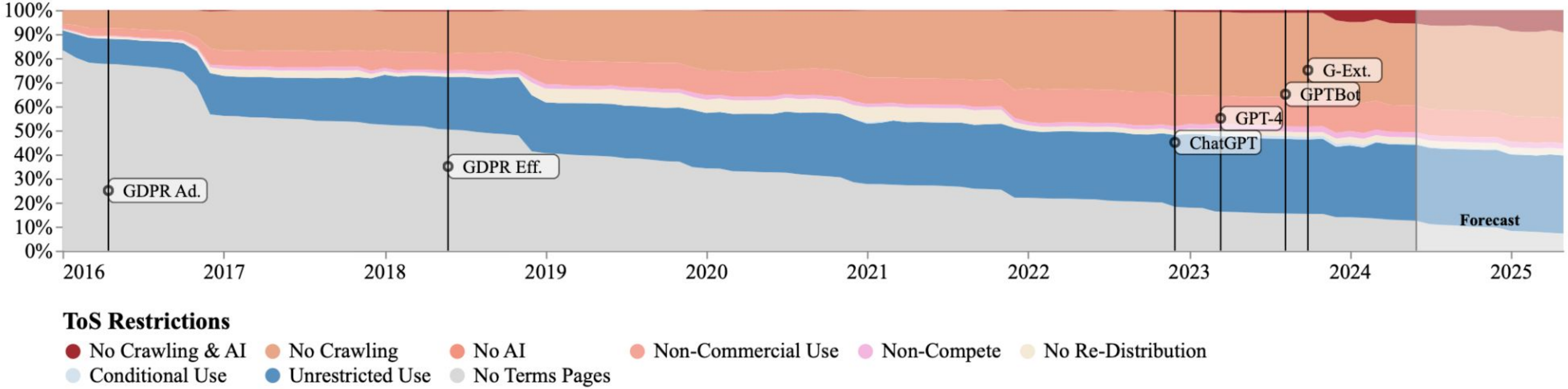
# The "What" - An AI Data Commons in Crisis

- Core Finding:
  - The public web data used to train AI models—the "AI Data Commons"—is rapidly shrinking
- The Evidence
  - A large-scale, longitudinal audit of 14,000 web domains from major training corpora (like C4) shows a massive increase in data use restrictions.
- Key Statistics (2023-2024):
  - Robots.txt: ~5% of all tokens in C4 are now fully restricted from AI crawlers. This number jumps to 28% for the most critical, actively maintained domains.
  - Terms of Service (ToS): A staggering 45% of C4 is now restricted by clauses in their ToS that forbid crawling or AI use.
  - This trend accelerated dramatically after AI crawlers like GPTBot were publicly announced.

# The "What" - An AI Data Commons in Crisis



Robots.txt Restrictions



ToS Restrictions

# The "Why" - Broken Protocols & Mismatched Incentives

- Why is this happening?
  - Web protocols like robots.txt were not designed for the age of AI, leading to chaos and confusion
- Ineffective Communication:
  - **Asymmetry:** Restrictions are inconsistent. OpenAI is blocked far more often (25.9% of top domains) than other developers like Meta (4.1%).
  - **Contradiction:** Websites' machine-readable robots.txt often conflicts with their human-readable Terms of Service. For example, 35% of sites forbid crawling in their ToS but have no robots.txt to enforce it.
- Shifting Incentives:
  - Content creators fear their work will be used to train models that compete with them, leading to a defensive stance.
  - This crisis impacts everyone, including non-commercial and academic researchers who are caught in the crossfire.

# The "How" (Impact) & Future Outlook

- How does this impact AI?
  - If these restrictions are respected, it will fundamentally alter AI development
  - Biasing Data: The available training data will become less diverse and fresh, skewing away from news and forums towards older, less dynamic content.
  - Challenging Scaling Laws: The foundational principle that "more data leads to better models" is under threat
  - End of the Open Web?: The trend incentivizes creators to move content behind paywalls or logins, further closing off the web.
- The Call to Action:
  - The paper argues for the urgent need for new, standardized web protocols that allow creators to express nuanced consent (e.g., allow for non-commercial use, or require attribution) rather than the current all-or-nothing approach.