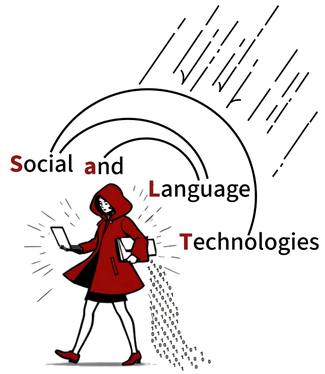
CS 329X: Human-Centered LLMs

Data, Data, and Data

Collection Filtering Curation

William Held



Who am I?

Research

Understanding and improving the robustness of LLMs to linguistic variation

Work

Contributor to the Llama 3 & 4 LLMs as part of the Pretraining Data Team at Meta GenAI

Open-Source

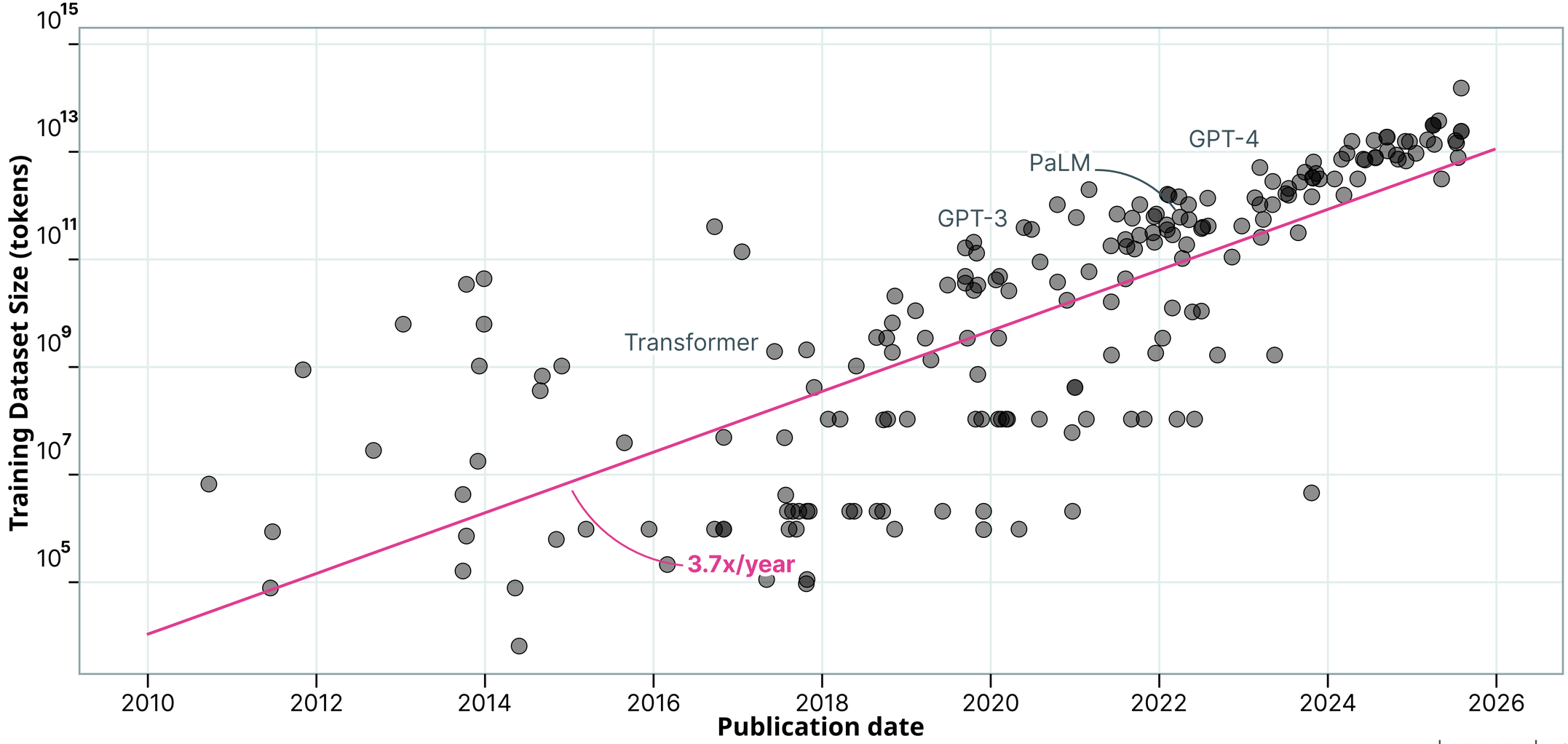
One of the Pretraining Leads for the open-source Marin LLMs as part of Stanford CRFM

Learning Goals for Today

What is the data “supply chain” for today’s LLMs?

What are the human impacts of this pipeline?

Data Drives LLM Progress



Technical Challenges in LLM Data

Collection

Data does not fall from the sky! How is the data used to train LLMs collected and structured?

Filtering

Not all data is created equal! How do we distinguish “good” data from “bad” data?

Curation

Only some good data is “useful”! How do we balance the diverse capabilities expected in LLMs?

Data in Early LLMs

BERT
English Wikipedia + BooksCorpus

Size
$3x10^9$ Tokens

Devlin et al. 2018

Collection

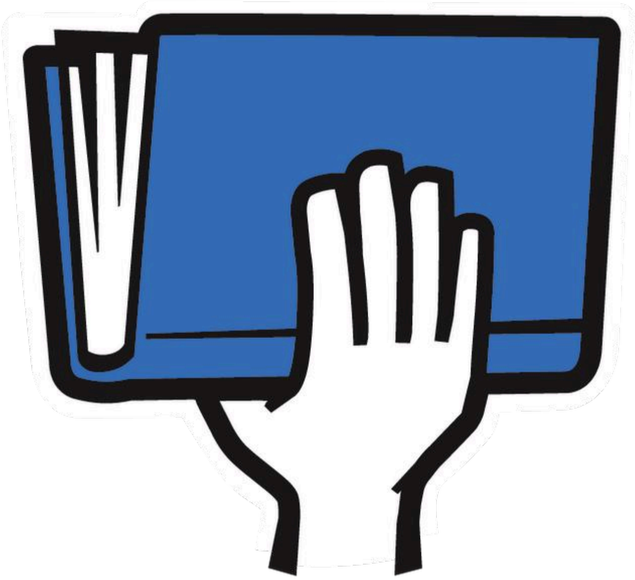
Both BooksCorpus and Wikipedia are released as single website dumps with clean text!

Filtering

Inherently, we assume books and Wikipedia to be of “high-quality” so no filtering is done.

Curation

Uniform sampling across all documents from the corpus to capture the natural distribution.



GPT-2
Reddit Outbound Links with
at least 3 upvotes

Size
$1x10^{10}$ Tokens

Radford et al. 2019

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%

Collection

Documents require processing to convert HTML to text and remove boilerplate. GPT-2 uses both DragNet and Newspaper

Filtering

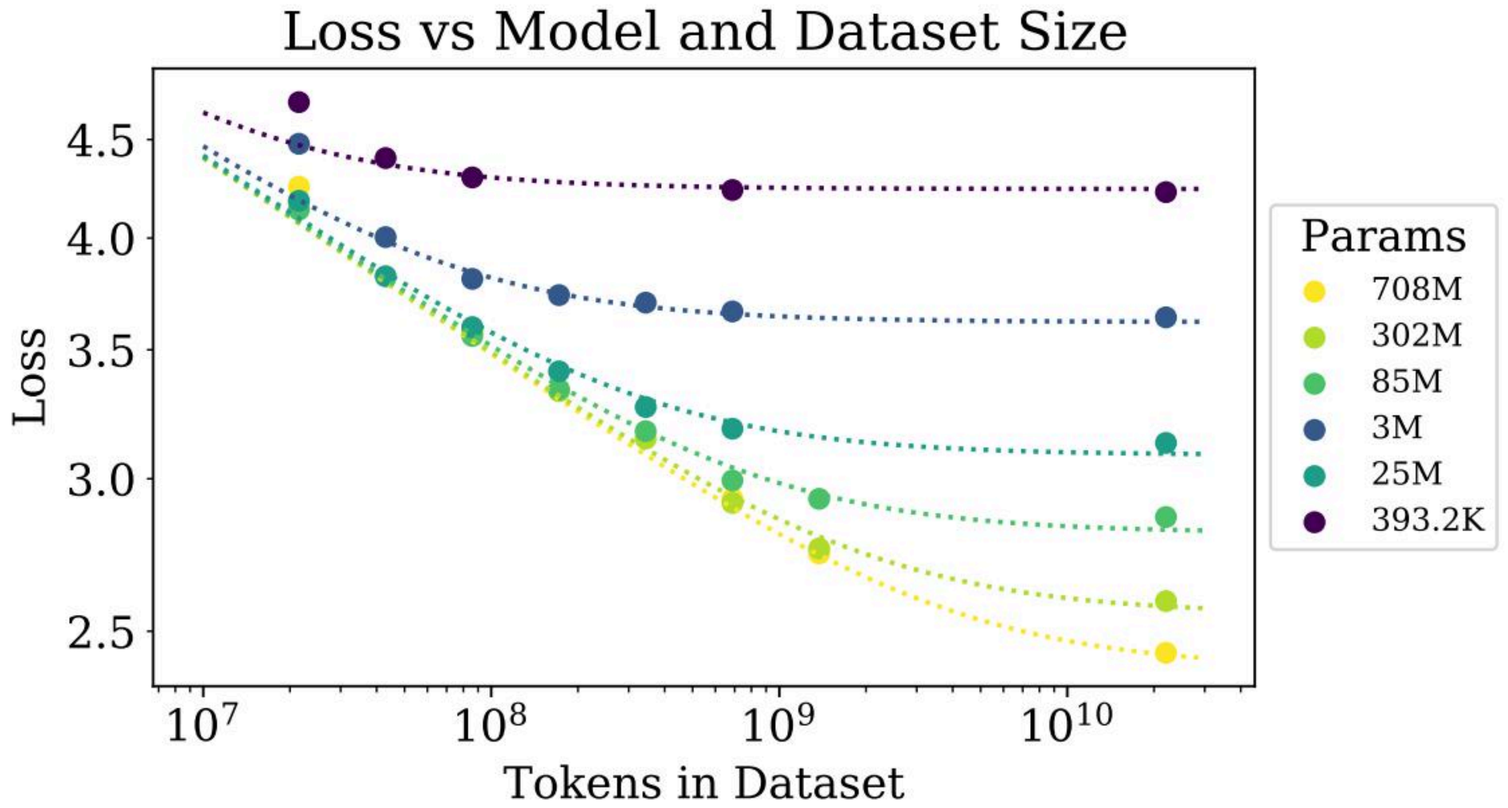
Using human-sourced quality signals (Reddit Upvotes) to determine which webpages are high-quality for training!

