

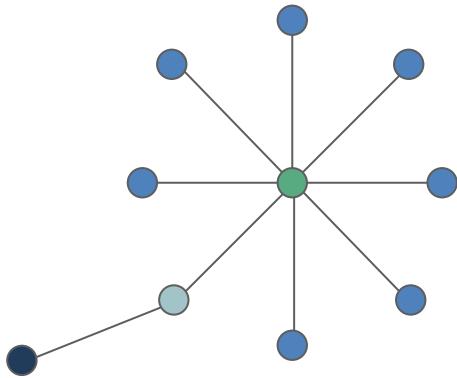
Introduction to Common Crawl

AI Alliance @ IBM One Madison
UN Open Source Week 2025

What is Common Crawl?

What is Common Crawl?

- Free and open corpus containing > 275 billion pages over > 18 years
- 501(c)(3) nonprofit, started in 2007 by Gil Elbaz
- Hosted on AWS S3 as an Open Data set, enabled by the Open Data Sponsorship Program
- Cited in > 10,000 research papers
- 3–4 billion new pages added each month
- Steered by link-based Harmonic Centrality ranks
- Web Graphs showing the structure and connectivity of the web
- We also release the accompanying ranking data for hosts and domains



$$H(v) = \sum_{u \neq v} \frac{1}{d(v, u)}$$

Where $H(v)$ is the **Harmonic Centrality** of vertex v ,
and $d(v, u)$ is the shortest path distance between vertices v and u .

Top 1000 Ranks					
Domain					
<input type="text" value="cc-main-2025-mar-apr-may"/> ▼					
Search table...					
#harmonicc_pos	#harmonicc_val	#pr_pos	#pr_val	#host_rev	#n_hosts
1	3.1209718E7	2	0.01395099273569498	com.googleapis	3319
2	3.0959424E7	3	0.008752723837883454	com.facebook	3504
3	3.0553926E7	1	0.01460402646460373	com.google	16077
4	2.8184966E7	5	0.005740515998492531	com.instagram	822
5	2.730949E7	4	0.006533040668666358	com.googletagmanager	46
6	2.7058282E7	7	0.004505815248064307	com.youtube	1765
7	2.6530742E7	9	0.004187459812149493	com.twitter	700
8	2.5745642E7	6	0.004723508361219754	com.gstatic	202
9	2.5189044E7	11	0.003225250938409697	com.linkedin	723
10	2.5145868E7	8	0.004266825384363396	org.gmpg	3

« Page 1 of 100 »

