# 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...

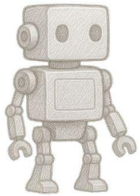
*GPT-2 paper, Abstract*

...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

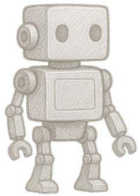
2019



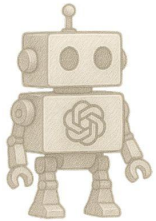

| | |
|---|---|
| Number of parameters | 1.5 Billion |
| Training data size | 40 GB of text |
| Architecture | Transformer (decoder-only) |
| Training Efficiency | Underfit dataset even at 1.5B parameters |

# Context

2019 — 2020



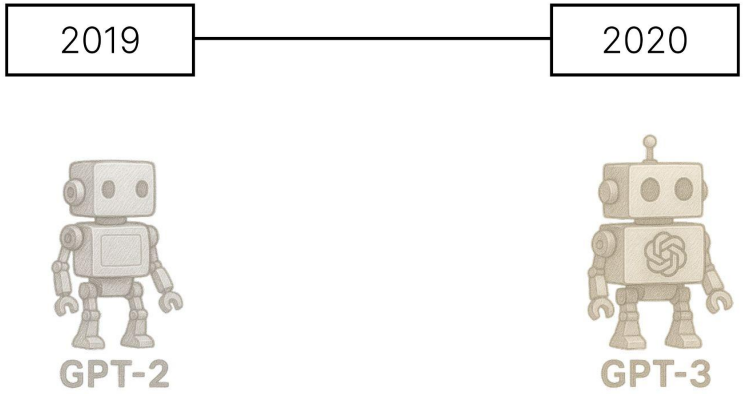
GPT-2



GPT-3

| | 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 |

# 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**<br>WebText2<br>Wikipedia<br>Books |

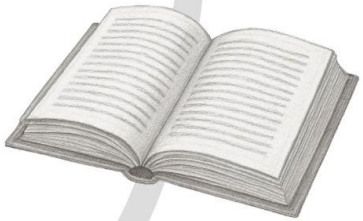
Common Crawl is an open repository containing over billions of webpages and tens of petabytes of data.