Curation

Uniform sampling across high-quality documents, which naturally upsamples genres popular on Reddit (e.g. news).

The first widely known GPT!

Questions?



Kaplan et al. 2020

Increasing data is a reliable lever

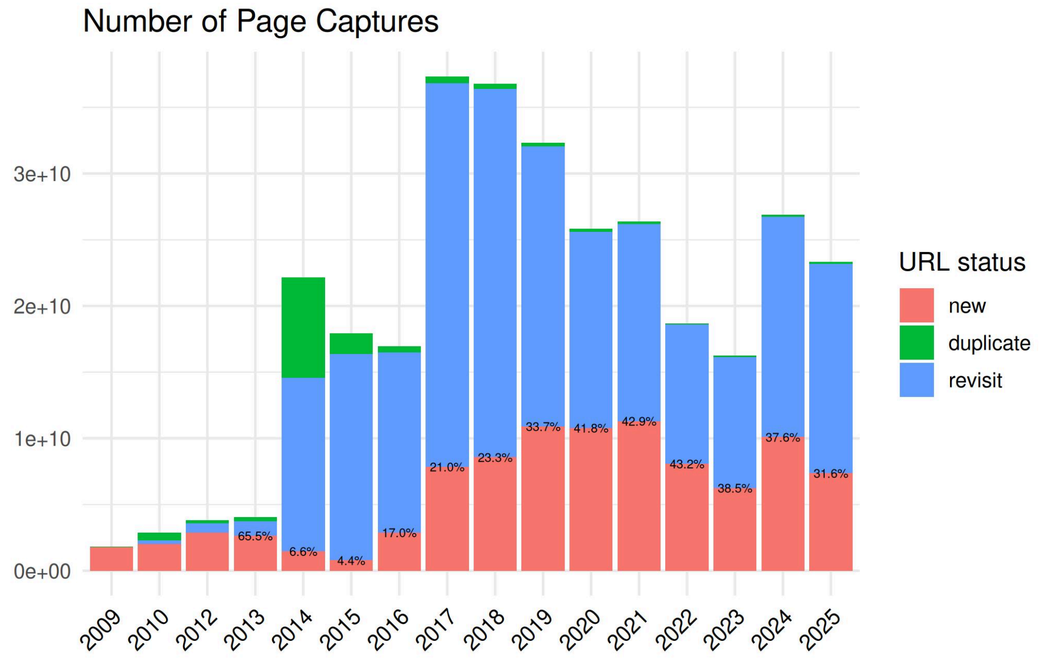
CommonCrawl
An Open Non-Profit Crawl of the Internet

Size
$1x10^{14}$ Tokens

Collection

Indiscriminate web crawling from a seed set of websites to try and cover as much of the web as possible!

CommonCrawl.org



Common Crawl as the “raw” foundation for LLM data

There’s clearly a lot more data in CC!
Most of it is garbage (or worse) though…

Random Sample of Common Crawl Data

How do we get usable training data at scale from the web?

Three Common Types of LLM Data Filtering

Perplexity Based Filtering

Given curated “good” data, use a lightweight language model to remove OOD data.

Rule Based Filtering

Researchers define programmatic rules of what “good” data looks like!

Classifier Based Filtering

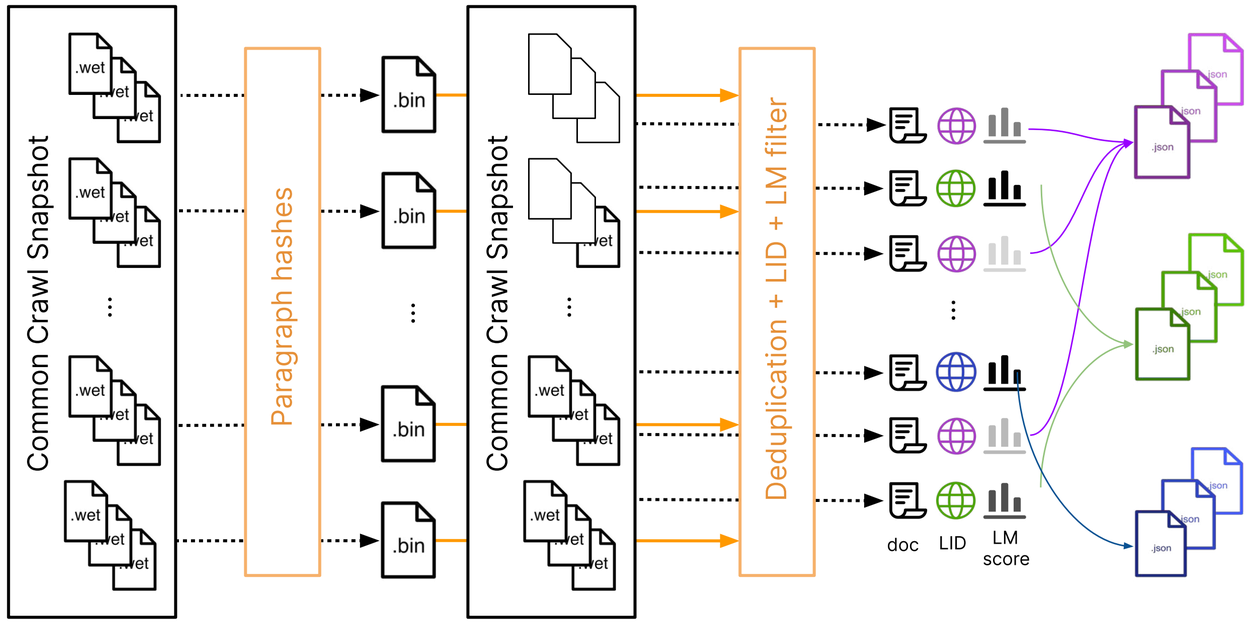
Given an expensive heuristic for “good” data, use a lightweight classifier as a proxy to remove!

CCNet

CommonCrawl Filtered by Language + PPL Filter

Size

$5x10^{11}$ Tokens

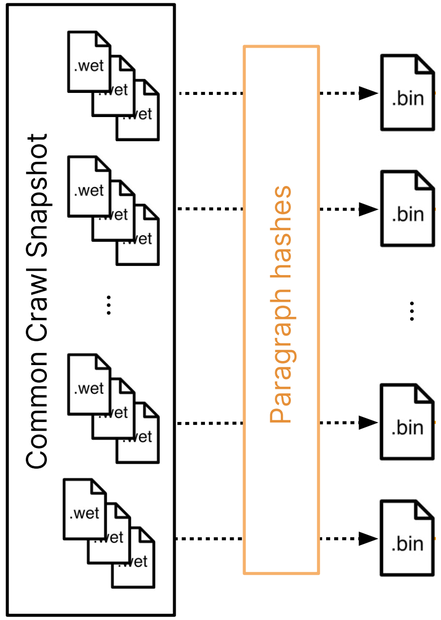


Filtering

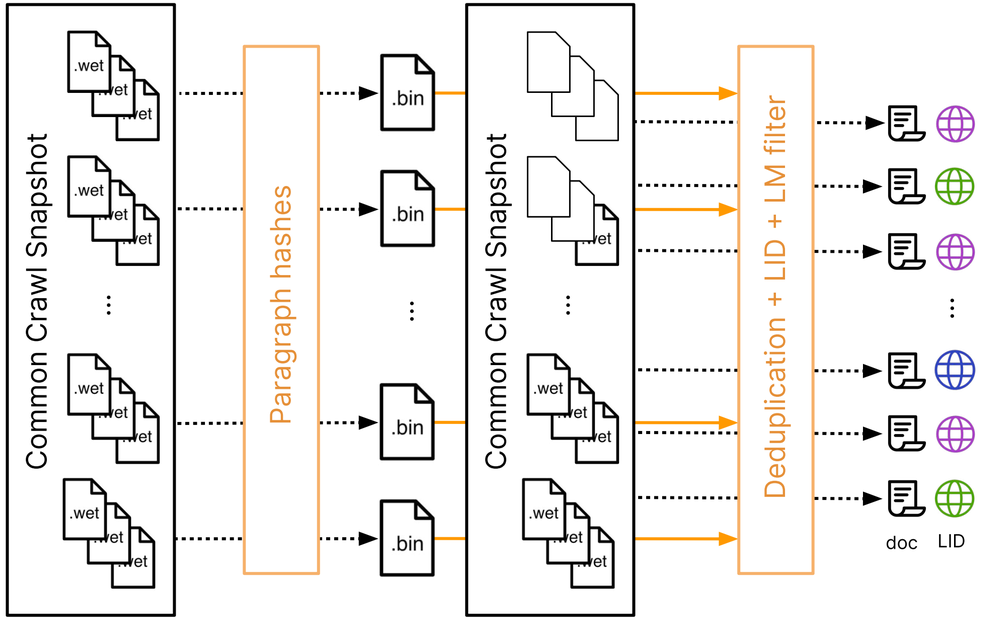
Given a CommonCrawl snapshot, removes documents that are overly “surprising” to a KenLM model trained on Wikipedia