9.5 PiB

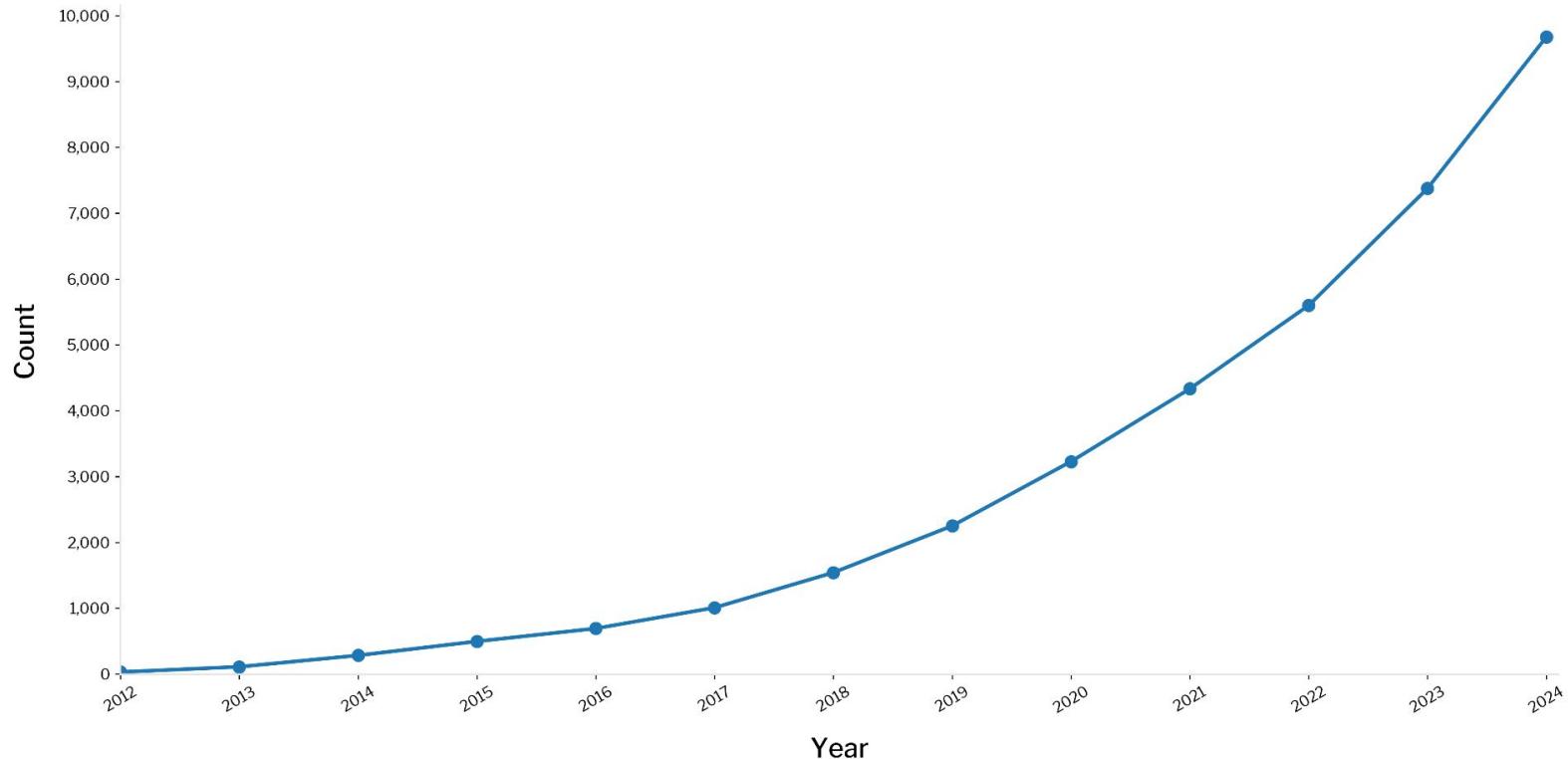
Dataset size as of May 2025.

About 3.45 billion copies of “War and Peace”.
9.5 PiB ≈ 10.696 PB.

What's it used for?

- Natural language processing
- Web science
- Information retrieval
- Semantic web
- Security research
- Language modelling ...

