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

Paper Analysis

Mohamed Amine Kina

Context

2019

...F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested lan...

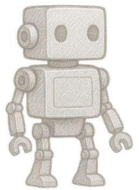
GPT-2 paper, Abstract

...vised objective to convergence. Preliminary experiments confirmed that sufficiently large language models are able to perform multitask learning in this toy-ish setup but learning is much slower than in explicitly supervised approaches.

GPT-2 paper, Section 2

Context

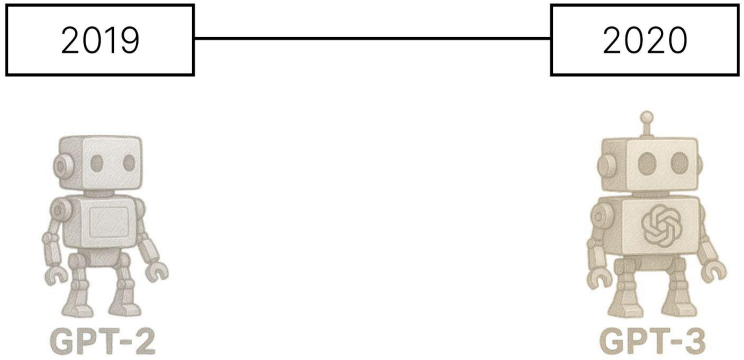
2019



Number of parameters	1.5 Billion
Training data size	40 GB of text
Architecture	Transformer (decoder-only)
Training Efficiency	Underfit dataset even at 1.5B parameters

[1]

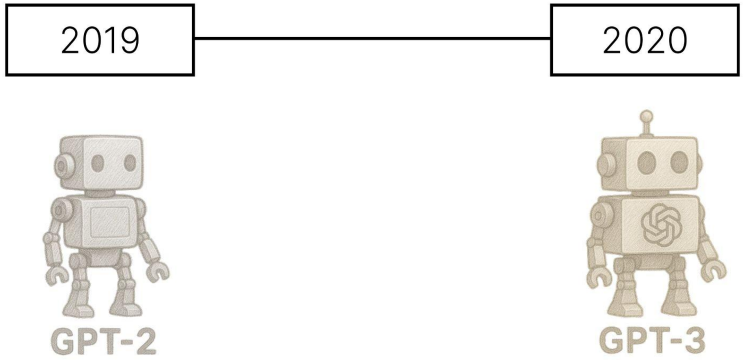
Context



	GPT-2	GPT-3
Number of parameters	1.5 Billion	175 Billion x116
Training data size	40 GB of text	570 GB of text x14
Architecture	Transformer (decoder-only)	Transformer (decoder-only)
Training Efficiency	Underfit dataset even at 1.5B parameters	Trained with larger data to match model capacity

[1] [2]

Training Data



	GPT-2	GPT-3
Training data size	40 GB of text	570 GB of text
Data source	Reddit-linked high-quality web pages	Common Crawl WebText2 Wikipedia Books

Common Crawl is an open repository containing over billions of webpages and tens of petabytes of data.

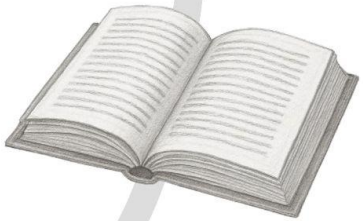
[1] [2]

Crawlers

```
robots.txt

User-agent: *
Disallow: /private/
```

url
https://climatewatch.org/
https://thoughtarchive.net/
https://devpulse.io/
https://musiclens.org/
https://historyatlas.co/



Store in Database

[3]