Wenzek et al. 2019

Open-Data Era: Perplexity Filtering

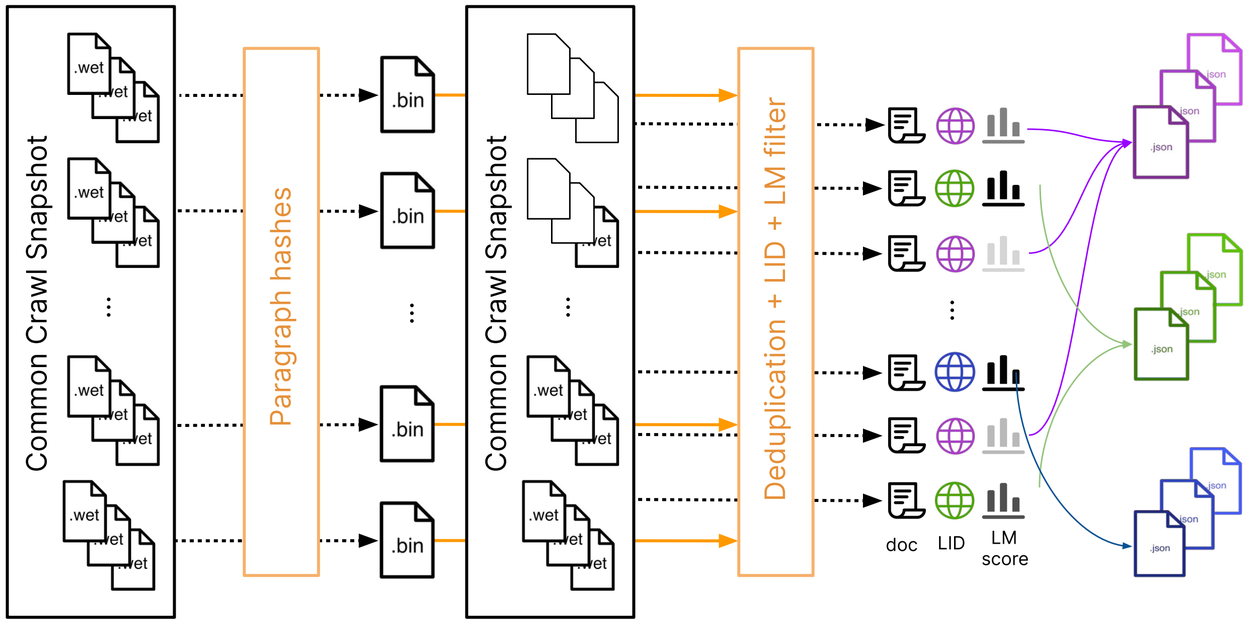


Take CommonCrawl WET (WARC Encapsulated Text) Files and Deduplicate Paragraphs Using a HashMap

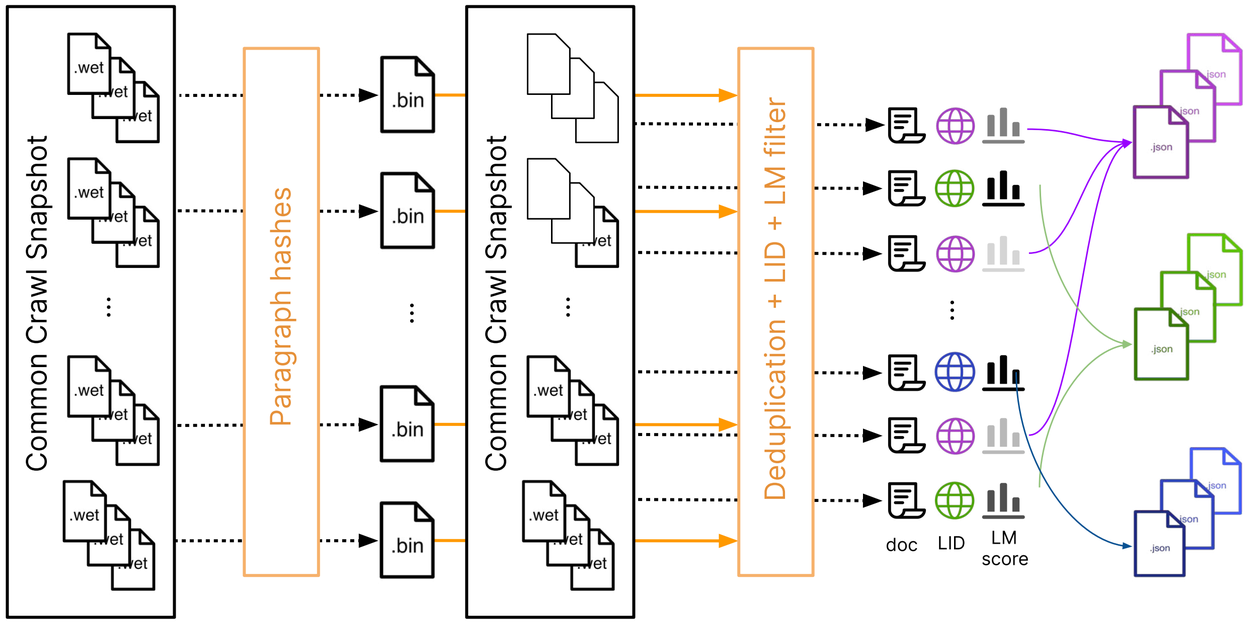


Then, classify the languages in the document using a FastText* Classifier trained on Translation data

* FastText is a simple bag of words classifier, we need something cheap here since it's large scale.



Then, remove documents that have “un-natural” language as measured by a Wikipedia N-Gram LM



Then, remove documents that have “un-natural” language as measured by a Wikipedia N-Gram LM

Questions?

Open-Data Era: Rule Based Filtering

C4

CommonCrawl Filtered by Human-Written Heuristics

Size

$3x10^{11}$ Tokens

Filtering

Given a CommonCrawl snapshot, trains removes documents & passages that fail any of the following rules.

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 3 sentences and only retained lines that contained at least 5 words.
- We removed any page that contained any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words".[6]
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder "lorem ipsum" text; we removed any page where the phrase "lorem ipsum" appeared.
- Some pages inadvertently contained code. Since the curly bracket "{" appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any pages that contained a curly bracket.
- Since some of the scraped pages were sourced from Wikipedia and had citation markers (e.g. [1], [citation needed], etc.), we removed any such markers.
- Many pages had boilerplate policy notices, so we removed any lines containing the strings "terms of use", "privacy policy", "cookie policy", "uses cookies", "use of cookies", or "use cookies".
- To deduplicate the data set, we discarded all but one of any three-sentence span occurring more than once in the data set.

Raffel et al. 2019

Open(ish)-Data Era: Classifier Based Filtering

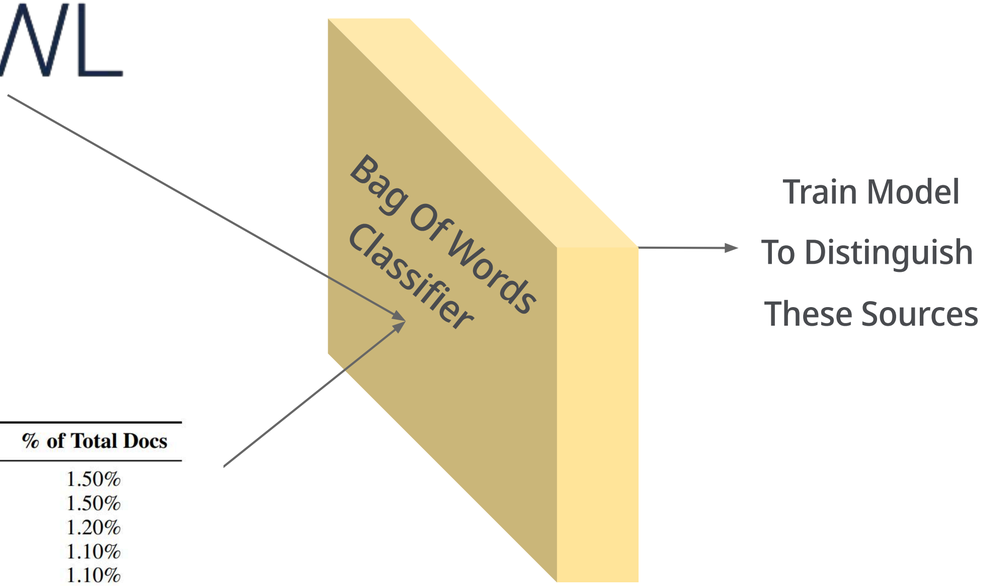
Mostly Low Quality

GPT-3 Corpus

CommonCrawl Filtered by a Classifier trained on the GPT-2 Corpus

Size

$4x10^{11}$ Tokens



Filtering

Given a CommonCrawl snapshot, keep only the "false positives" that are classified as similar to GPT-2 training data.

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 High Quality

Radford et al. 2019

Open(ish)-Data Era: Classifier Based Filtering

Three Common Types of LLM Data Filtering

Perplexity Based Filtering

Given curated “good” data, use a lightweight language model to remove OOD data.
Examples:

- Dolma CC (AllenAI) (2024)
- LLama 1 Training Data (2022)

Rule Based Filtering

Researchers define programmatic rules of what “good” data looks like!
Examples:

- RefinedWeb (2023)
- FineWeb (2024, Reading Today!)

