COMMON CRAWL

Preserving Humanity's Knowledge and Making it Accessible

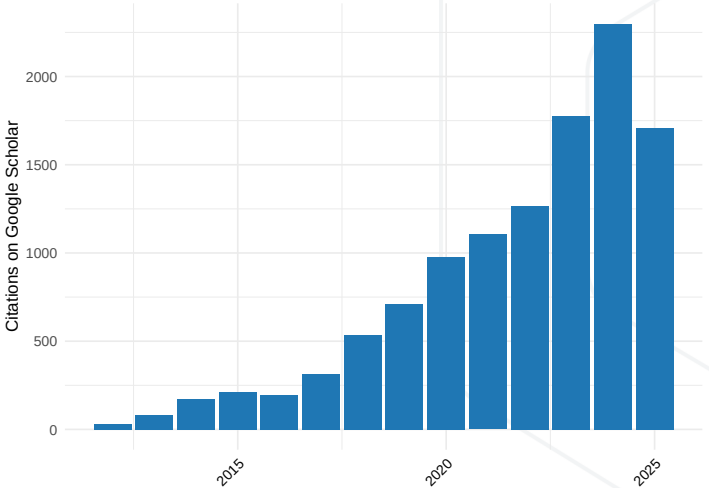
Addressing Challenges of Public Web Data

Stanford HAI, 22 October 2025

About Common Crawl

- We're a non-profit making web data accessible to programmers and data scientists
- Free archive of the public web since 2007, founded by Gil Elbaz
- Hosted as Open Dataset on Amazon Web Services [1, 2]
- Cited in over 10,000 research papers

Papers Citing Common Crawl (through 9/25)



Common Crawl Use Cases

- Search indexes
- Sociopolitical research
- Linguistic analyses
- Internet security research
- Economic research
- AI/ML training
- A lot more…

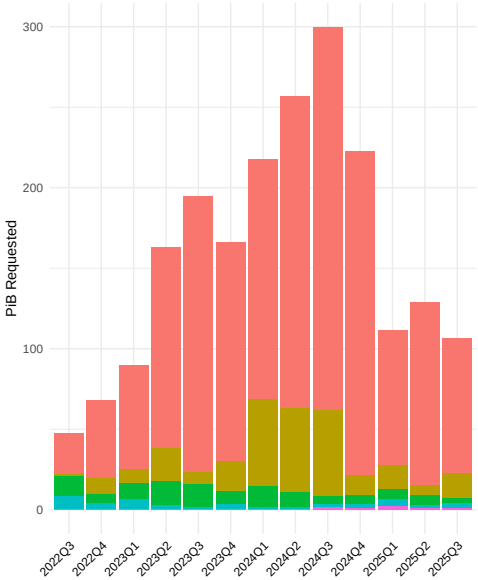
Our Mission

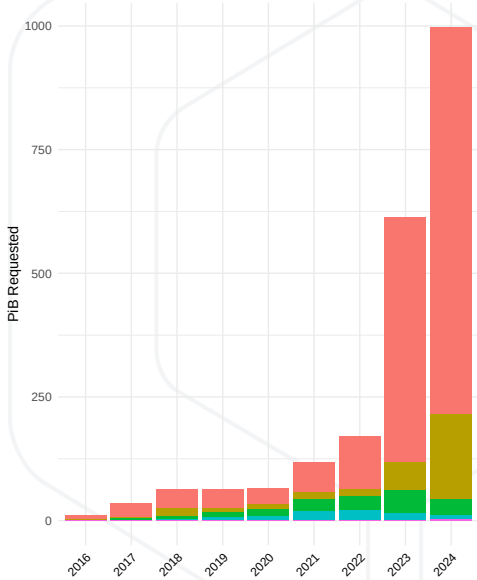
- Make high-quality public web crawl data (once limited to large search engines) available to everyone
- Support people, governments, and organizations make data-driven decisions and tackle global issues
- Build an open knowledge ecosystem that encourages sharing and innovation
- Support understanding and preservation of language, community, and culture