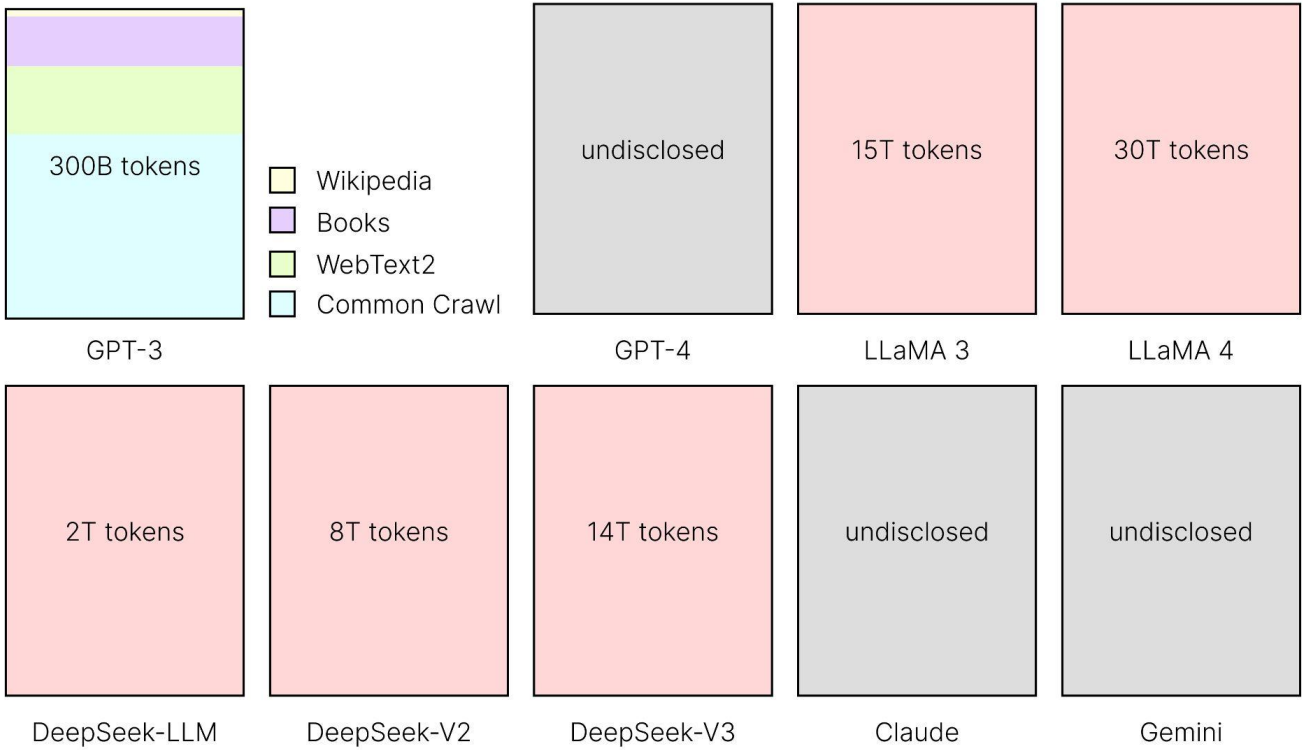
Corpora

Dataset	Source	Created by	Web Domains
C4	Common Crawl	Google	15,928,138
RefinedWeb	Common Crawl	LAION	33,210,738
Dolma	Mixed(web,books,code)	AI2	45,246,789

Overview of Dataset Structure

Text	url
This is a preliminary analysis of regional climate patterns...	https://climatewatch.org/
A curated list of the most influential philosophic...	https://thoughtarchive.net/
Packed with updated frameworks and performance...	https://devpulse.io/
Long before the digital age, music served...	https://musiclens.org/
Trade between ancient civilizations was more complex...	https://historyatlas.co/

Data Scale



[2] [4] [5] [6]

AI Crawlers

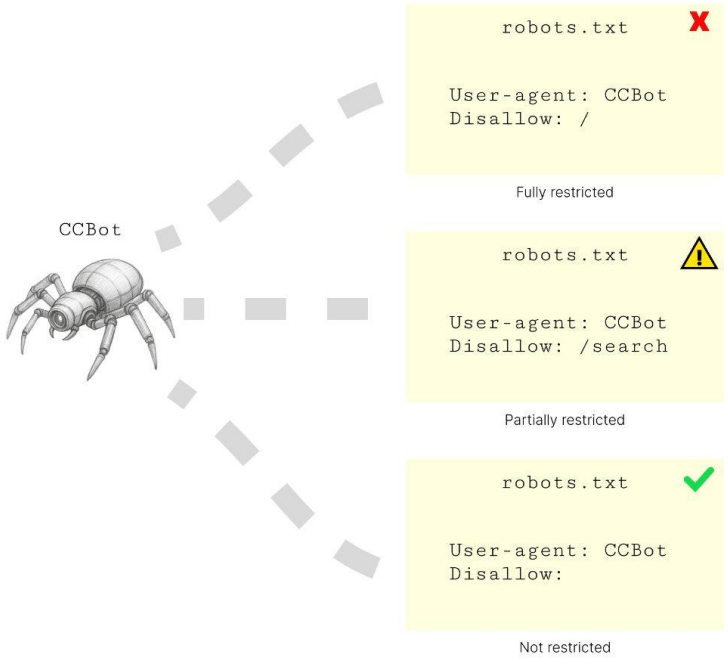
	Identifier	Status	Purpose
OpenAI	GPTBot	●	Training
	ChatGPT-User	●	Retrieval
Google	Google-Extended	●	Training
	Googlebot	●	Web Search
Anthropic	ClaudeBot	●	Training & Retrieval
	anthropic-ai	○	Training
	Claude-Web	○	Retrieval
Meta	FacebookBot	●	Training & Retrieval
Cohere	cohere-ai	○	Training & Retrieval
Common Crawl	CCBot	●	Training & Retrieval
Internet Archive	ia_archiver	●	Training & Retrieval

● Official crawler ○ Unofficial crawler

How are websites telling AI companies whether they can or can’t use their content, and how is that changing over time?

Restrictions

Restrictions by robots.txt



Restrictions by ToS

You may not use, copy, or reproduce any content from this website for the purpose of training, fine-tuning, or evaluating machine learning models or artificial intelligence systems.

in any means. Content on this site is provided for personal, educational, or non-commercial use only. Any commercial reproduction, redistribution, or repurposing is strictly prohibited.

Automated access or scraping of this website, including but not limited to web crawlers or bots, is prohibited unless explicitly authorized in writing by the site owner.

Methodology

2016 — 2024

ia_archiver

WayBack Machine

https://www.nytimes.com/robots.txt

https://www.nytimes.com/termsofservice

JAN	FEB	MAR	APR
MAI	JUN	JUI	SEP
OCT	NOV	DEC	JAN

Extracting robots.txt and ToS

Subset

Dataset	Source	Created by	Web Domains
C4	Common Crawl	Google	15,928,138
RefinedWeb	Common Crawl	LAION	33,210,738
Dolma	Mixed(web,books,code)	AI2	45,246,789

Subset	Source	Size	Criteria
$HEAD_{All}$	$HEAD_{C4,RW,Dolma}$	3,900	★
$HEAD_{C4}$	C4	2,000	★
$HEAD_{RW}$	RefinedWeb	2,000	★
$HEAD_{Dolma}$	Dolma	2,000	★
$RANDOM_{10k}$	Intersection*	10,000	🎲
$RANDOM_{2k}$	$RANDOM_{10k}$	2,000	🎲