Classifier Based Filtering

Given an expensive heuristic for “good” data, use a lightweight classifier as a proxy to remove!
Examples:

- RedPajama (2023)
- FineWeb-Edu (2024)
- Nemotron-CC (2024)
- Llama 3 & 4 (2024 & 2025)
- DCLM (2024)

There’s clearly a lot more data in CC!
We can also clearly get some good data from it!

Is that all we need for pretraining?

Dataset	Percentage
Dolmino DCLM HQ	67.8%
Dolma peS2o	10.8%
FineMath 3+	6.3%
Dolma Arxiv	5.2%
Dolma StackExchange	3.2%
StarCoder	2.2%
Dolma Algebraic Stack	2.1%
Dolma Open Web Math	0.9%
Dolma Megawika	0.8%
Dolma Wikipedia	0.7%

Marin 2025

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

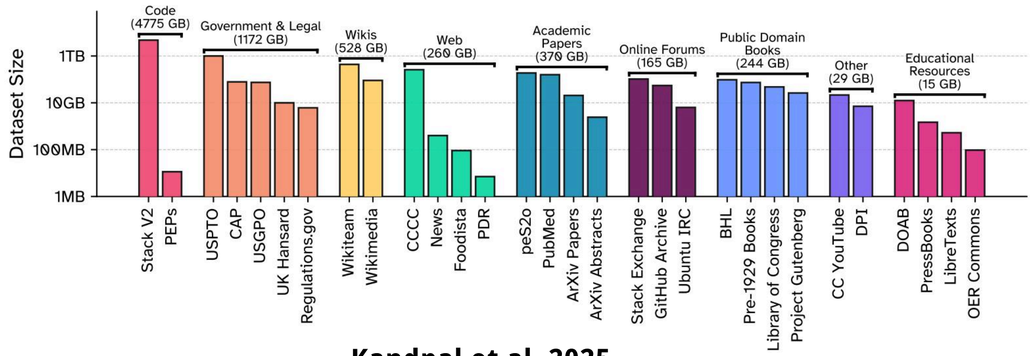
Touvron et al. 2023

Source	Type	Tokens	Words	Bytes	Docs
Pretraining ✦ OLMo 2 1124 Mix					
DCLM-Baseline	Web pages	3.71T	3.32T	21.32T	2.95B
StarCoder filtered version from OLMoE Mix	Code	83.0B	70.0B	459B	78.7M
peS2o from Dolma 1.7	Academic papers	58.6B	51.1B	413B	38.8M
arXiv	STEM papers	20.8B	19.3B	77.2B	3.95M
OpenWebMath	Math web pages	12.2B	11.1B	47.2B	2.89M
Algebraic Stack	Math proofs code	11.8B	10.8B	44.0B	2.83M
Wikipedia & Wikibooks from Dolma 1.7	Encyclopedic	3.7B	3.16B	16.2B	6.17M
Total		**3.90T**	**3.48T**	**22.38T**	**3.08B**

Soldaini et al. 2024

At a big company, many data engineers are working on collecting & filtering new datasets for particular use-cases!

In the open, the same is true for different labs and research groups!



Kandpal et al. 2025

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**

Gao et al. 2020

At a big company, many data engineers are working on collecting & filtering new datasets for particular use-cases!

In the open, the same is true for different labs and research groups!

Curation

Given all these datasets, how do we actually sample data in order to train our model?

Types of LLM Data Curation for Pretraining

Heuristic Curation

We can sample data according to our beliefs about it's relative quality or based purely on high-level metrics such as size.

Training Dynamics

We can use the degree to which our model is able to learn different training domains to inform curation!

Resource Optimization

Given some definition of our downstream goals, we can curate our data to optimize this definition of final model quality.

Types of LLM Data Curation for Pretraining

Heuristic Curation

We can sample data according to our beliefs about it's relative quality or based purely on high-level metrics such as size.

Training Dynamics

We can use the degree to which our model is able to learn different training domains to inform curation!

- DoReMi (2023)
- Online Data Mixing (2024)
- Aioli (2025)

Resource Optimization

Given some definition of our downstream goals, we can curate our data to optimize this definition of final model quality.

- Perplexity Correlations (2025)
- UtiliMax (2025)
- Scaling Laws for Optimal Data Mixtures (2025)

Lots of research here!

Types of LLM Data Curation for Pretraining

Heuristic Curation

We can sample data according to our beliefs about it's relative quality or based purely on high-level metrics such as size.

- Proportional Sampling
- Uniform Sampling
- UniMax (2023)

Most open models are here!

Training Dynamics

We can use the degree to which our model is able to learn different training domains to inform curation!

Resource Optimization

Given some definition of our downstream goals, we can curate our data to optimize this definition of final model quality.

Questions?

Instruction Tuning

Templated Training Data

Super-Natural Instructions
- 1.6K+ tasks
- Tasks contributed by community (via GitHub)

Flan
- 1.8K+ tasks
- Scraped by Google

Ultra-Curated Web Data

LIMA [Reading for Today]
- Hand Curated Small set of 1000 examples from high quality web data.
- Performance surprisingly close to frontier models at the time!

MAmmoTH2
- Curated WebInstruct, 10M instructions from Common Crawl
- Filter: train fastText classifier on quiz sites
- Extract: use GPT-4 and Mixtral to extract QA pairs

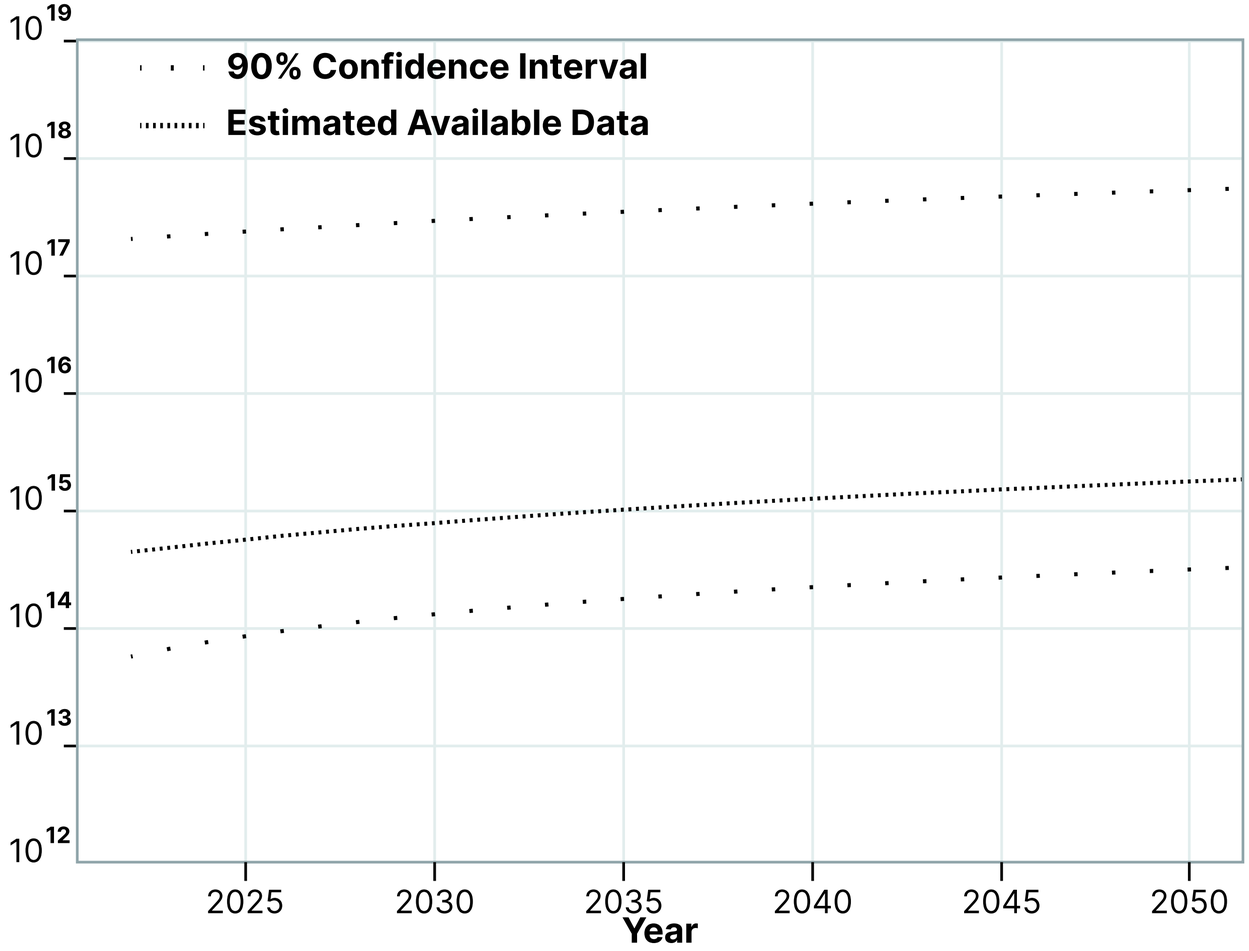
Synthetic Data From Industry Models

Alpaca
- Dataset of 52K examples from text-davinci-003 using self-instruct

Nvidia Nemotron post-training data
- Prompts: public datasets (e.g., WildChat) or synthetically-generated, then filtered
- Generated synthetic responses from Llama, Mixtral, DeepSeek r1, Qwen

Discussion time!

People Create The Data



Human Challenges in LLM Data

Ownership

Data does not fall from the sky! Who owns data and how much control do they have over the use of their data in LLMs?