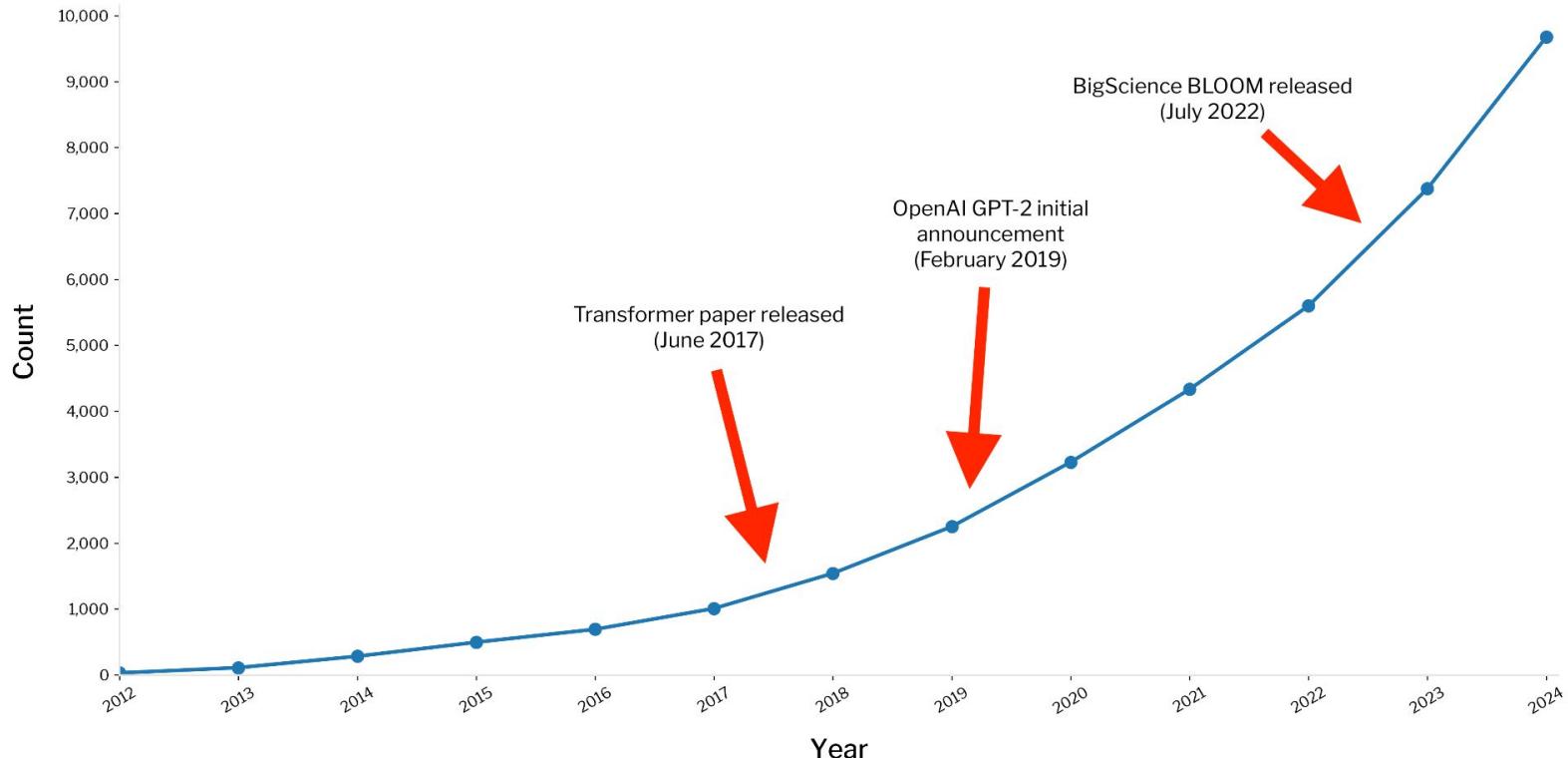
Plot of Common Crawl citations (cumulative) in Google Scholar until January 2025