Data Overview

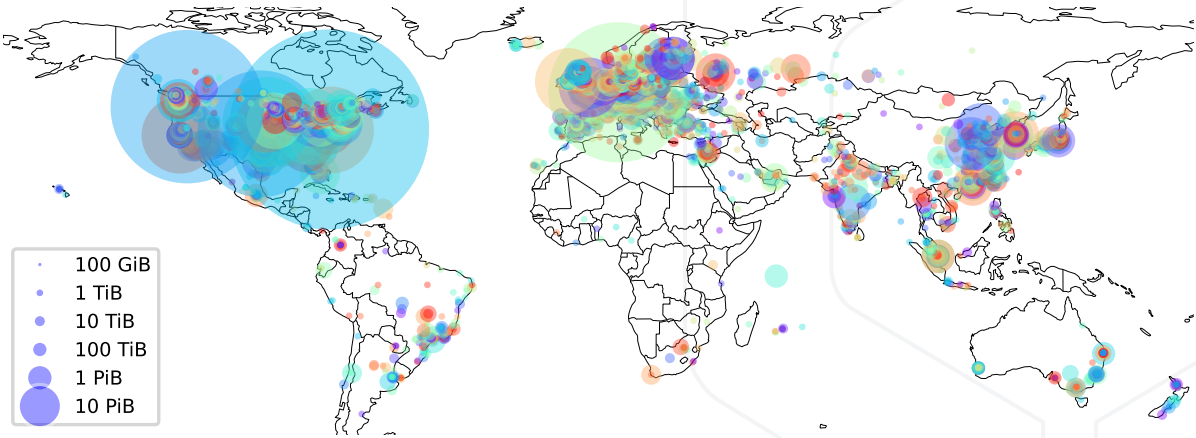
- Over 300 billion web pages spanning 17 years (2008 – 2025)
- Around 2.5 billion pages added each month
- More than 100 crawl archives released to date
- 10.1 PiB of data (Sept 2025)
- Additional data products: web graph, host index, etc.

Data Downloads





Downloads by Geographical Location



Data Collection – A Look Back In Time

Four phases of data collection using different

- Crawler implementations
- Approaches to find and sample (prioritize) seeds and URLs
- Page revisit policies

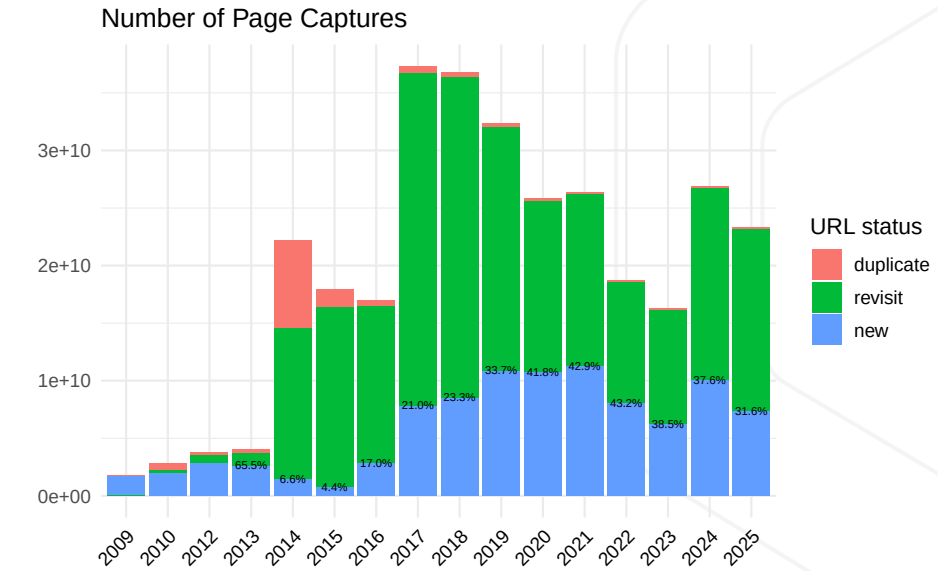
	crawler	seeds / link prioritization	revisit policy
2008–2009	Nutch	list based	
2012	in-house	page rank	
2013 – 2016	Nutch	seed donations, list based	
2017 – now	Nutch	harmonic centrality	30% new, 70% revisits based on page relevance

What Is Representative?

Aspects of representativity:

- Breadth: coverage of unique domains (web sites)
- Depth: per-site coverage
- Freshness (new content)
- Amount of (near-)duplicates (per crawl and over multiple crawls)
- Regional coverage (top-level domains, content languages)
- Content quality
- Applicable for a given data use case?