Bias & Representations

Not all data is created equal! When we distinguish “good” data from “bad” data, who benefits and who could be harmed?

Utility to All

Only some good data is “useful”! When we prioritize data from certain distributions, who is likely to benefit most?

Ownership

As Language Models become more widely used, what rights do we have as the sources of data they were trained on?

Data Ownership - Copyright is King!

Copyright protection applies to 'original works of authorship fixed in any tangible medium of expression, now known or later developed, from which they can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device'

- Copyright applies to specific works, not to the ideas underlying those works!
- Copyright applies retroactively once registered.
- Since copyright applies only to the specific instance of work, the threshold is relatively low.

Most of the internet is either explicitly (due to website terms of service) or implicitly copyrighted!
This means if you want to train a LLM you only have two routes

- Pay the rights holders money for a license!
- Argue that you don’t need to pay them due to fair use

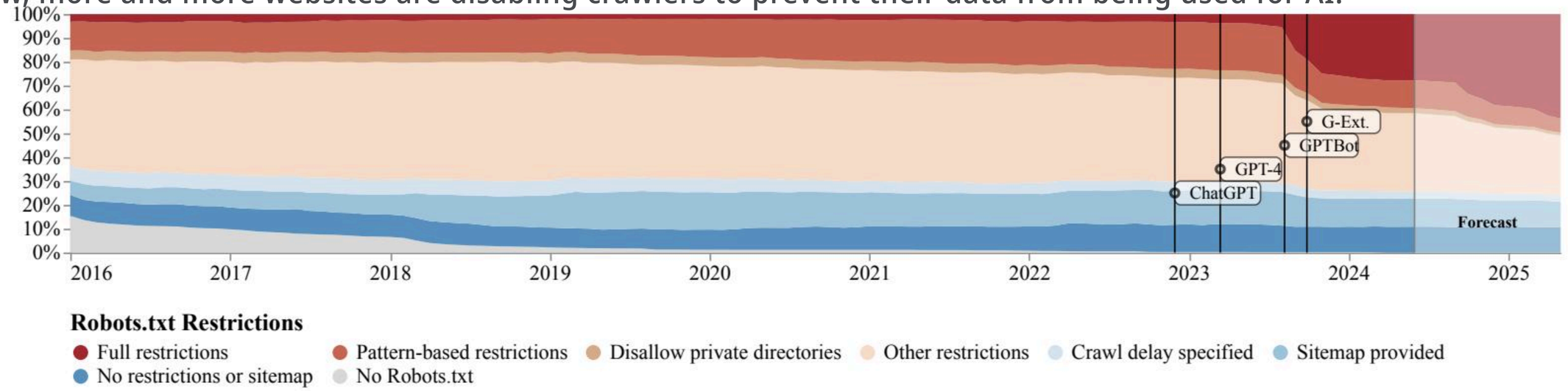
I am not a lawyer, but some real smart lawyers are looking at this:
https://simons.berkeley.edu/news/large-language-models-meet-copyright-law

Copyright is not the only data ownership mechanism

Robots.txt is a file websites can host at their root that defines their rules for how web-scrapers should access them.

Before LLMs, the vast majority of websites had relatively permissive scraping rules so that they could be indexed by search engines!

Now, more and more websites are disabling crawlers to prevent their data from being used for AI.

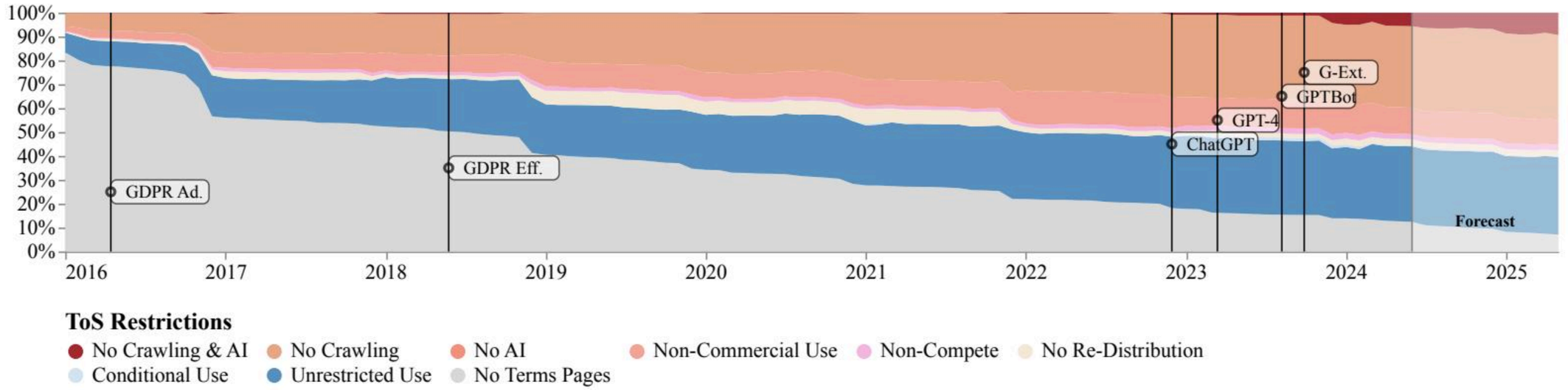


Longpre et al. 2024

Copyright is not the only data ownership mechanism

Even for data which is determined to be fair use, a website may have terms of service which make crawling the website illegal (since you add additional load to their servers or other concerns)

For example, YouTube's terms of service prohibits downloading videos, even if the videos are licensed under Creative Commons. ToS is often not enforced, but is commercially restrictive!



Longpre et al. 2024

Data in Early LLMs

BERT
English Wikipedia + BooksCorpus

Size
$3x10^9$ Tokens

Devlin et al. 2018

Collection

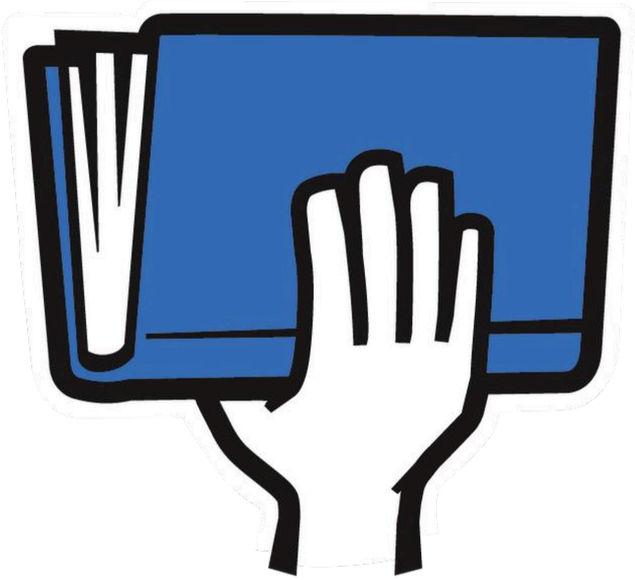
Both BooksCorpus and Wikipedia are released as single website dumps with clean text!

Filtering

Inherently, we assume books and Wikipedia to be of “high-quality” so no filtering is done.

Curation

Uniform sampling across all documents from the corpus to capture the natural distribution.



Data in Early LLMs

BERT
English Wikipedia + BooksCorpus

Size
$3x10^9$ Tokens

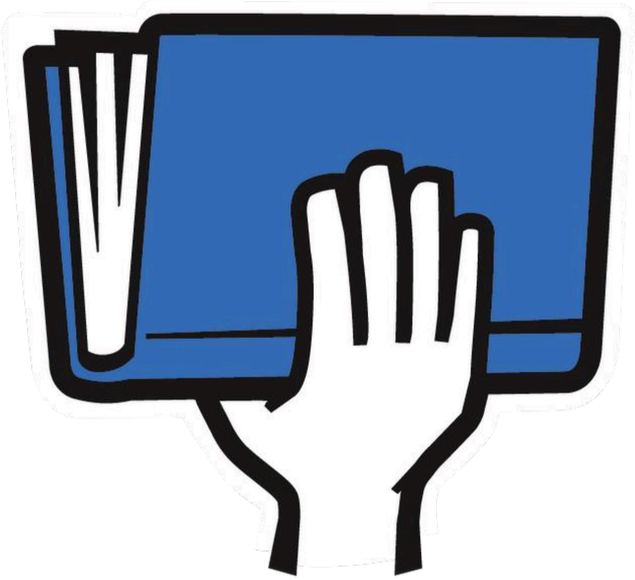
Devlin et al. 2018

Collection
Both BooksCorpus and Wikipedia are released as single website dumps with clean text!

Filtering
Inherently, we assume books and Wikipedia to be of “high-quality” so no filtering is done.

Curation
Uniform sampling across all documents from the corpus to capture the natural distribution.