<https://commoncrawl.org/research-papers>

<https://huggingface.co/datasets/commoncrawl/citations>

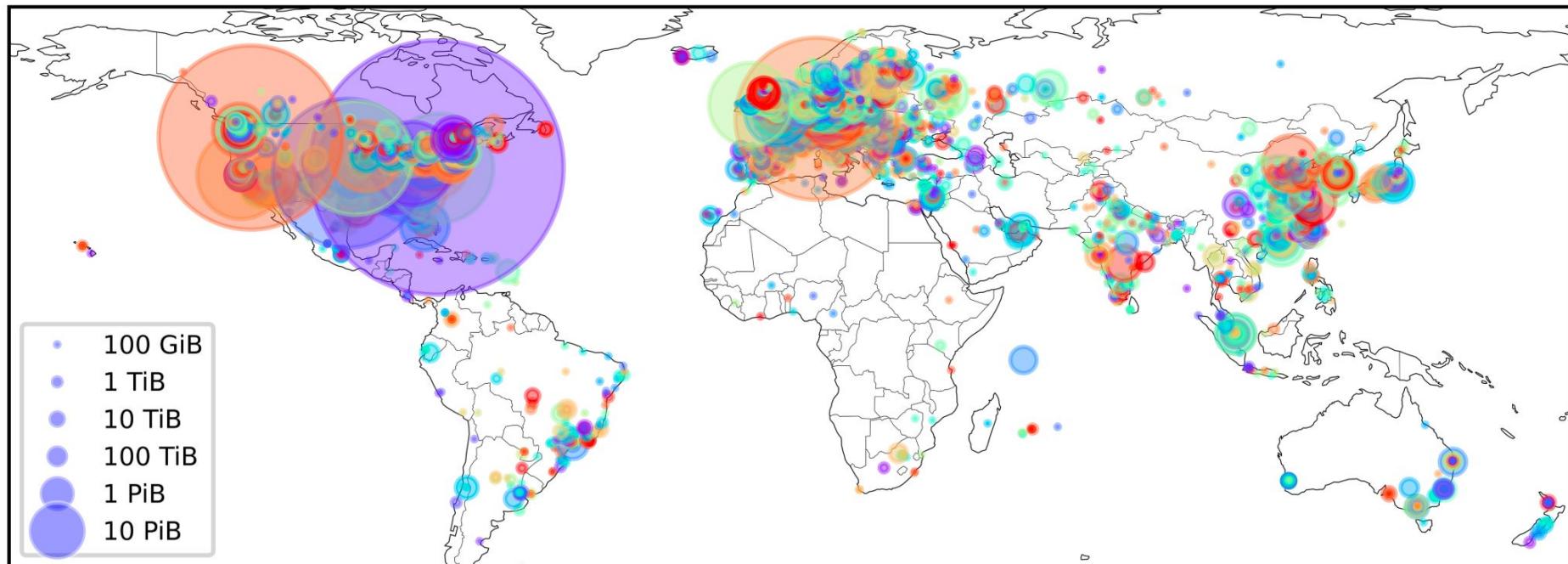
Plot of Common Crawl citations (cumulative) in Google Scholar until January 2025