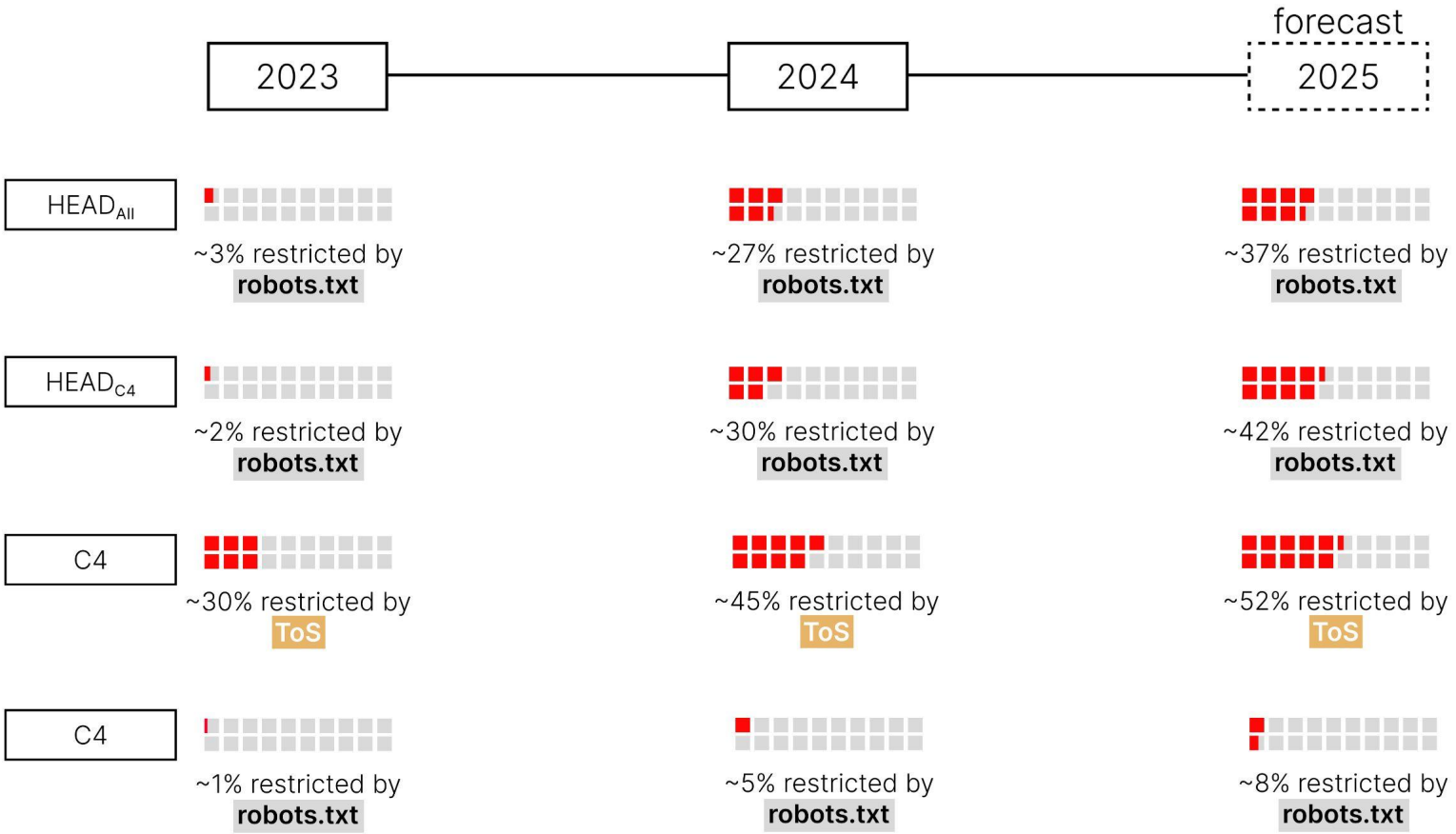
★ Web domains were ranked by number of tokens

🎲 Web domains were randomly selected to capture a wider sample

* The sample is drawn from the intersection of C4, RefinedWeb, Dolma.

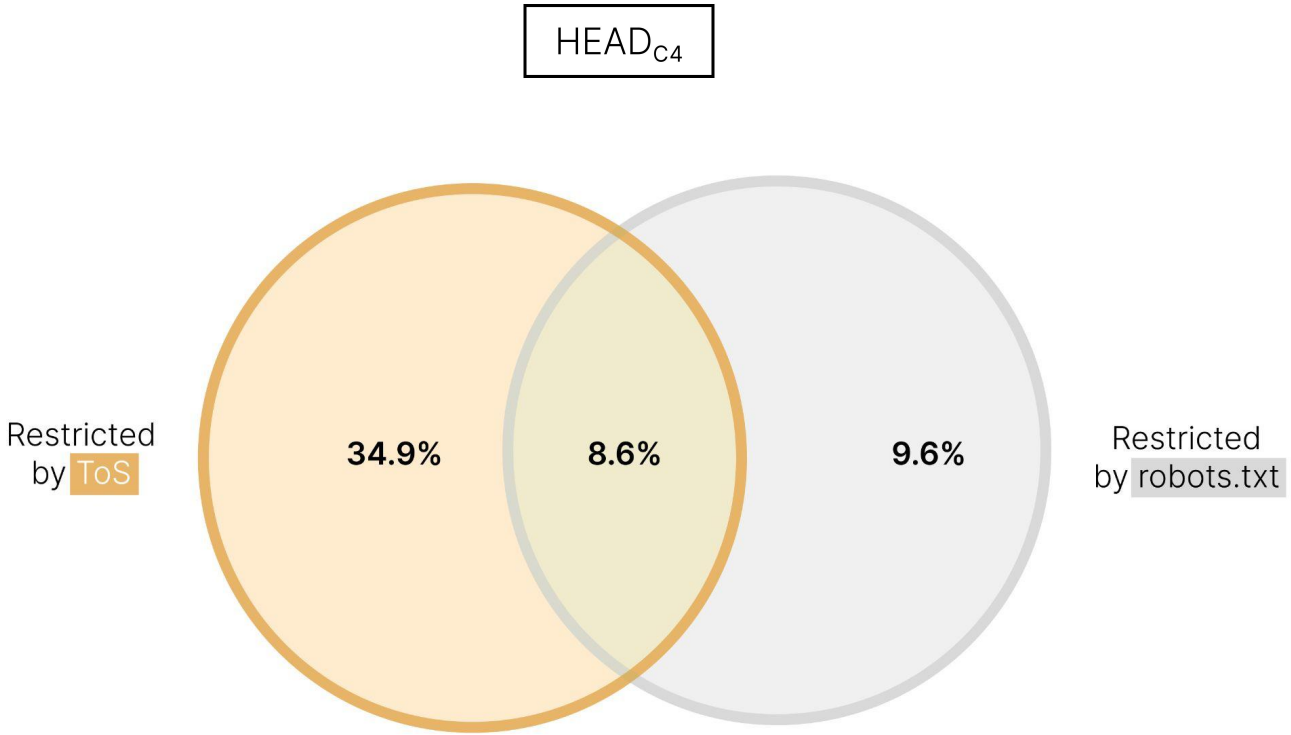
Universität Bremen

Consent in Crisis



Universität Bremen

Mixed Signals



Authors' Suggestions

...We observe robots.txt instructions which allow some AI organizations to crawl while restricting others, references to non-existent crawlers, and contradictions between the robots.txt and Terms of Service. Together, these issues point to the need for better preference signaling protocols.