Size and Freshness



Crawler Politeness

- Crawl slowly
  - Further slow down (exponential backoff) if a site responds with errors
- Respect robots.txt rules (Robots Exclusion Protocol – [3]) and
- URI-level metatags (`<meta name=robots value=nofollow>`) [4, 5]

- CCBot identifies itself
  - User-agent string and contact information sent along with requests
  - Crawling from fixed list of IP addresses, publicly announced and verifiable via reverse DNS

Robots Exclusion Protocol (REP) - robots.txt

- A text file `robots.txt` is deployed in the root folder of a web site (eg. `https://example.org/robots.txt`)
- Readable for web crawlers (“robots”)
- Contains policies whether and how crawlers shall access the site’s content

- A technical solution to coordinate different interests between the owners of content and robots
- Standardized 2022 in RFC 9309 [3]
- A convention based on consensus not a legally binding regulation [6]
- Yet a widely acknowledged opt-out protocol [7]

robots.txt Example (2022)

```
User-agent: Googlebot
Disallow: /cgi-bin
Disallow: /cgi-perl
Disallow: /includes/
Disallow: /nav/
Disallow: /search/
Disallow: /search$
Disallow: /userinfo/
Allow: /scotland$
Allow: /scotland/
Allow: /wales$
Allow: /wales/
...

User-Agent: bingbot
...

User-agent: *
...
Disallow: /food/menus/*/shopping-list


User-agent: QuerySeekerSpider
Disallow: /

User-agent: magpie-crawler
Disallow: /
```

- BBC robots.txt, archived January 2022, simplified
- Rules exclude templates, dynamic content, user pages
- Improve quality of crawled content and search results!
- Dedicated rules for Googlebot and Bingbot
- Default rule set (wildcard user-agent) marginally different
- A few “bad” bots are disallowed

robots.txt Example (2025)

```
User-agent: *
Disallow: /search/
Disallow: /search$
Disallow: /search?
Disallow: /userinfo/
Disallow: /food/menus/*/shopping-list
...

User-agent: Amazonbot
Disallow: /

User-agent: magpie-crawler
Disallow: /

User-agent: CCBot
Disallow: /

User-Agent: omgili
Disallow: /

User-Agent: omgilibot
Disallow: /

User-agent: ClaudeBot
Disallow: /

User-agent: Claude-Web
Disallow: /

User-agent: anthropic-ai
Disallow: /

User-agent: cohere-ai
Disallow: /

User-agent: Bytespider
Disallow: /

User-agent: PetalBot
Disallow: /

User-agent: Scrapy
Disallow: /

User-agent: Applebot-Extended
Disallow: /

User-agent: GPTBot
Disallow: /

User-agent: ChatGPT-User
Disallow: /

User-agent: Google-Extended
Disallow: /

User-Agent: PerplexityBot
Disallow: /

User-agent: Perplexity-User
Disallow: /

User-agent: Google-CloudVertexBot
Disallow: /

User-agent: meta-externalagent
Disallow: /

User-agent: OAI-SearchBot
Disallow: /

User-agent: YandexAdditional
Disallow: /

User-agent: YandexAdditionalBot
Disallow: /

User-agent: TurnitinBot
Disallow: /
```

Many more user-agents entirely disallowed.

What happened?

ChatGPT and other tools based on Large Language Models.

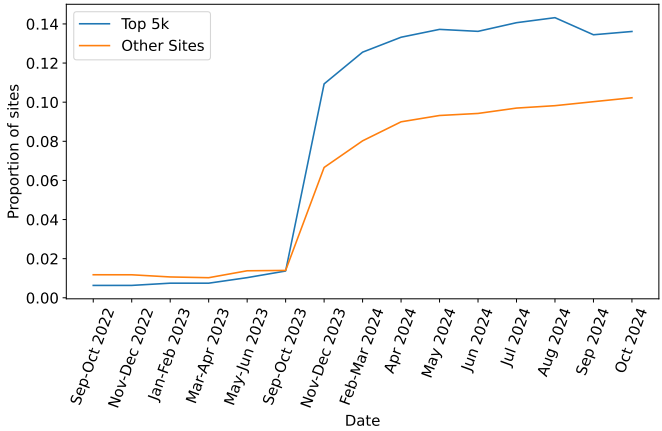
The AI Panic

- ChatGPT's web interface launched in November, 2022
- The initial backlash frequently mentioned CCBot
- The backlash grew over time thanks to abusive, anonymous that appear to be AI-related
- Website providers (e.g. Cloudflare, Wordpress) saw this as a marketing opportunity
- This caused many changes in both robots.txt and "bot defenses"
  - Cloudflare manages robots.txt for 3.8 million websites
  - Wordpress is blocking CCBot with a low rate limit
  - We have the raw data for bot blocking but have not analyzed it yet