<https://commoncrawl.org/research-papers>

<https://huggingface.co/datasets/commoncrawl/citations>

Data Requested by GeoIP

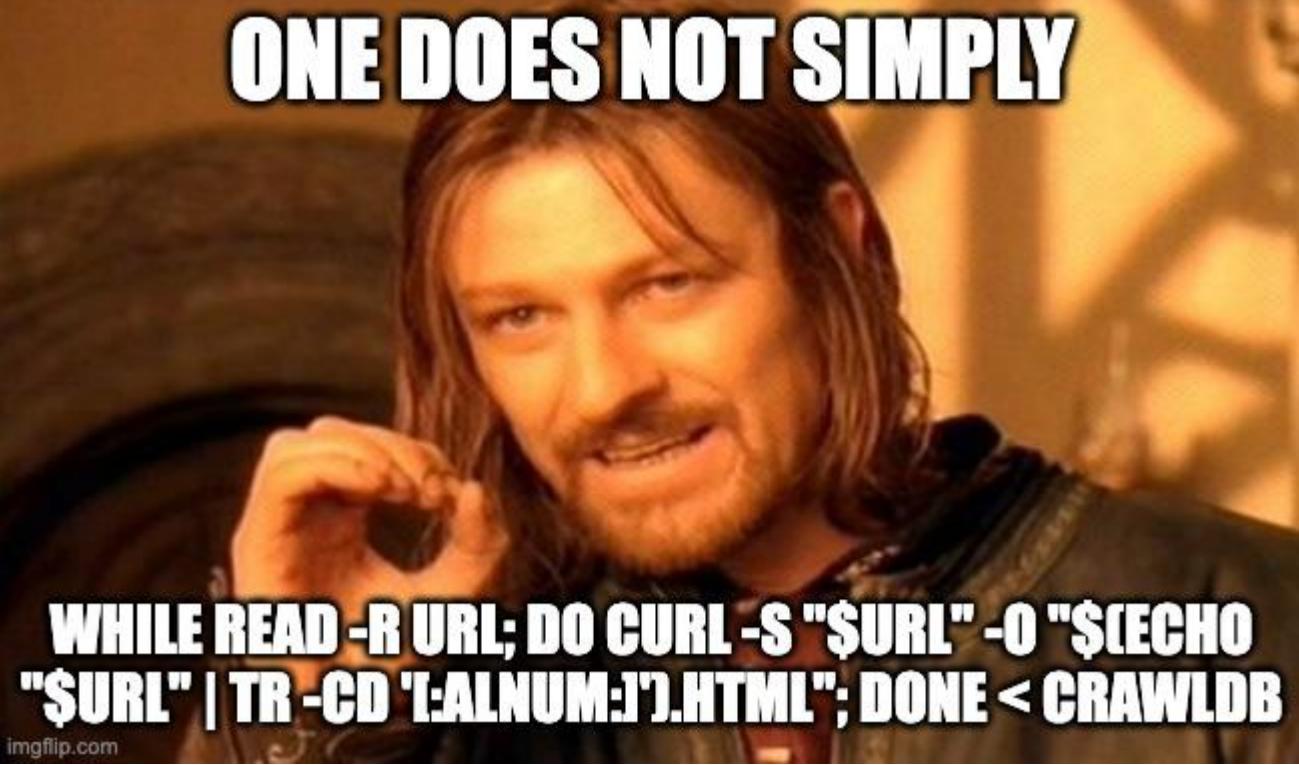


>1 EiB

Total egress in 2024.

One quintillion bytes, or roughly 372 billion copies of "War and Peace".

Crawler Perspective



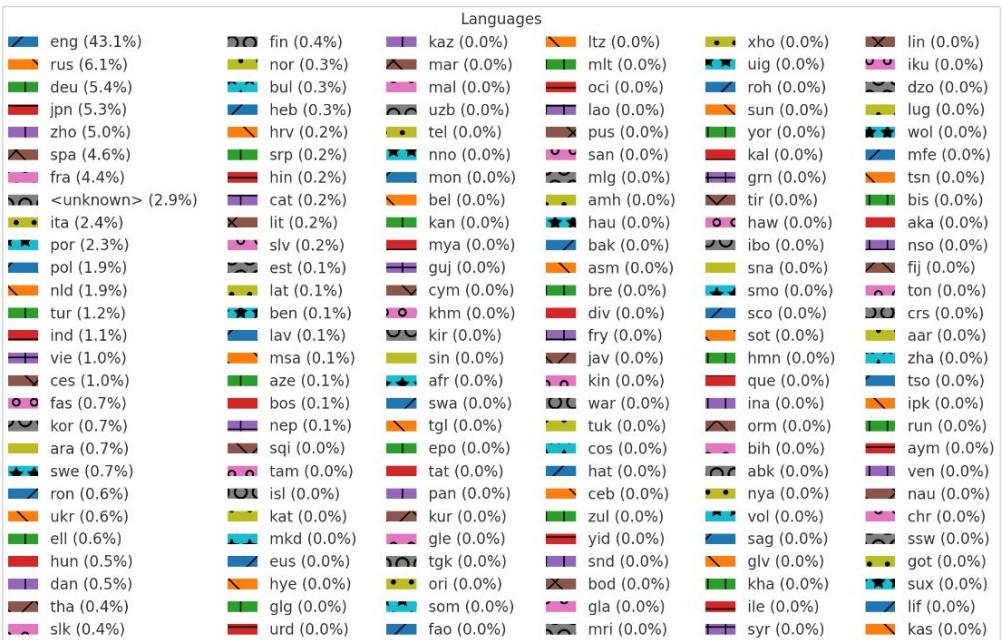
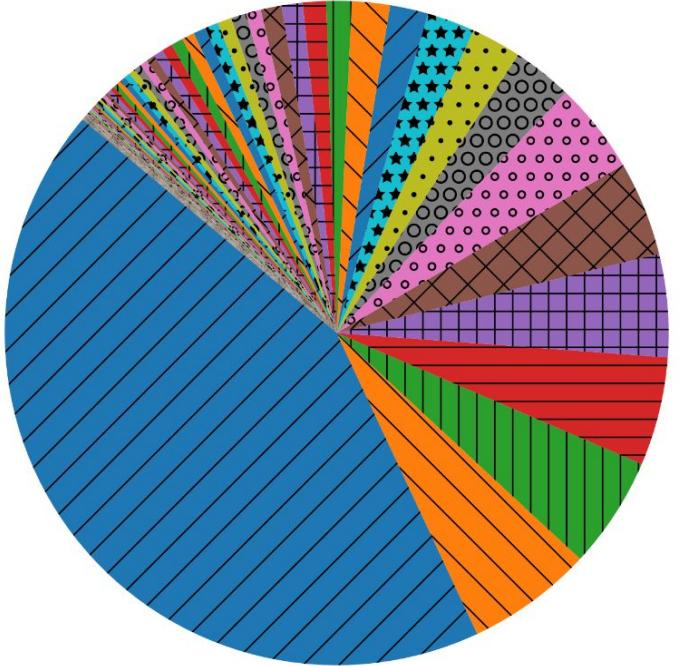
ONE DOES NOT SIMPLY