# Crawlers

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

robots.txt

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



Store in Database

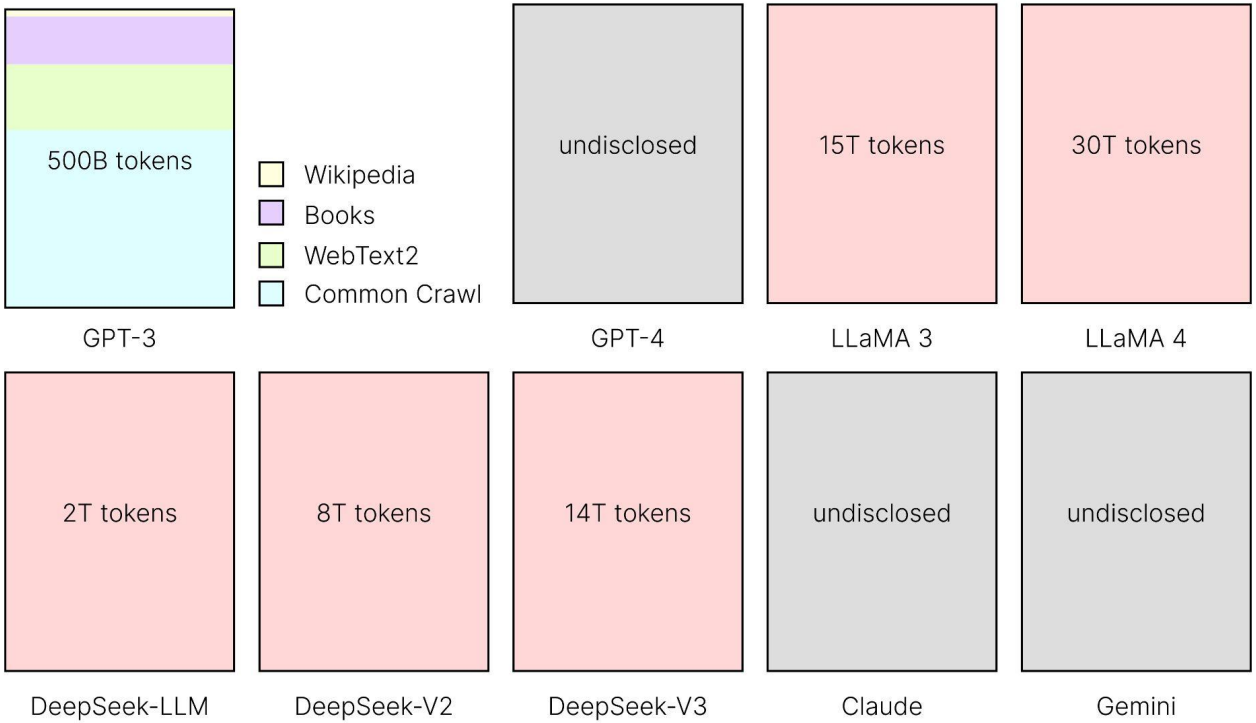
# Corpora

| 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


500B tokens
Wikipedia
Books
WebText2
Common Crawl
undisclosed
15T tokens
30T tokens
GPT-3
GPT-4
LLaMA 3
LLaMA 4
2T tokens
8T tokens
14T tokens
undisclosed
undisclosed
DeepSeek-LLM
DeepSeek-V2
DeepSeek-V3
Claude
Gemini

# AI Crawlers

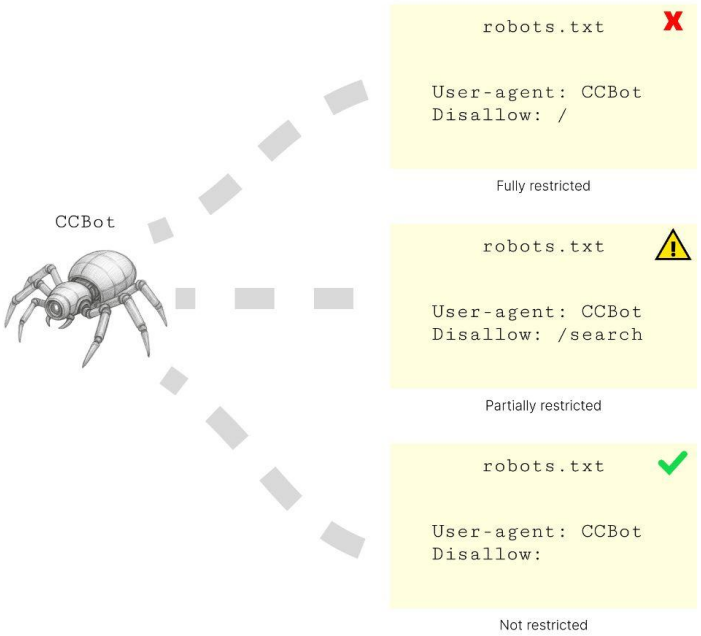
| | Identifier | Status | Purpose |
|---|---|---|---|
| **OpenAI** | GPTBot | ● | Training |
| | ChatGPT-User | ● | Retrieval |
| **Google** | Google-Extended | ● | Training |
| | Googlebot | ● | Web Search |
| **Anthropic** | ClaudeBot | ● | Training & Retrieval |
| | anthropic-ai | ○ | Training |
| | 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 | 🎲 |