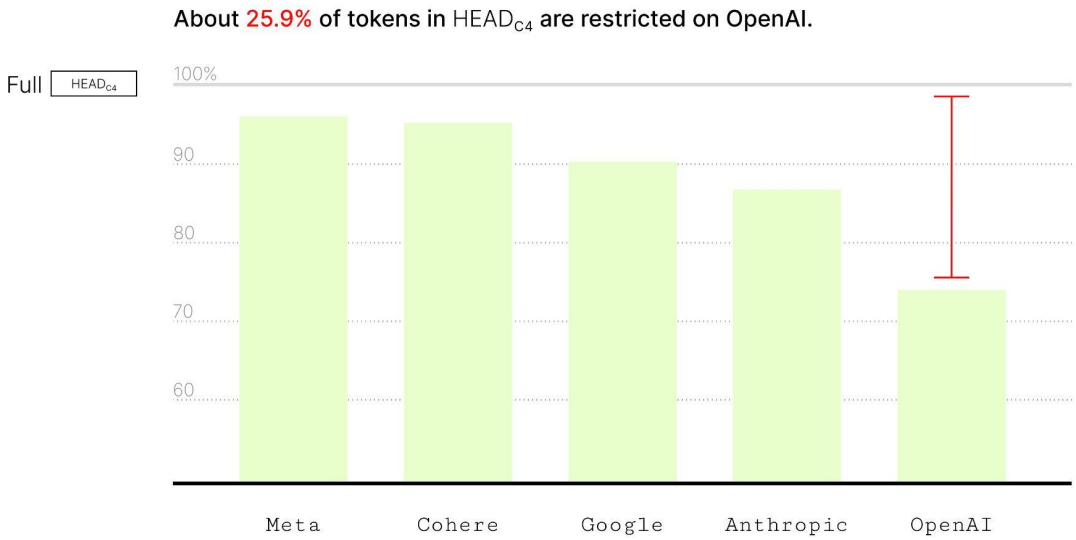
Consent in Crisis, Section 3.2

... An alternative scheme might give website owners control over how their webpages are used rather than who can use them. This would involve standardizing a taxonomy that better represents downstream use cases, e.g. allowing domain owners to specify that web crawling only be used for search engines, or only for non-commercial AI, or only for AI that attributes outputs to their source data. New commands could also set extended restriction periods given dynamic sites may want to block crawlers for extended periods of time, e.g. for journalists to protect their data freshness.

Consent in Crisis, Section 4. Discussion

Targeted Crawlers

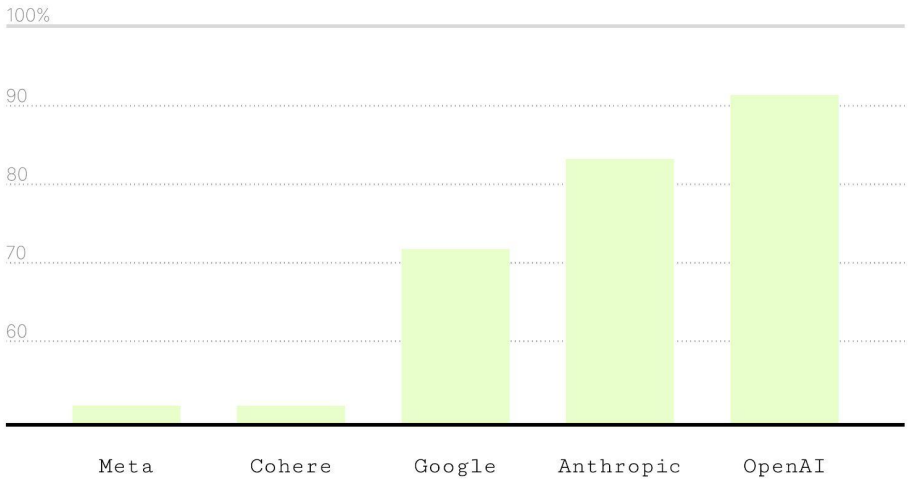
2024



Targeted Crawlers

2024

If any AI developer is restricted, OpenAI is also restricted in 91.5% of those cases.



Impact

The rise in restrictions will skew data representativity, freshness, and scaling laws. Prior work has forefronted scaling data as essential to frontier model capabilities. While the declining trend in consent will protect content creators' intentions, it would also challenge these data scaling laws. Not only would these restrictions reduce the scale of available data, but also the composition (away from news and forums), diversity, and representativeness of training data —biasing this data toward older content and less fresh content.

Recently, multiple AI developers have been accused of bypassing robots.txt opt-outs to scrape publisher websites. While it is not possible to confirm, in each case it appears AI systems may be distinguishing between crawling data for training, and crawling data to retrieve information for user questions at inference time. [...] However, creators may feel this violates the spirit of the opt-outs, especially if the opportunity to attribute sources is not taken.

Consent in Crisis, Section 4. Discussion

Plaintiff	Defendant	Key Issue
Authors, NYT et al.	OpenAI, Microsoft	Use of copyrighted text in LLM training
Ziff Davis	OpenAI	Use of proprietary articles despite robots.txt
Universal Music	Anthropic	Use of copyrighted lyrics in Claude AI
Authors	Meta	Use of pirated books (LibGen) in LLaMA training

[7] [8] [9] [10]

What makes the most commonly used websites in AI training different from the rest of the internet?