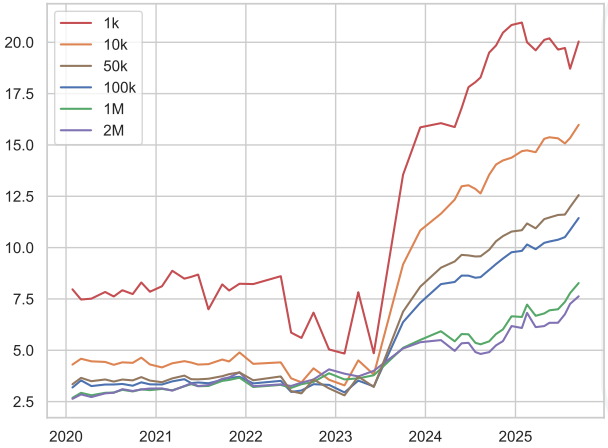
robots.txt – Impact on Web Crawling (i)

Liu et al. 2024, Somesite I used to crawl [8]

- Proportion of sites that fully disallow any AI-related user agent, broken down by site rank,
- Data: top-5k vs. top-100k Tranco domains [9], Common Crawl robots.txt captures [10]



robots.txt – Impact on Web Crawling (ii)

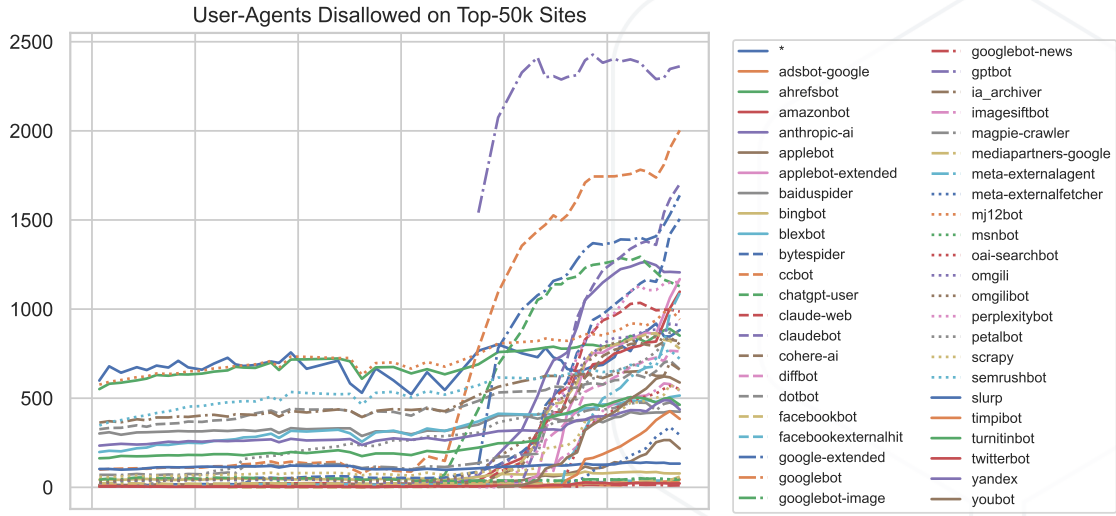


- Common Crawl's crawler (CCBot)
- Share of sites blocking CCBot has grown since 2023
- Higher ranking sites: $> 10\%$
- Long-tail: disallowed share up from 3% to 6–10%
- Does the trend continue?
- CCF robots.txt dataset, Tranco domains, see [11, 12]

robots.txt – Impact on Web Crawling (iii)

What about other User-agents?

robots.txt – Impact on Web Crawling (iii)