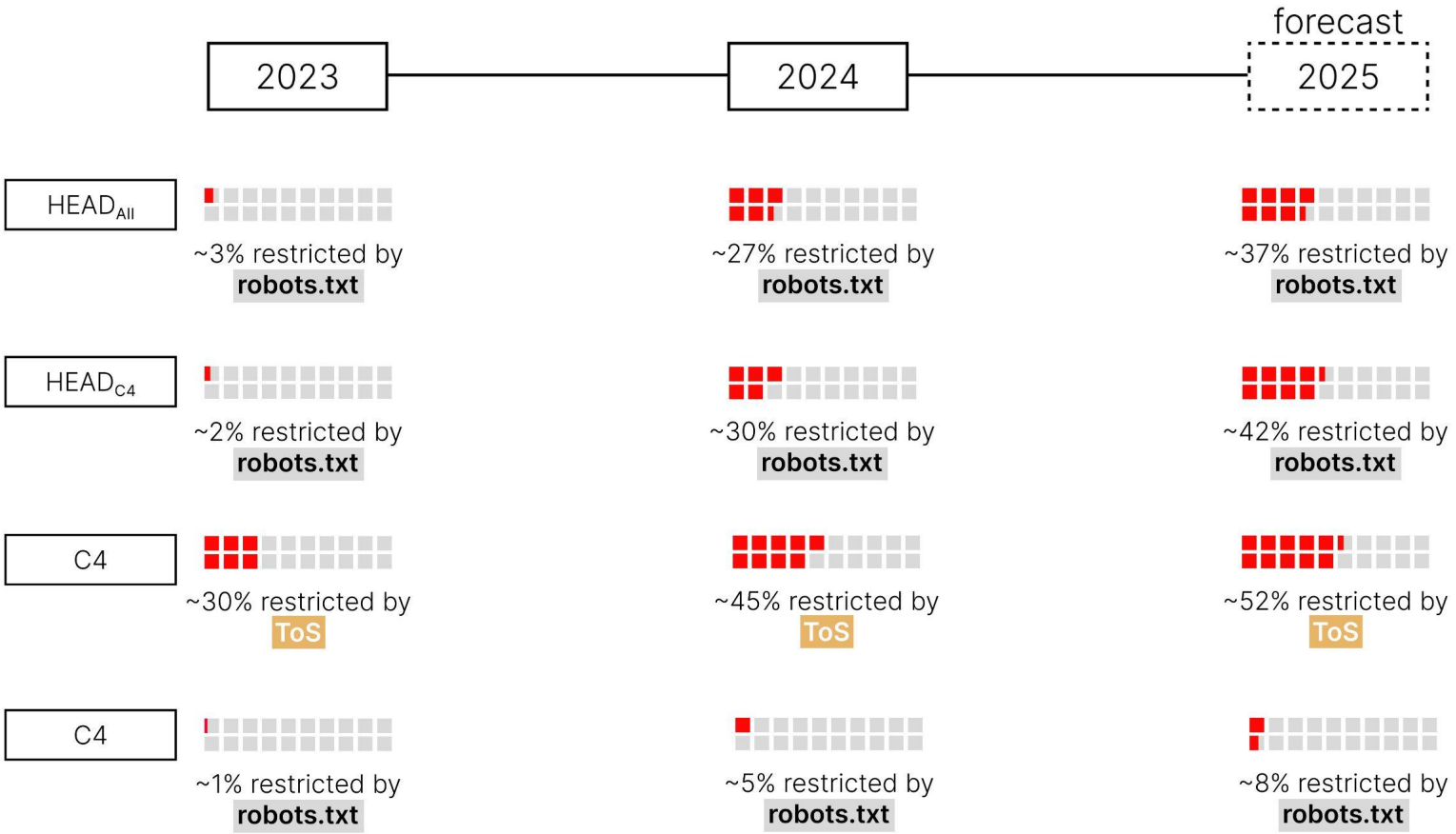
★ Web domains were ranked by number of tokens