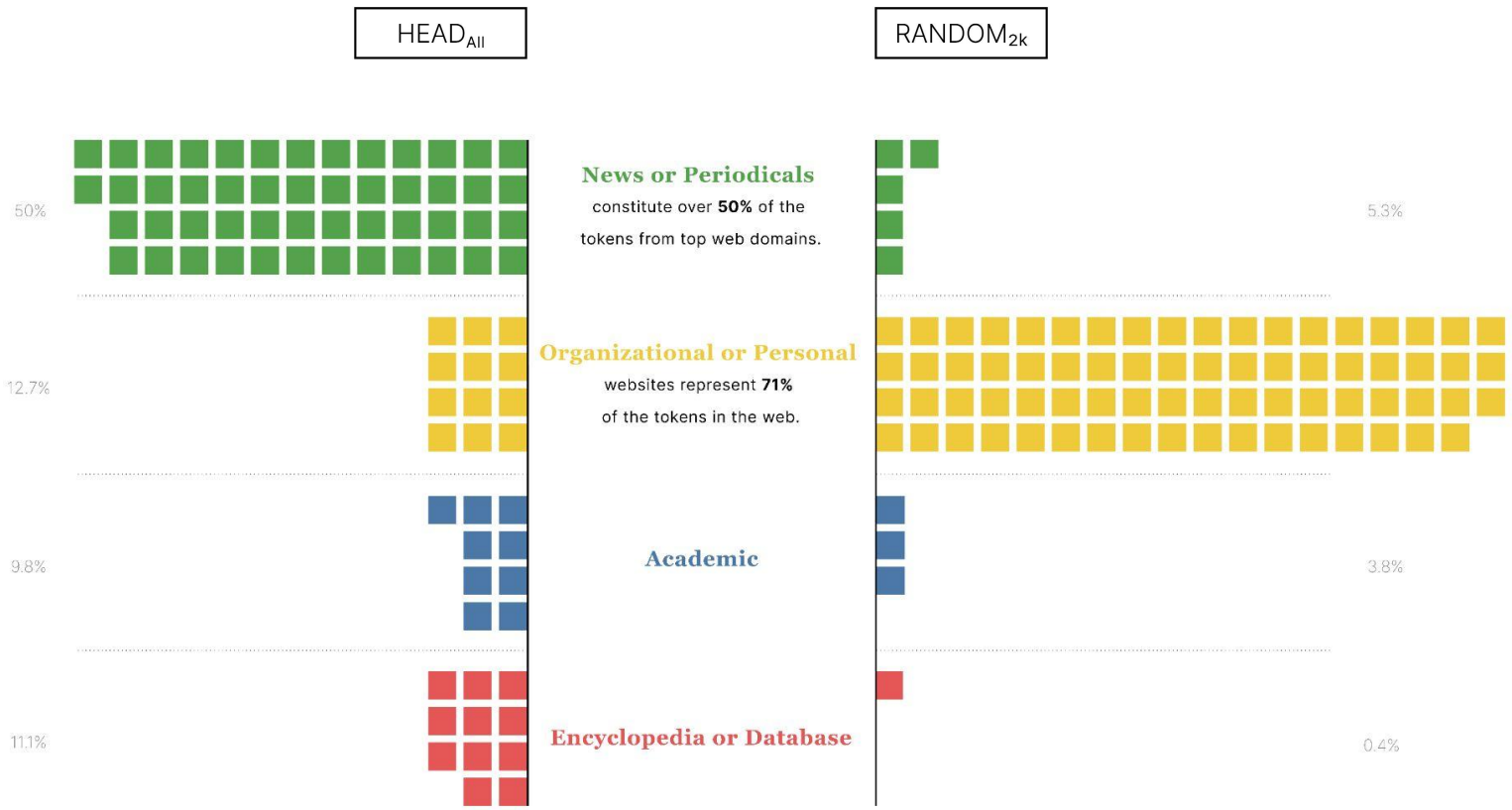
Annotators

Human annotators were hired to manually collect specific details on the websites in HEAD_{all} **and** RANDOM_{2k}

Attribute	Details
Content Modalities	Whether the web domain has images, videos, and standalone audio in addition to text
User Content	Whether the web domain hosts primarily content provided by users
Sensitive Content	Whether explicit, illicit, pornographic, or hate speech content is clearly present.
Paywall	Whether the web domain has use limits or any access gating behind a paywall
Advertisements	Whether the web domain has automatic advertisements embedded into any of its pages.
Purpose & Service	Website purpose: News, Academic, Social Media...

Universität Bremen

Content distribution



Features

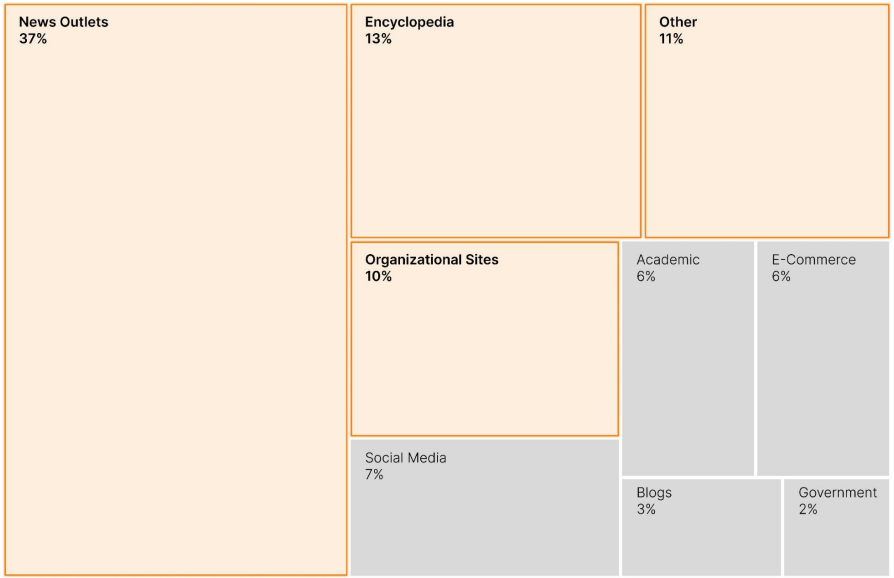
$HEAD_{All}$		$RANDOM_{2k}$
24.6%	15.4x Paywall	1.6%
41.8%	12.3x Audio	3.4%
53.2%	9.8x Advertisements	5.4%
58.7%	3.1x Video	18.9%

Is there a disconnect between the web content used to train AI models and the tasks users expect these models to perform?