**WHILE READ -R URL; DO CURL -S "\$URL" -O "\${ECHO
"\$URL" | TR -CD '[:ALNUM:]'.HTML}"; DONE < CRAWLDB**

Crawler Perspective

- Politeness is crucial to long-term sustainability, ignoring consent leads to backlash [1]
- Be 🙌 transparent 🙌 to 🙌 build 🙌 trust
- Preference signals must be machine-readable [2]
- Purposes differ widely in the “crawl space” [3]
- Opt-Out vs. Opt-In [4]

Language Initiatives



Detected language distribution (averaged) in the last three crawls using CLD2 as the language identifier
(CC-MAIN-2024-46, CC-MAIN-2024-51, and CC-MAIN-2025-05)

<https://commoncrawl.github.io/cc-crawl-statistics/plots/languages>
<https://github.com/CLD2Owners/cld2>

commoncrawl/web-languages

github.com/commoncrawl/web-languages/tree/main

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

web-languages Public Edit Pins Unwatch 8 Fork 44 Starred 43

main Go to file Code

About

Crowd-sourced lists of urls to help Common Crawl crawl under-resourced languages. See <https://github.com/commoncrawl/web-languages-code/> for the code

commoncrawl.org/

crawling language-detection dataset

Readme Activity Custom properties 43 stars 8 watching 44 forks

Report repository

Contributors 38

+ 24 contributors

Web Languages Project

Welcome! This is a crowd-sourced effort to improve crawling of low-resource languages. This dataset is public.

Common Crawl recognizes a lot of languages, and we can see that we don't have enough of languages like Hindi (500 million speakers!), smaller country languages like Hungarian, and regional languages like Catalan. We are interested in languages from all over the world. If you choose to help, you'll be helping create lists of websites related to languages that you read or speak.

Web Languages



The screenshot shows a web browser window for Dynabench, specifically the 'Text Language Identification' task page. The header includes the Dynabench logo, a search bar, and navigation links for 'About' and 'Communities'. The main visual is a colorful graphic featuring large letters (A, a, B, B, C) and binary code (01110010, 11100111). Text on the graphic includes 'Others', 'Text Language Identification', 'Common Crawl's Lang ID', '01', '1', '1120', 'ROUNDS', and 'EXAMPLES'. Below the graphic are buttons for 'Leaderboard', 'Overview' (which is selected), and 'Create Examples'. The 'Overview' section has a 'Description' tab selected, containing the title 'Common Crawl - MLCommons Language Identification task'. The 'Instructions' section welcomes users to the task and explains the goal: to produce a new LangID dataset based on Common Crawl data for as many languages as possible. It also describes the task as annotators selecting a language they are proficient in from a search bar. At the bottom, there is a text input field labeled 'TEXT LANGUAGE IDENTIFICATION' with the placeholder 'Label the text with the languages you think it is written in'.