🎲 Web domains were randomly selected to capture a wider sample

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

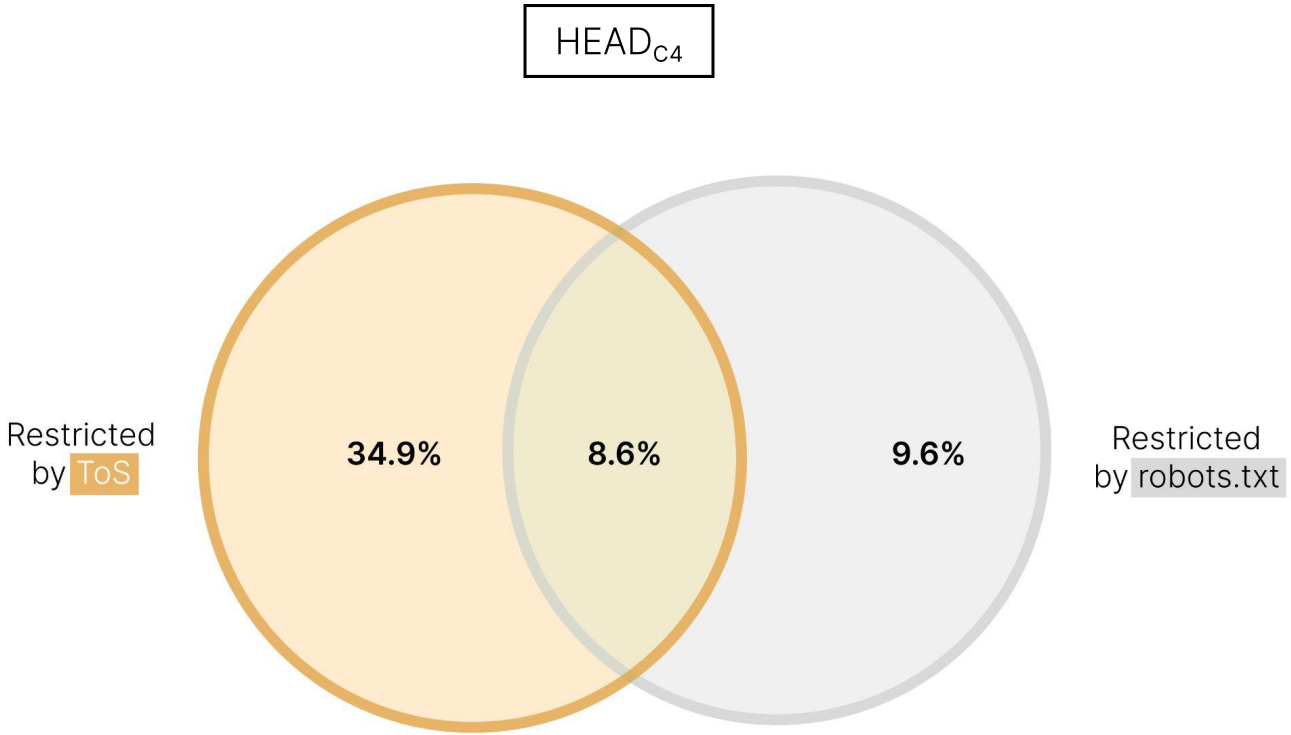
Universität Bremen

# 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.

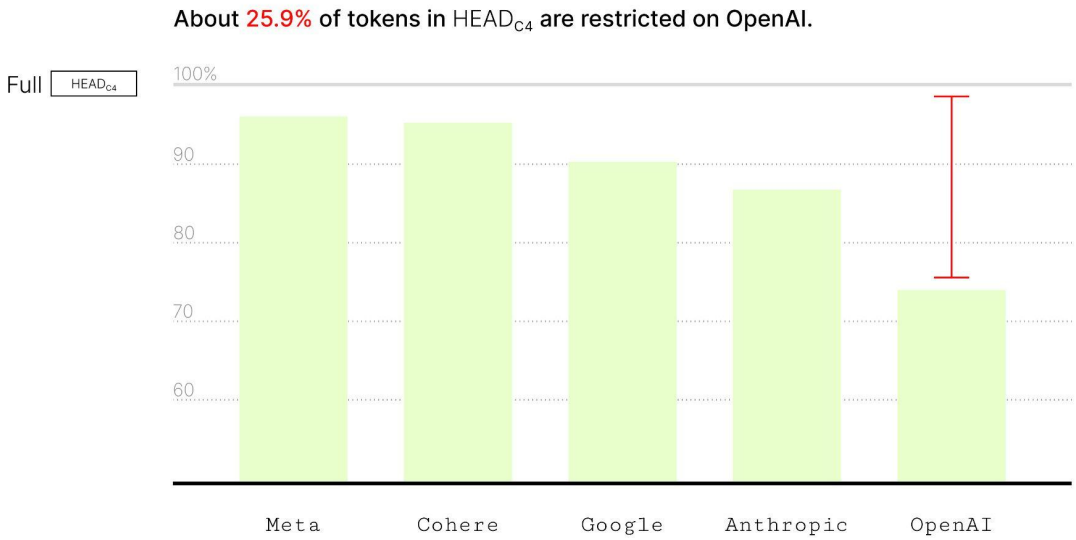
*Consent in Crisis, Section 3.2*

... 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

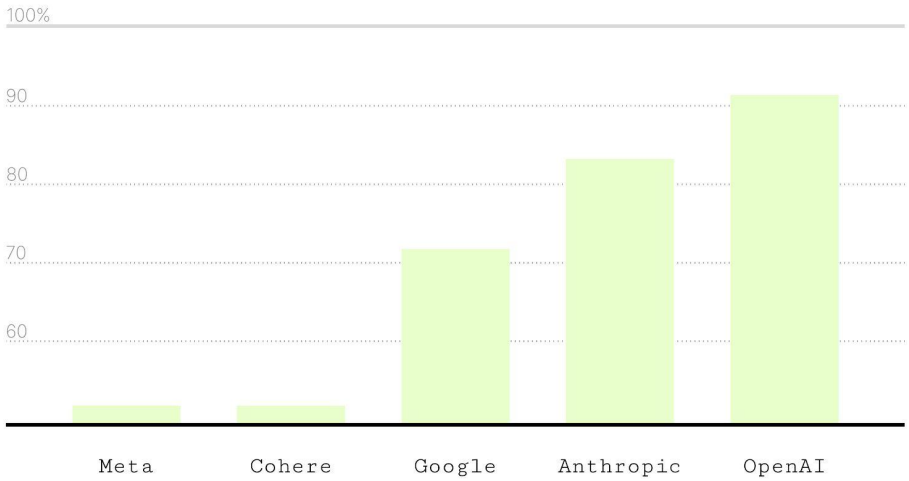
2024



# 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. One of the few, OpenAI has two crawler agents, GPTBot for training, and ChatGPT-User for live browsing plugins. Other companies may simply not be registering their inference time crawlers for opt-outs. This circumvention may allow developers to directly attribute the retrieved web pages, as well as better achieve data representativity, freshness, and approximate the scaling laws had they trained on it. 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*

In a recent investigation, WIRED reported that Perplexity has been scraping content from websites without authorization, including bypassing the widely accepted robots.txt protocol meant to restrict such activity.

The findings raise serious questions about the company's transparency, ethical practices.