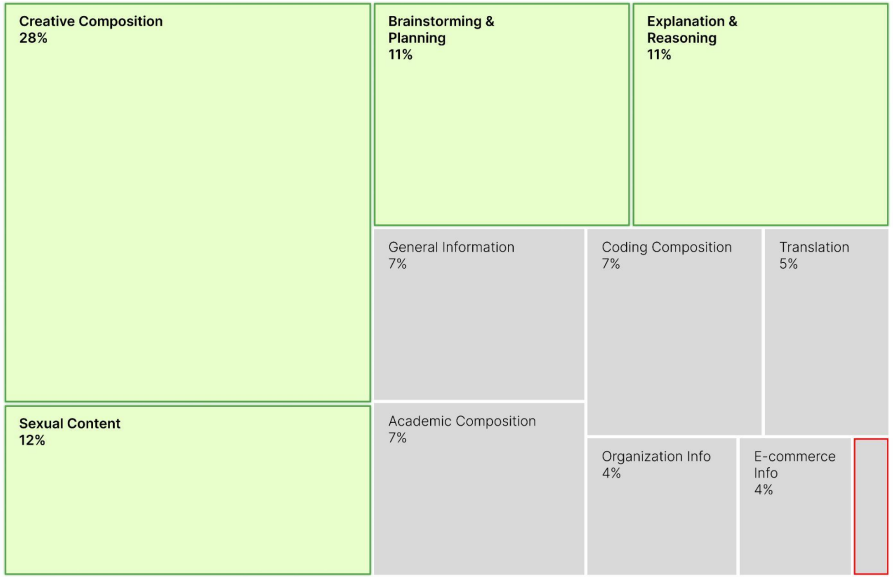
Universität Bremen

Training Gap

Training data tends to consist of content from commercial websites — not creative or conversational content.



In actual ChatGPT conversations, news represent only 1% of all the queries.



Fair Use

Although AI models like ChatGPT are trained heavily on news content, **their primary real-world uses are in creative and general inquiries—not news.**

This mismatch suggests that AI may not directly compete with some of its training sources, potentially lessening market harm—though this is not conclusive.

Impact

However, content creators precieve Generative AI as a competitor that will potentially undermining their source of revenue.

Smaller websites, lacking resources to block unwanted data scraping, **may withdraw from the open web or shift to closed platforms to protect their work.**

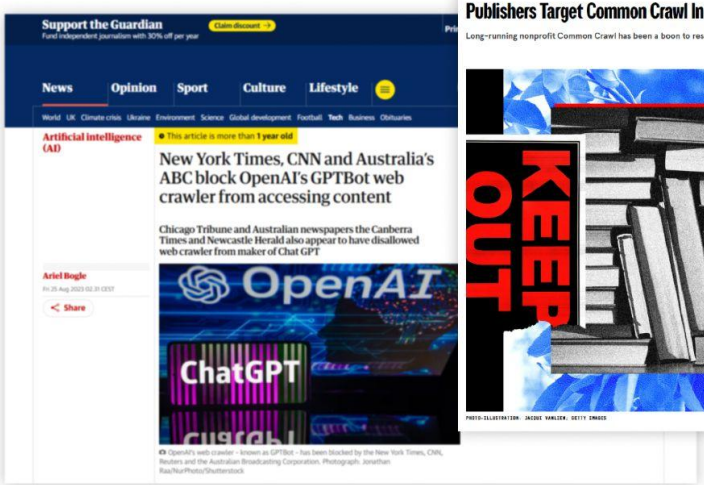
Open Web

Behind Paywalls

Impact

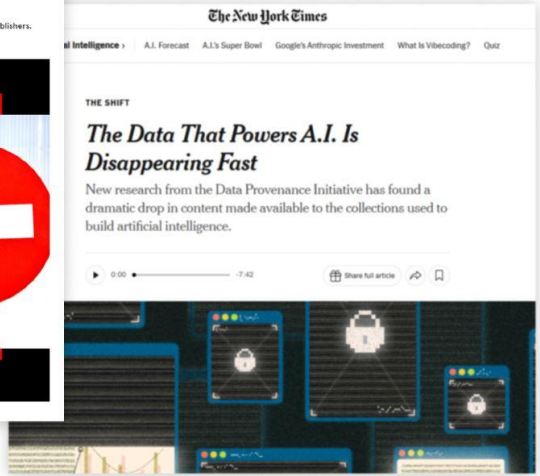
Websites are increasingly using blanket robots.txt rules to block all crawlers, as it's **difficult to separate commercial AI scrapers from non-commercial ones.**

These sweeping blocks often affect non-profit organizations like Common Crawl and the Internet Archive, **despite their role in public knowledge and research.**



The Guardian

WIRED



New York Times

Discussion

Will less data hurt model performance?

Scaling Laws

Our results foretell significant changes not only to AI data collection practices and data **scaling laws**, but also the structure of consent on the open web, which will impact more than AI developers.