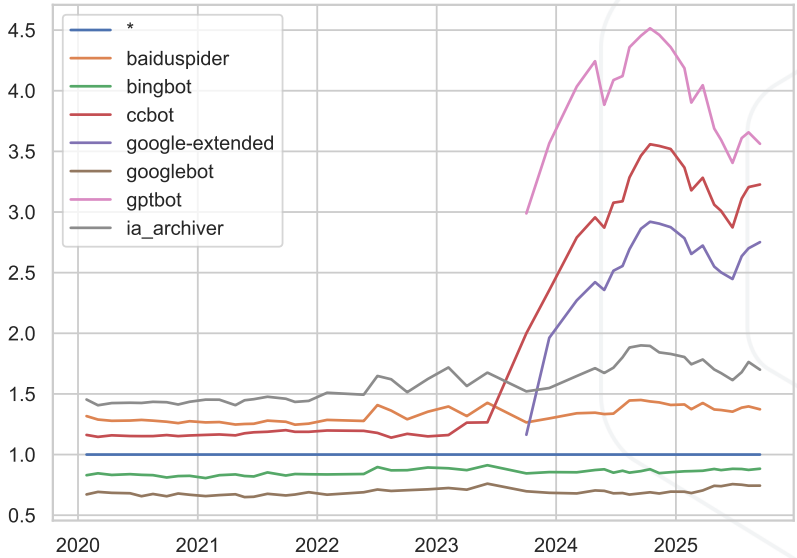
robots.txt – Impact on Web Crawling (iv)

- Starting 2023 more and more user-agents are addressed
- To understand how this affects individual user-agents
  - We take the wildcard user-agent as baseline
  - And use the ratio

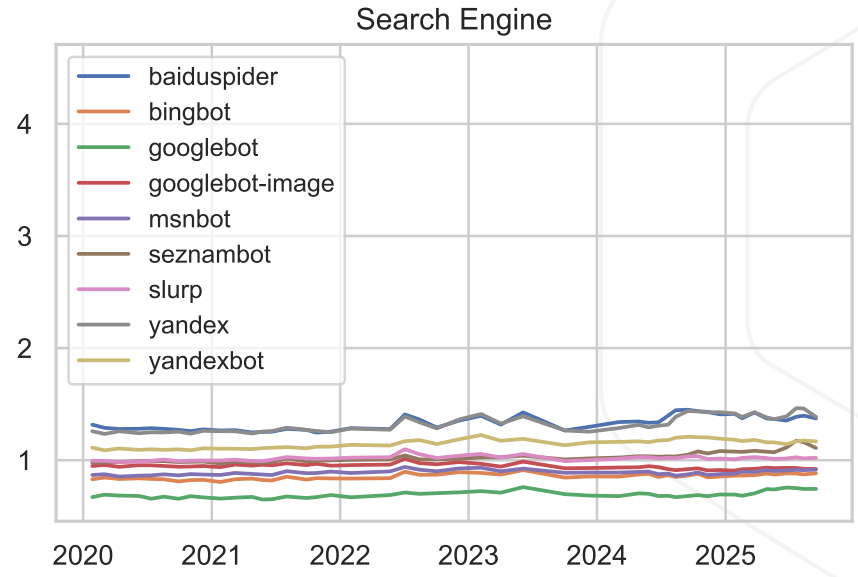
$$\frac{\text{disallowed for user-agent } x}{\text{disallowed for wildcard user-agent}}$$

  - On the top-50k stratum

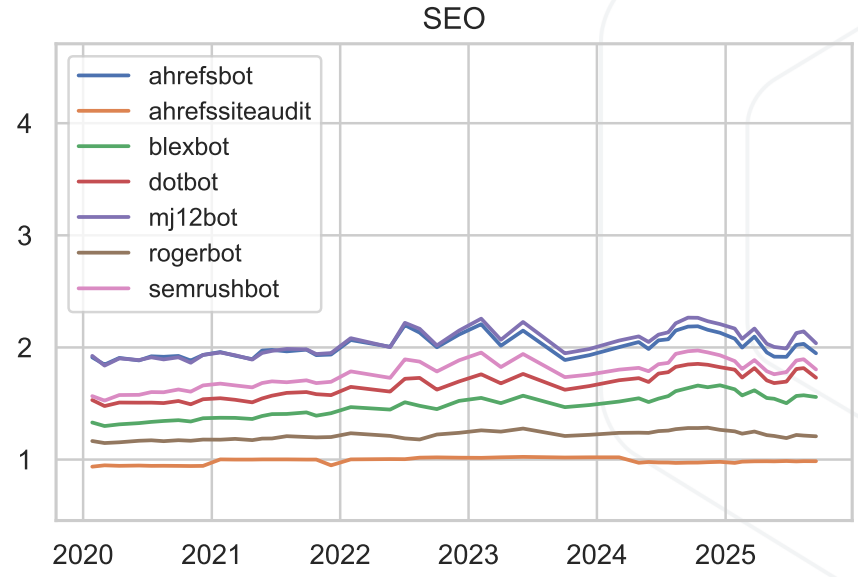
robots.txt – Impact on Web Crawling (v)