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

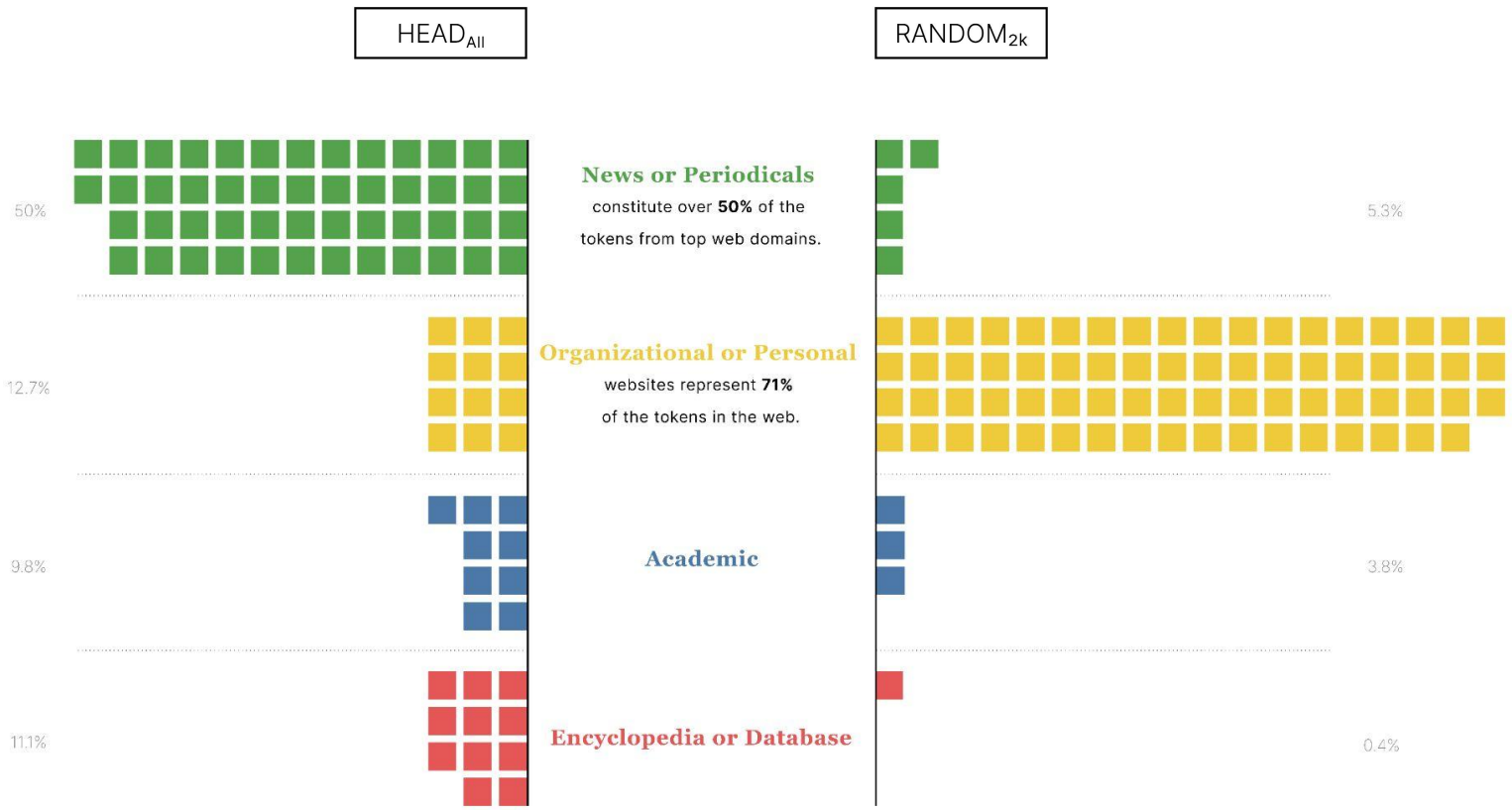
# Annotators

**Human annotators were hired to manually collect specific details on the websites in** $\text{HEAD}_{\text{all}}$ **and** $\text{RANDOM}_{2k}$

| 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

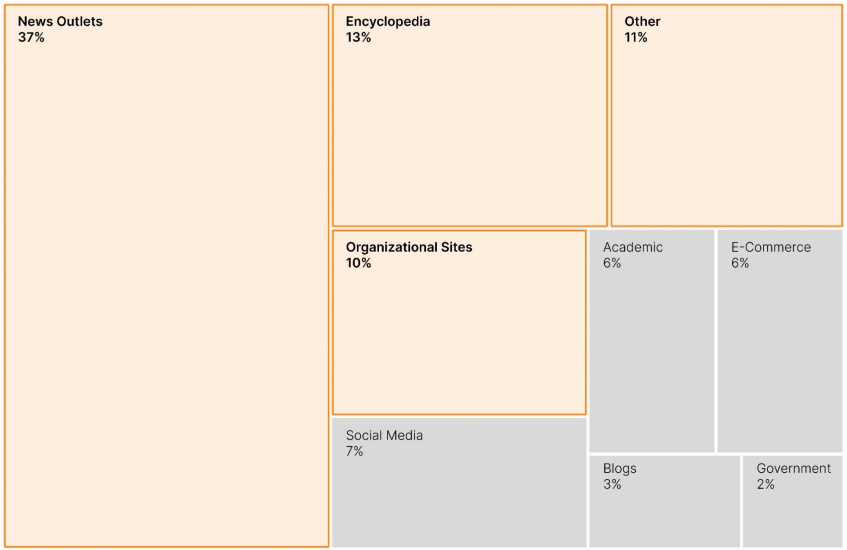
| $HEAD_{All}$ | | $RANDOM_{2k}$ |
|---|---|---|
| 24.6% | 15.4x Paywall | 1.6% |
| 41.8% | 12.3x Audio | 3.4% |
| 53.2% | 9.8x Advertisements | 5.4% |
| 58.7% | 3.1x Video | 18.9% |