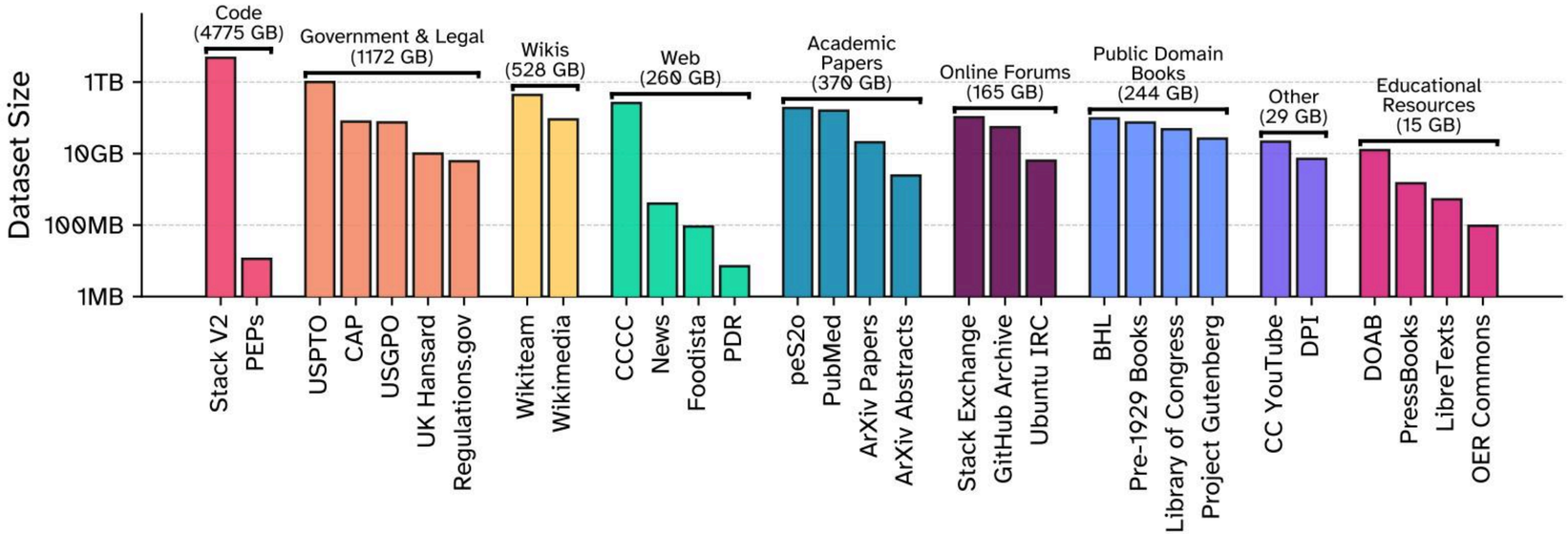
The training data for BERT was in violation of Smashwords terms of service, but not copyright!



How far can we get with truly open data?

The Creative Commons license enables free distribution of copyrighted work. In theory, anyone is free to utilize these sources to train models (in addition to anything else)

Examples: Wikipedia, Open Courseware, Khan Academy

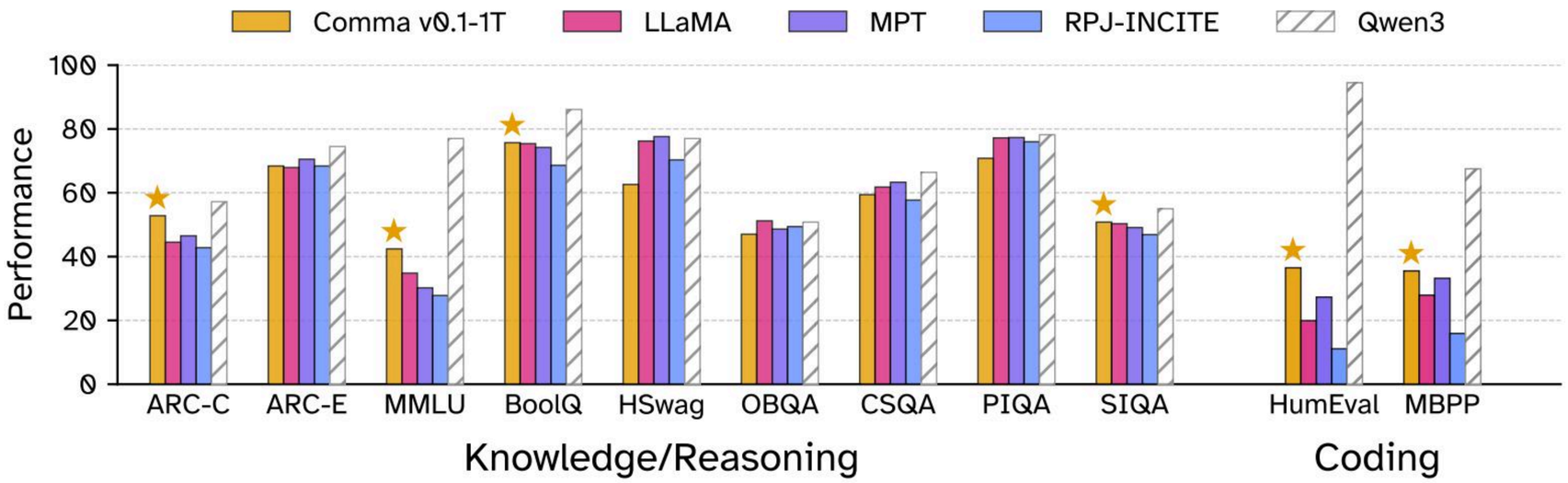


Kandpal et al. 2025

How far can we get with truly open data?

The CommonPile focuses on collecting as much data as possible that is released under this permissive license!

The resulting model trained on this data is ~3 years behind the “open-weights” state-of-the-art.

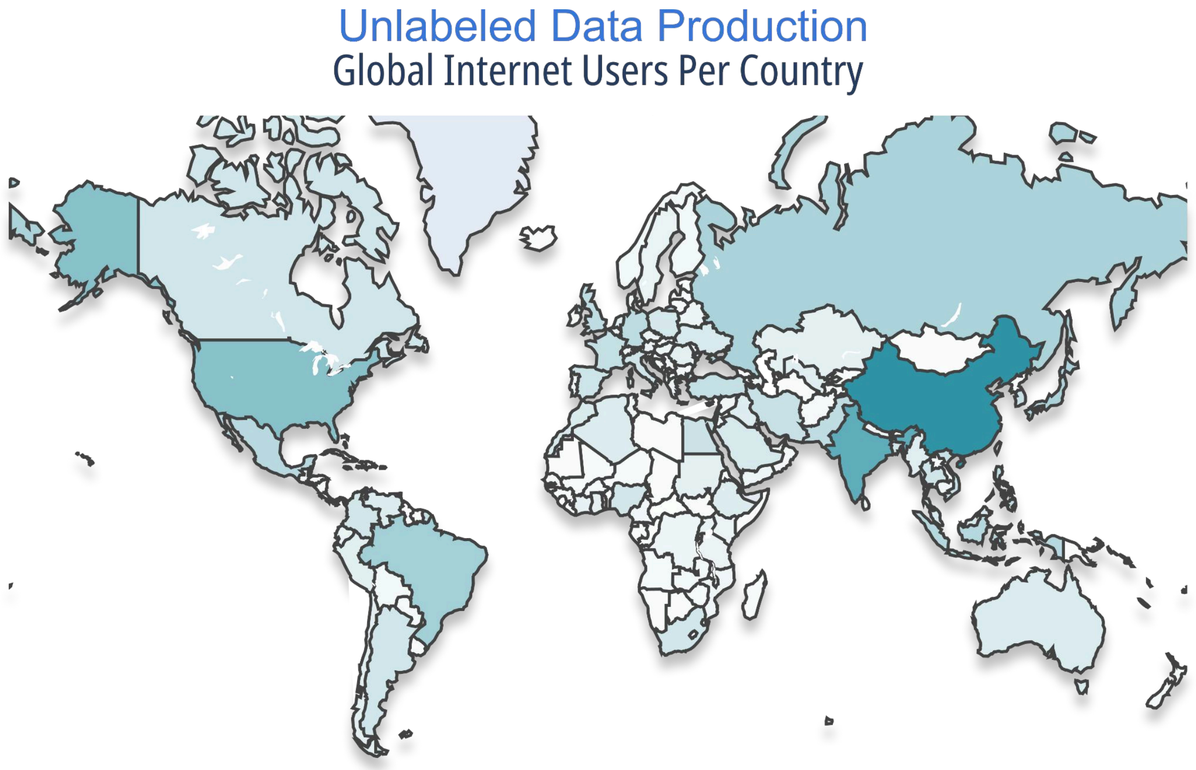


Kandpal et al. 2025

Bias & Representation

As Language Models become more widely used, whose perspectives do Language Models represent?

Before filtering, bias exists from available web data



Held et al. 2024

Before filtering, bias exists from available annotated data

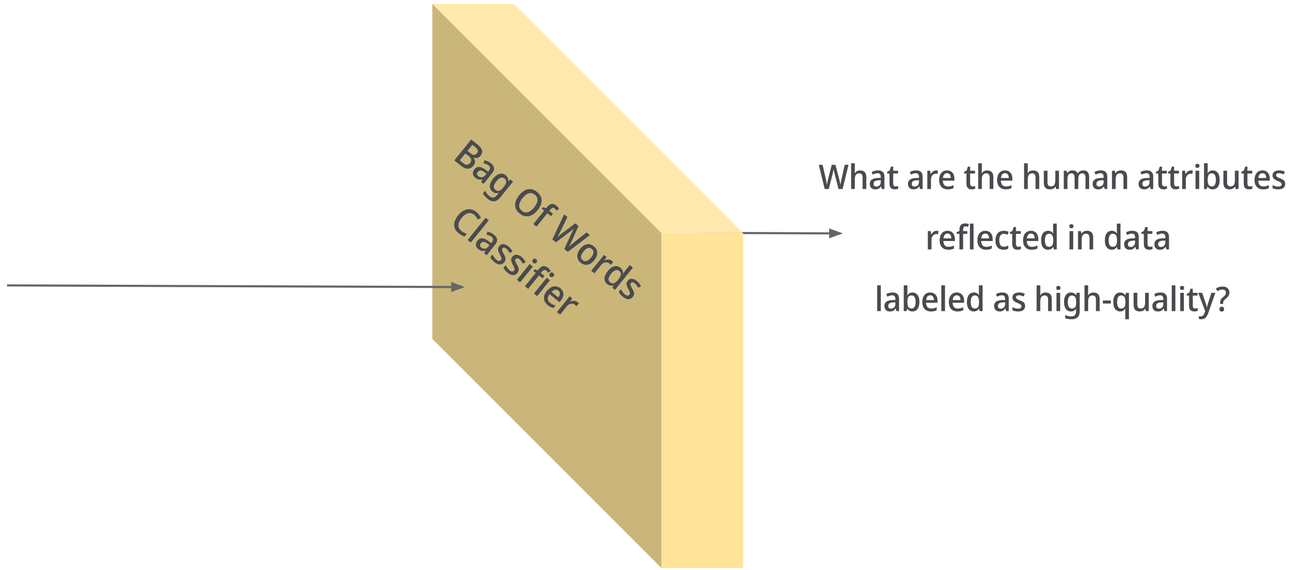
Held et al. 2024

Bias & Risks: A Case Study on Quality Filters

Bias & Risks: A Case Study on Quality Filters

SnoSites.com
High School News WordPress Template used by 2483 Schools across the United States

910K Articles
1410 schools
1329 ZIP codes
All 50 States & Washington D.C.