Consent in Crisis, Section 1. Introduction

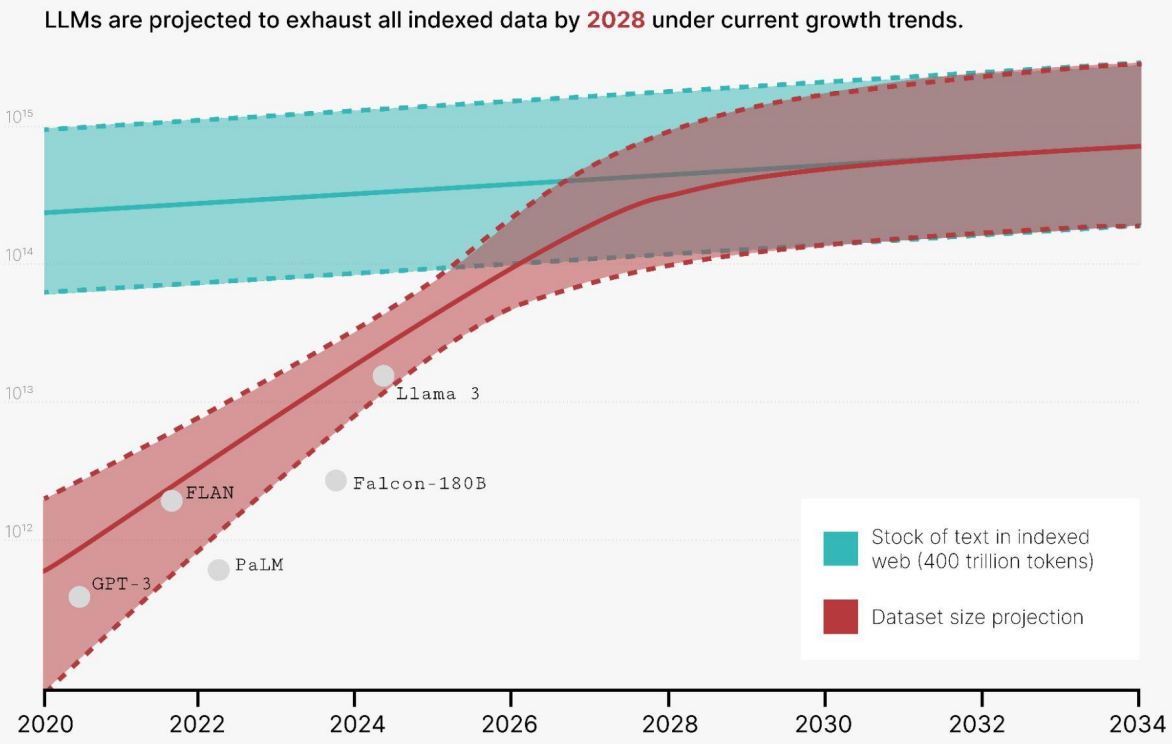
These trends illustrate a systematic rise in restrictions on data sources, which, where enforced or respected, will severely hamper the data **scaling practices** in the coming years—which have thus forth been responsible for the remarkable capability improvements.

Consent in Crisis, Section 3. Findings

Model Size *Dataset Size* *Compute*

[11]

Data Stock



Source: Will we run out of data? (2024)

But there are other ways to improve performance

Quality Tokens

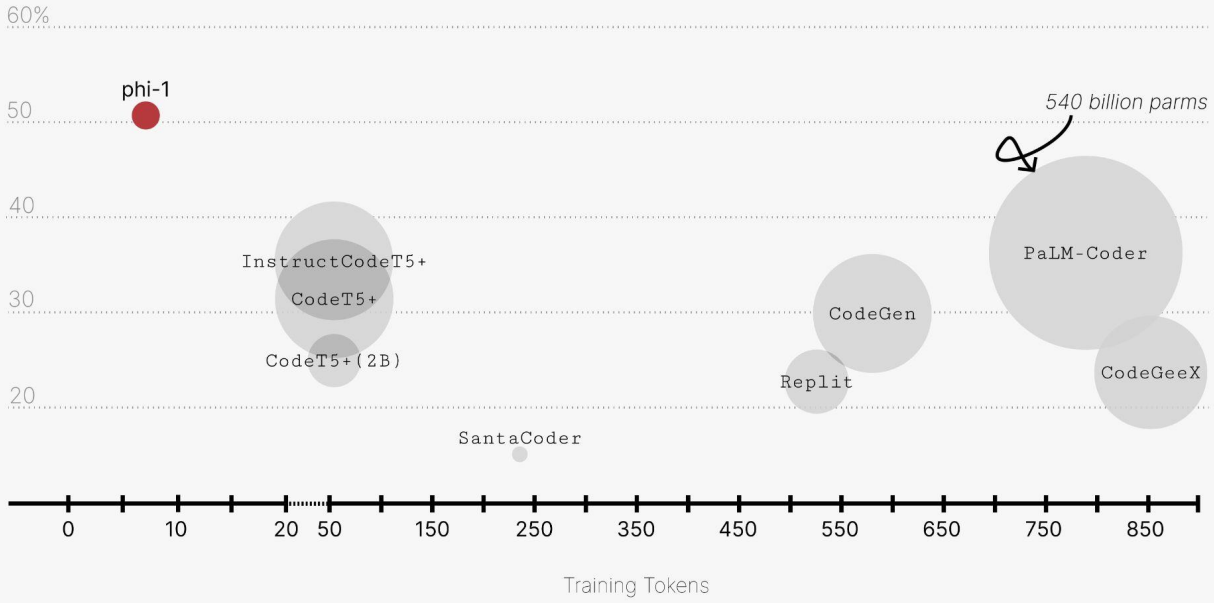
2023

We demonstrate **the power of high quality data in breaking existing scaling laws** by training a 1.3B-parameter model, which we call phi-1, for roughly 8 passes over 7B tokens (slightly over 50B total tokens seen) followed by finetuning on less than 200M tokens. Roughly speaking we pretrain on "textbook quality" data, both synthetically generated (with GPT-3.5) and filtered from web sources, and we finetune on "textbook-exercise-like" data. **Despite being several orders of magnitude smaller** than competing models, both in terms of dataset and model size (see Table 1), we attain 50.6% pass@1 accuracy on HumanEval and 55.5% pass@1 accuracy on MBPP (Mostly Basic Python Programs), which are one of the best self-reported numbers using only one LLM generation.

Textbooks Are All You Need, Section 1

Performance

Trained on only **7 billion** tokens, phi-1 out performers other models in HumanEval.