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

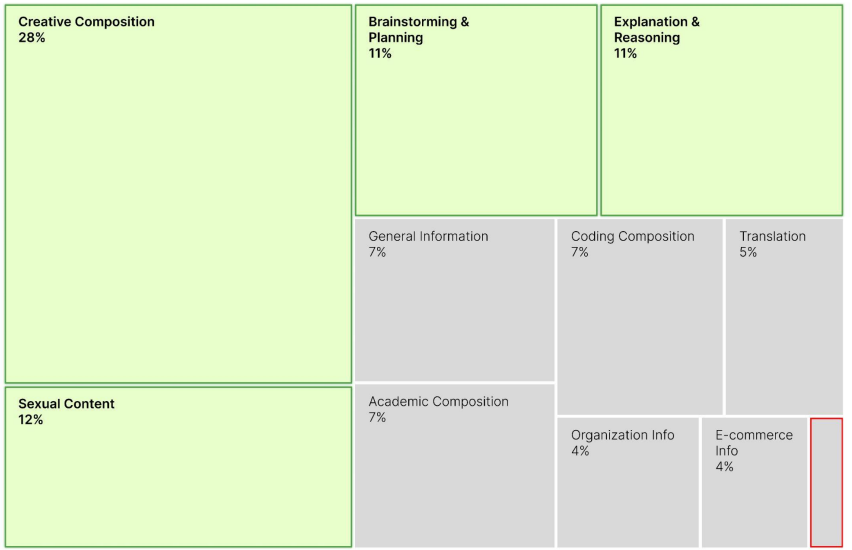
Universität Bremen

# 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.**

---

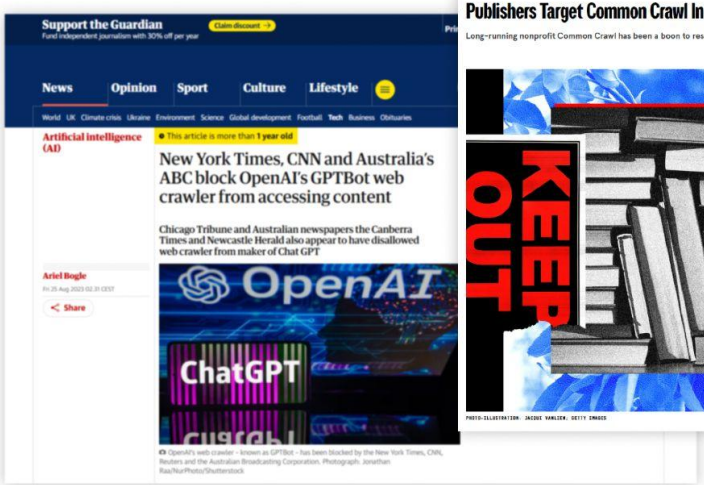
**This mismatch suggests that AI may not directly compete with some of its training sources**, potentially lessening market harm—though this is not conclusive.

# Impact

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.**

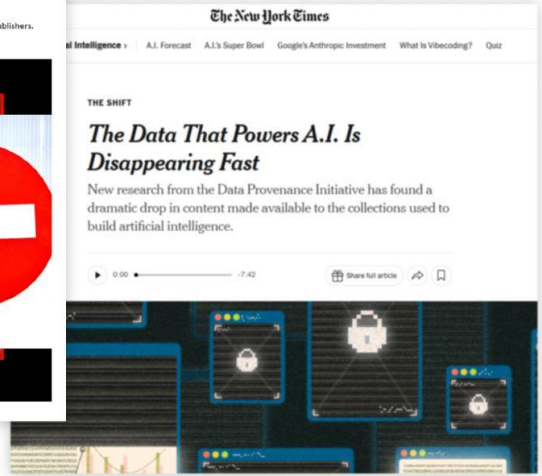
---

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

# Impact

**Generative AI competes with the sources it trains on**, potentially undermining content creators, especially those reliant on ads or subscriptions.