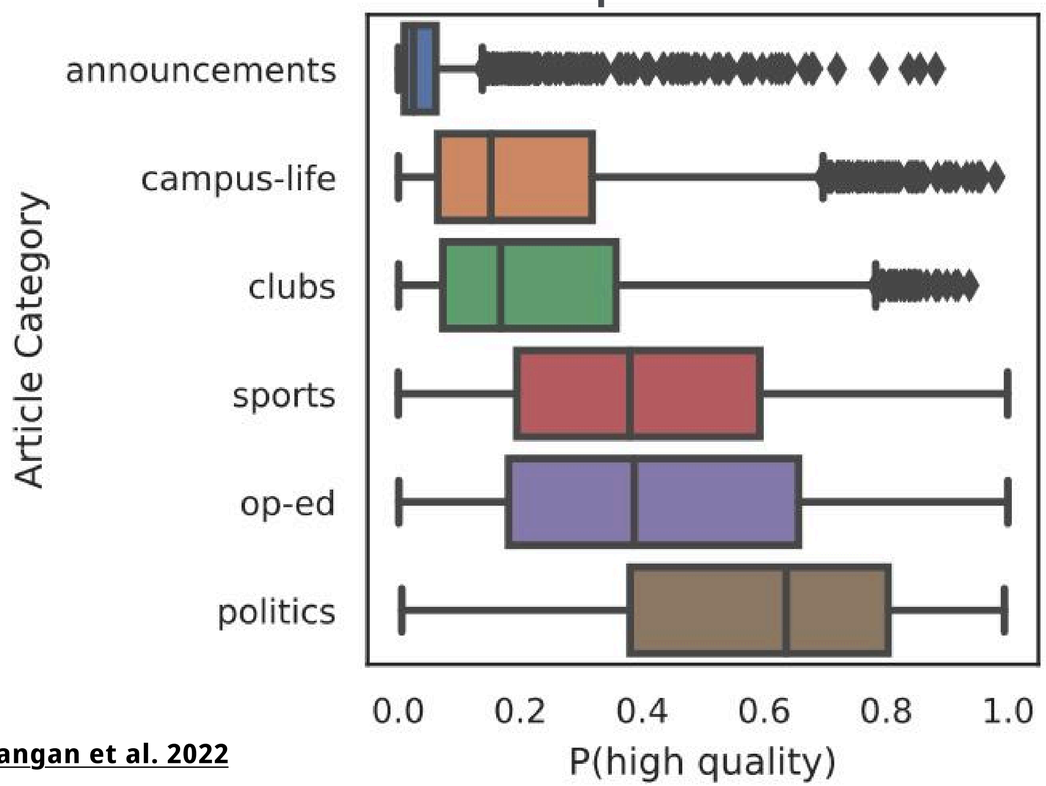
Gururangan et al. 2022

Topical Bias in Quality Classification

Top Source Domains

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%

Topical Biases



Gururangan et al. 2022

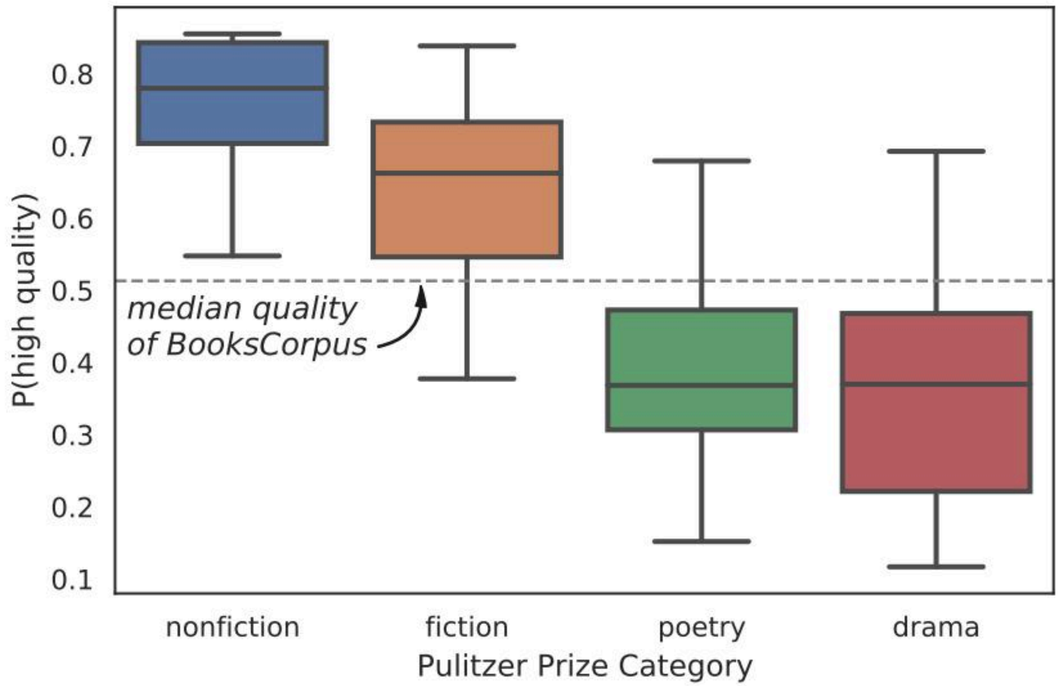
Socioeconomic Bias in Quality Classification

Dependent variable: $P(\text{high quality})$
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

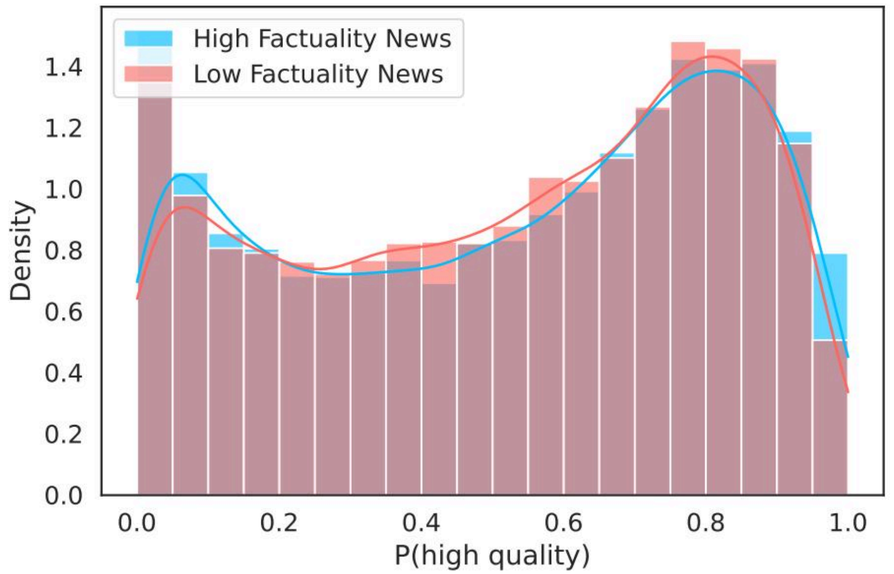
Gururangan et al. 2022

Length Bias in Quality Classification



Gururangan et al. 2022

What does this definition of “quality” end up capturing?



Gururangan et al. 2022

Utility to All

As people invest more in LLM data, do these investments map uniformly to what people want LLMs to be good at?