Language Annotations





1st Workshop on
Multilingual Data Quality Signals

Palais des Congrès
Montréal, Canada
10 October 2025

[Call for Papers](#) [Shared Task](#)

Workshop



Global Safety

- Multilingual (especially LOTE) content is comparatively sparse [1]
- There are serious risks to deploying LLMs in other languages [2]
- GPT, PaLM2, LLaMA-2-Chat, and Vicuna give unsafe responses for LOTE queries [3]



*"Hey, buy this car, but the brakes and
seat belts only work in English-speaking
locales"*



Nobody, ever

Ethical & Trustworthy AI Data Pipelines

Ethical & Trustworthy AI Data Pipelines

- Crawler opt-outs are rising, so researchers and model builders must rethink how they gather data, and how that affects model safety and quality
- Sanitised or ambiguous opt-outs dilute user intent. AIPREF [1] aims to preserve real author preferences
- We need open, accountable systems that show *how* and *which* web content is used in model training
- We need a common vocabulary for expressing AI-related content preferences
- We need protocol-level attachments (e.g., metadata, HTTP headers, etc.) to signal those preferences
- We need a standard for resolving conflicting signals across sources
- We need all this in order to rebuild **trust**.

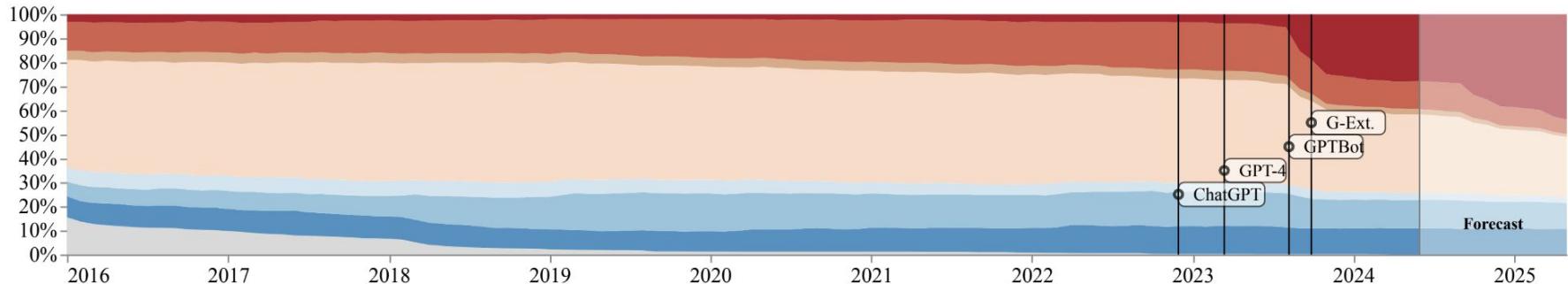


"If respected or enforced, these restrictions are rapidly biasing the diversity, freshness, and scaling laws for general-purpose AI systems. [...] The foreclosure of much of the open web will impact not only commercial AI, but also non-commercial AI and academic research."

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

Shayne Longpre et al., 2024

<https://arxiv.org/abs/2407.14933>



Robots.txt Restrictions

- Full restrictions
- Pattern-based restrictions
- Disallow private directories
- Other restrictions
- Crawl delay specified
- Sitemap provided
- No restrictions or sitemap
- No Robots.txt

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

Shayne Longpre et al., 2024

<https://arxiv.org/abs/2407.14933>



*"The only thing you absolutely
have to know is the location of
the library."*

Albert Einstein



Thank you!

You can access these slides with the QR code above.

Please feel free to join us on Discord or in our Google Group

<https://discord.gg/njaVFh7avF>

<https://groups.google.com/g/common-crawl>



I AM AWAKE

Thom Vaughan (Moderator)

Principal Technologist, Common Crawl

Lilith Bat-Leah

DMLR Working Group Co-chair, MLCommons

Dean Wampler

IBM Head of Technology, AI Alliance

- Ethics of large-scale data collection
- Preserving authenticity in user preference signals
- Governance and transparency in AI training data usage
- Collaborative standards across the AI ecosystem
- Building trust in public data pipelines

Jose Plehn-Dujowich

CEO, BrightQuery

Greg Lindahl

CTO, Common Crawl

Dave Buckley

Senior Policy Manager, OpenMined

Roberto di Cosmo

Director, Software Heritage