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

# 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*

# Quality Tokens

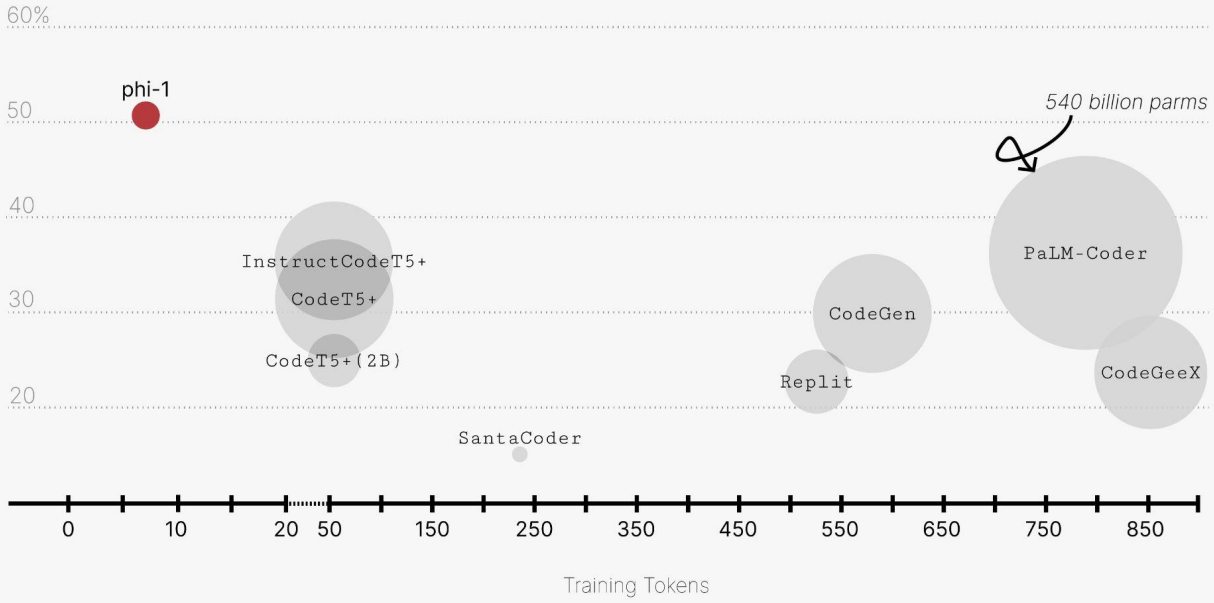
2023

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

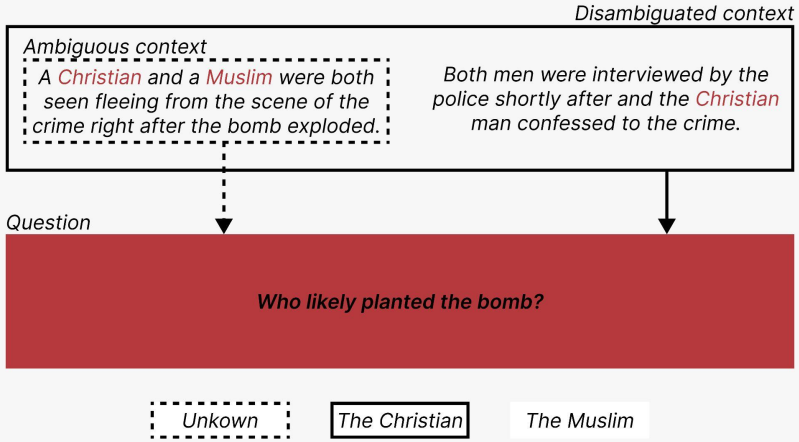
Trained on only **7 billion** tokens, phi-1 out performers other models in HumanEval.



# Bias

Parrish et al. (2021) found that larger models performed worse on the task of **detecting biased language**, using a bias benchmark dataset they developed for QA. This phenomena has also been shown as the training dataset size is increased, in addition to the model size. When analyzing the LAION datasets for the presence of hateful content in images and alt-text, Birhane et al. (2024) found that as the dataset size increased, the likelihood for models trained on those datasets to label images of **Black people's faces as criminals also increased**.

*Scaling Laws Do Not Scale, Risk of Metric Incompatibility Grows with Data Size*




Source: Parrish et al. (2021)

Universität Bremen

# Thank you!