Tool Learning

Given the fact that LLMs are trained on datasets comprising general knowledge, **they often exhibit deficiencies in specialized domains.** While LLMs demonstrate robust problem-solving capabilities for basic mathematical problems, excelling in operations such as addition, subtraction, and exhibiting reasonable proficiency in multiplication tasks, their abilities significantly decline when confronted with **division, exponentiation, logarithms, trigonometric functions,** and other more complex composite functions.

[...]For example, LLMs can use **online calculators** or mathematical tools to perform complex calculations, solve equations, or analyze statistical data. Additionally, the integration of external programming resources such as **Python compilers** and interpreters allows LLMs to receive code execution feedback, which is essential for refining code to align with user requirements and to optimize the code generation. Moreover, LLMs can also leverage tools in fields such as chemistry, biology, economics, medicine, and recommendation systems to **enhance their domain-specific expertise**.

Tool Learning with Large Language Models: A Survey

Reasoning Models

Chain of thought and extended inference time, used by Reasoning Models lead to better performance across benchmarks

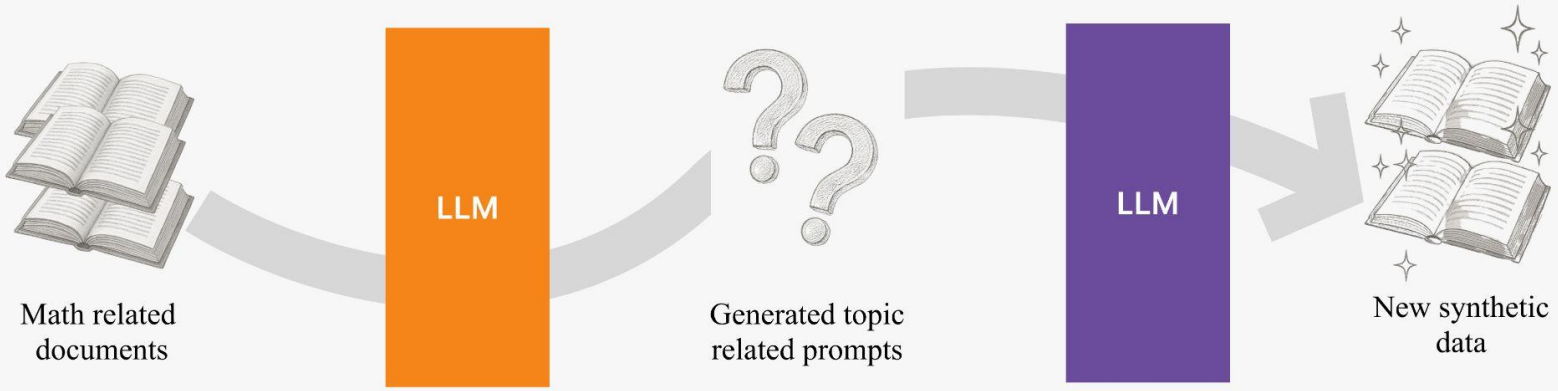
AIME	
o3 Mini	86.5%
Gemini 2.5 Pro Exp	85.8%
o3	85.3%
Grok 3 Mini Fast Beta High Reasoning	85.0%
o4 Mini	83.7%
DeepSeek R1	74.0%
o1	71.5%
Grok 3 Mini Fast Beta Low Reasoning	70.6%
Grok 3 Beta	58.7%
DeepSeek V3 03/24/2025	52.2%
GPT 4.1 mini	49.4%
Claude 3.7 Sonnet (Thinking)	44.4%
GPT 4.1	39.8%
Gemini 2.0 Flash (001)	29.8%
DeepSeek V3	27.5%

GPQA	
o3	83.6%
Gemini 2.5 Pro Exp	80.3%
Grok 3 Mini Fast Beta High Reasoning	79.0%
Claude 3.7 Sonnet (Thinking)	75.3%
o3 Mini	75.0%
o4 Mini	74.5%
Grok 3 Beta	73.7%
o1	73.0%
Grok 3 Mini Fast Beta Low Reasoning	72.7%
Llama 4 Maverick	67.7%
Claude 3.7 Sonnet	67.4%
GPT 4.1 Mini	67.4%
Gemini 2.0 Flash (001)	65.2%
GPT 4.1	64.6%
DeepSeek V3	61.1%

MMMU	
Gemini 2.5 Pro Exp	81.5%
o3	80.0%
o4 Mini	79.6%
o1	77.7%
Claude 3.7 Sonnet (Thinking)	76.0%
GPT 4.1	72.6%
Llama 4 Maverick	72.6%
Gemini 2.0 Flash (001)	71.9%
Claude 3.7 Sonnet	71.6%
Claude 3.7 Sonnet Latest	68.9%
GPT 4.1 Mini	68.9%
GPT 4o	68.1%
Grok 2 Vision	66.5%
Llama 4 Scout	66.5%
Gemini 1.5 Pro	66.1%

[12]

Synthetic Data



Source: Scaling Laws of Synthetic Data for Language Models (2025)

References

[1]	https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
[2]	https://arxiv.org/abs/2005.14165
[3]	https://en.wikipedia.org/wiki/Web_crawler
[4]	https://en.wikipedia.org/wiki/List_of_large_language_models
[5]	https://ai.meta.com/blog/meta-llama-3/
[6]	https://ai.meta.com/blog/llama-4-multimodal-intelligence/
[7]	https://www.theguardian.com/books/2025/apr/04/us-authors-copyright-lawsuits-against-openai-and-microsoft-combined-in-new-york-with-newspaper-actions
[8]	https://www.theverge.com/news/656044/ziff-davis-sues-openai-ign-cnet-pcmag
[9]	https://pitchfork.com/news/music-publishers-sue-ai-company-anthropic-for-copyright-infringement/
[10]	https://www.ft.com/content/b1f4965f-6ea6-4afd-968b-76bb2f7acfc4
[11]	https://arxiv.org/abs/2203.15556
[12]	https://www.vals.ai/benchmarks

Thank you!