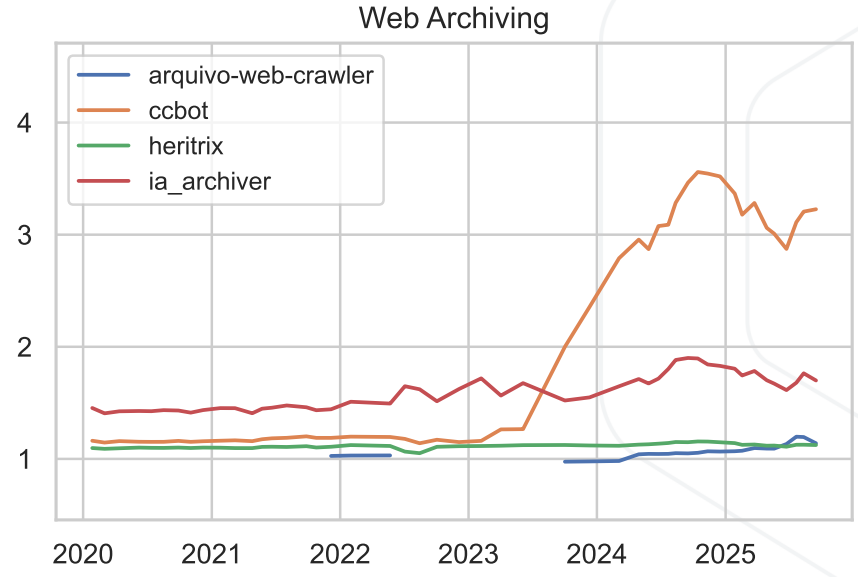
robots.txt – Impact on Web Crawling (vi)



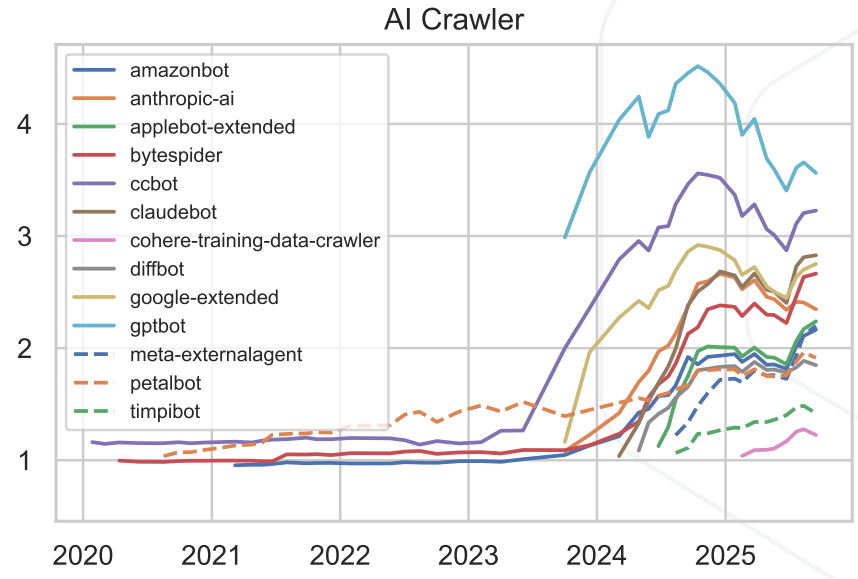
robots.txt – Impact on Web Crawling (vii)