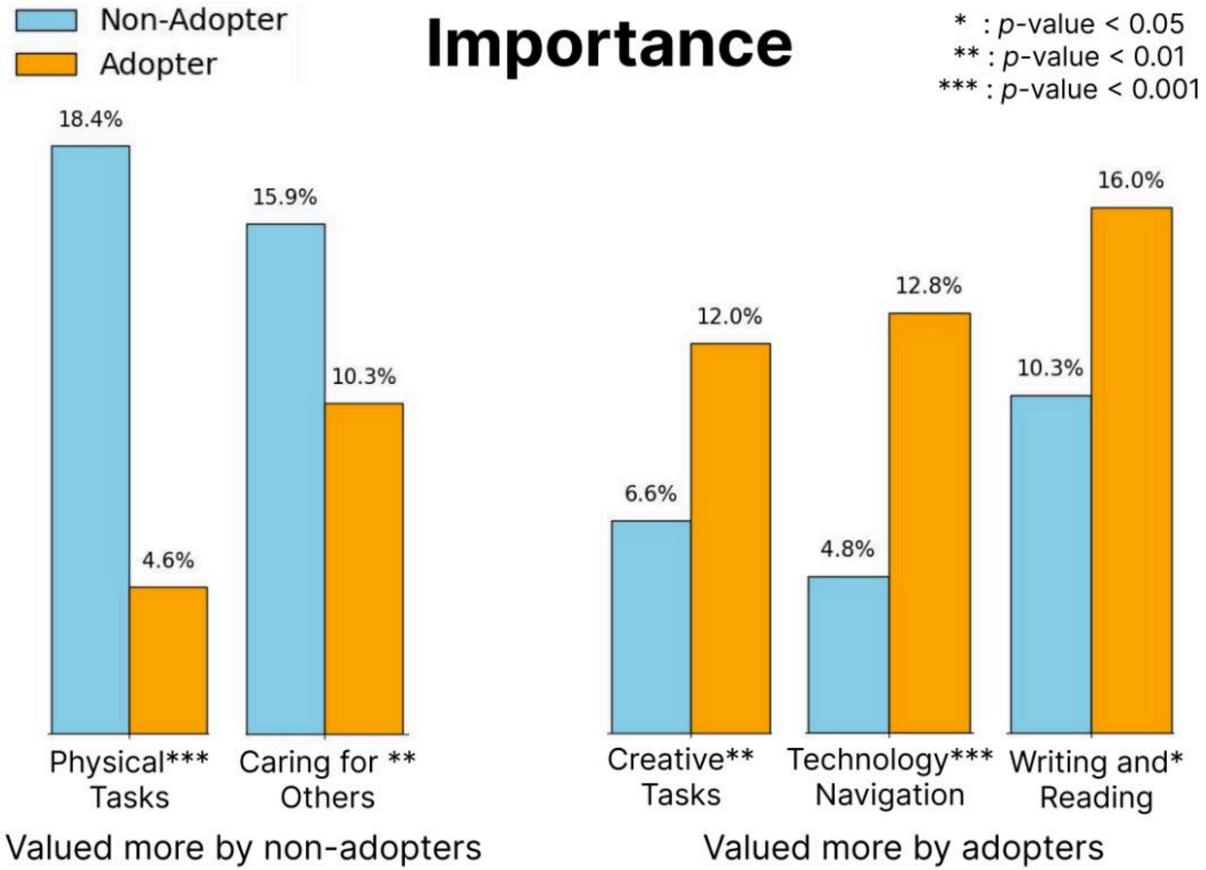
How is utility measured at big companies?

Category	Benchmark
General	MMLU (5-shot) MMLU (0-shot, CoT) MMLU-Pro (5-shot, CoT) IFEval
Code	HumanEval (0-shot) MBPP EvalPlus (0-shot)
Math	GSM8K (8-shot, CoT) MATH (0-shot, CoT)
Reasoning	ARC Challenge (0-shot) GPQA (0-shot, CoT)
Tool use	BFCL Nexus
Long context	ZeroSCROLLS/QuALITY InfiniteBench/En.MC NIH/Multi-needle
Multilingual	MGSM (0-shot, CoT)

	Claude Sonnet 4.5	Claude Opus 4.1	Claude Sonnet 4	GPT-5	Gemini 2.5 Pro
Agentic coding *SWE-bench Verified*	**77.2%** **82.0%** with parallel test-time compute	74.5% 79.4% with parallel test-time compute	72.7% 80.2% with parallel test-time compute	72.8% GPT-5 74.5% GPT-5-Codex	67.2%
Agentic terminal coding *Terminal-Bench*	**50.0%**	46.5%	36.4%	43.8%	25.3%
Agentic tool use *τ2-bench*	Retail **86.2%** Airline **70.0%** Telecom **98.0%**	Retail **86.8%** Airline 63.0% Telecom 71.5%	Retail 83.8% Airline 63.0% Telecom 49.6%	Retail 81.1% Airline 62.6% Telecom 96.7%	— — —
Computer use *OSWorld*	**61.4%**	44.4%	42.2%	—	—
High school math competition *AIME 2025*	100% (python) 87.0% (no tools)	78.0%	70.5%	99.6% (python) 94.6% (no tools)	88.0%
Graduate-level reasoning *GPQA Diamond*	83.4%	81.0%	76.1%	85.7%	**86.4%**
Multilingual Q&A *MMMLU*	89.1%	**89.5%**	86.5%	89.4%	—
Visual reasoning *MMMU (validation)*	77.8%	77.1%	74.4%	**84.2%**	82.0%
Financial analysis *Finance Agent*	**55.3%**	50.9%	44.5%	46.9%	29.4%

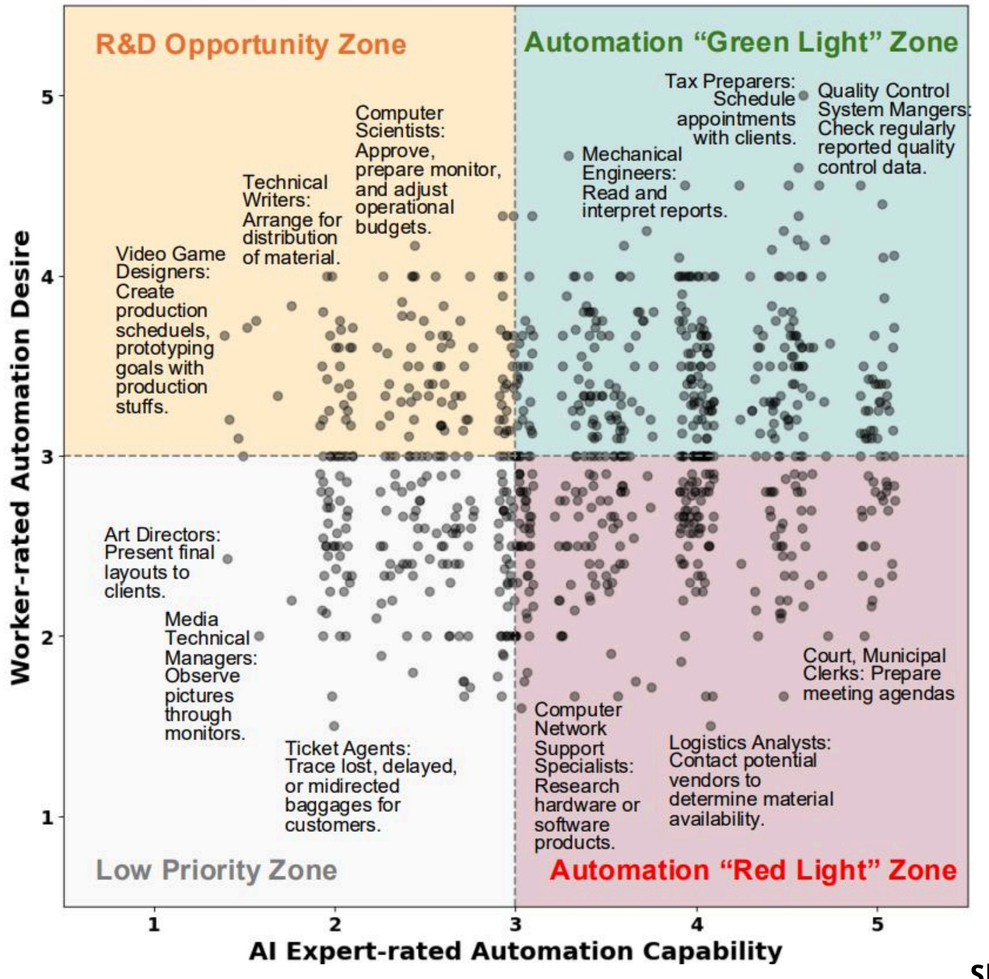
What would be useful to those not already using LLMs?

Task	Given Examples
Technology Navigation	uploading tax documents, scheduling online appointments
Obtain and Understand Information	understanding information related to health or politics
Calculation Tasks	budgeting, measurement conversions
Writing & Reading	writing emails, reading documents
Learning New Skills	recipes, languages, hobbies
Cooperating, Coordinating, & Negotiating	shared expenses, planning a family trip
Communication Tasks	giving instructions, explaining something
Physical Tasks	cleaning, cooking, yard work
Caring for Others	emotional support, caring for children, sick person, pet
Creative Tasks	designing, imagining, crafting

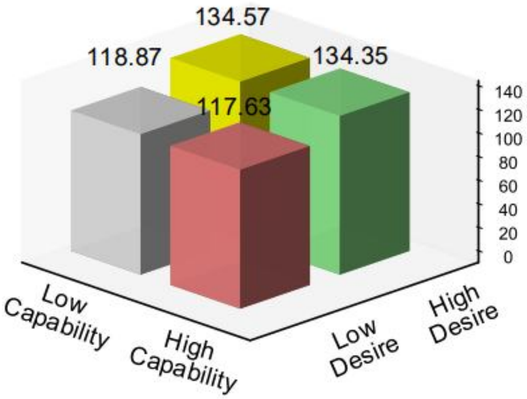


Zhou 2025

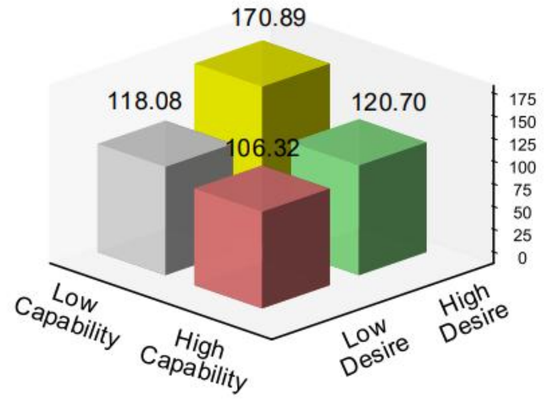
How well aligned are investments into research with these goals?



Industry Investment in AI Tasks



Academic Investment in AI Tasks



Shao et al. 2025

Questions?