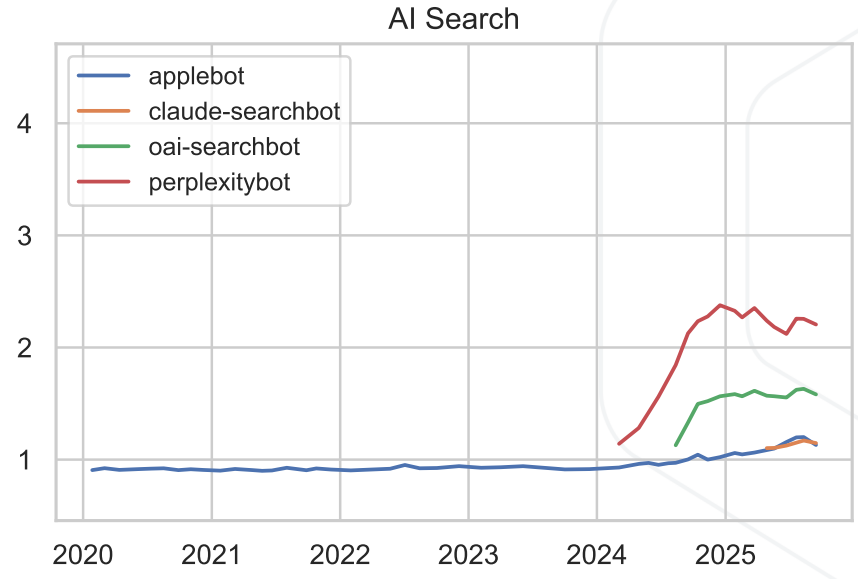
robots.txt – Impact on Web Crawling (viii)



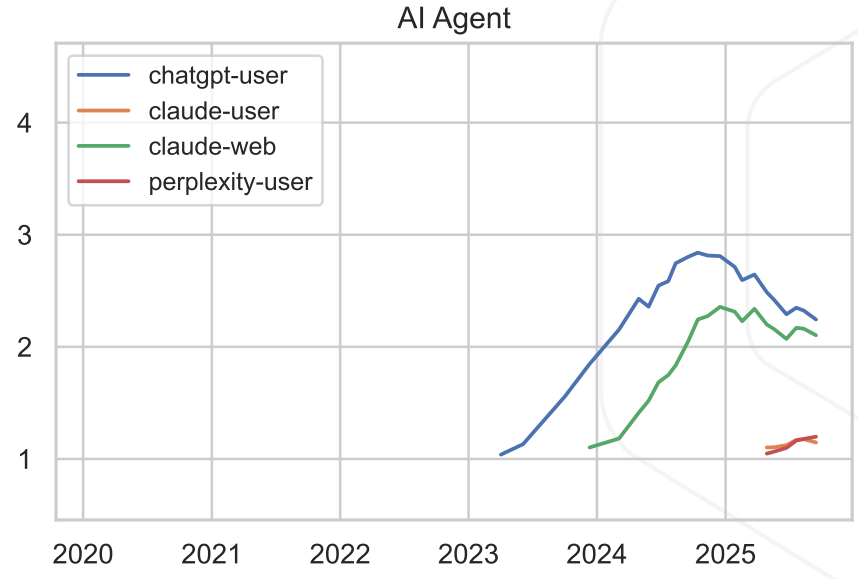
robots.txt – Impact on Web Crawling (ix)



robots.txt – Impact on Web Crawling (x)



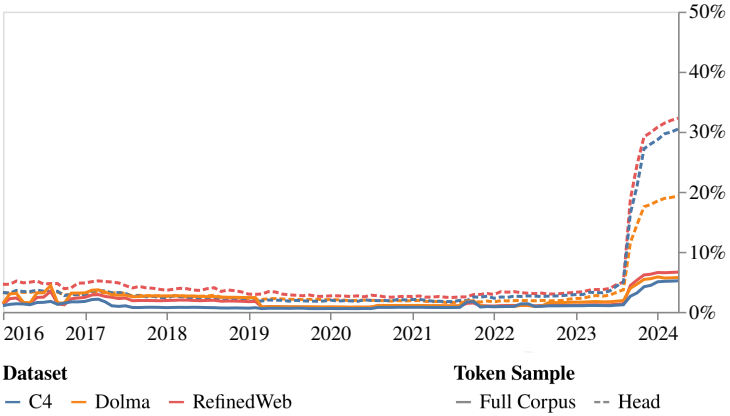
robots.txt – Impact on Web Crawling (xi)



robots.txt – Impact on Training Data (i)

Percentage of restricted tokens by robots.txt (Longpre et al. 2024, Consent in crisis [17])

- Head: top-2k web sites by token count in C4, Dolma and RefinedWeb
- Full Corpus: 10k randomly sampled sites

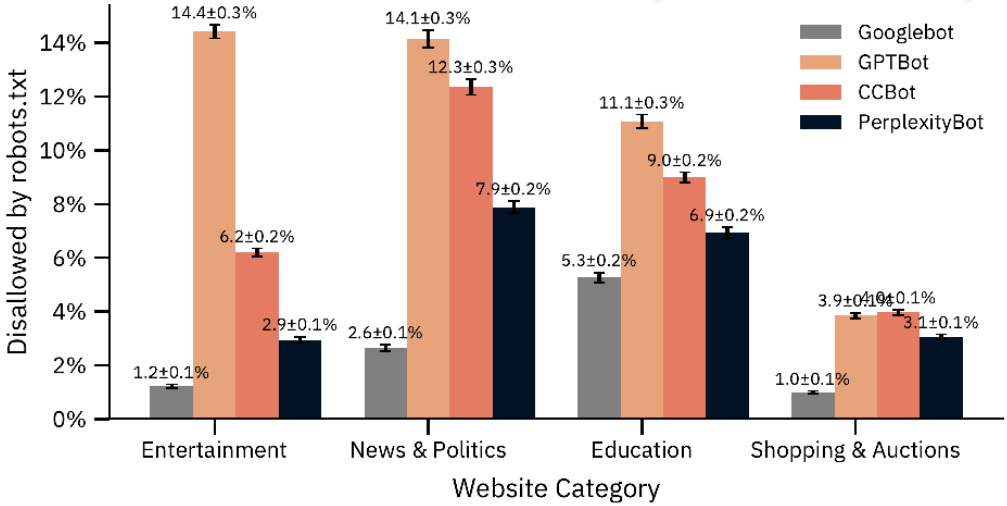


robots.txt – Impact on Training Data (ii)

“Pre-training on fully open data does not significantly impact general knowledge understanding. Even if all news publishers opt out, the effect remains minimal.”

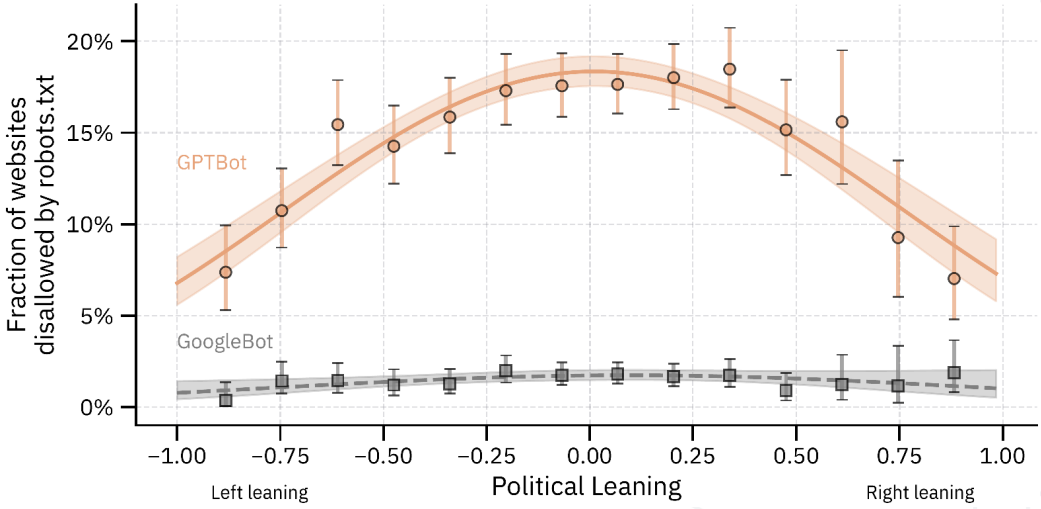
Can Performant LLMs be Ethical?
Quantifying the impact of web crawling opt-outs, 2025 [18]

robots.txt – Impact on Training Data (iii)



Web crawler restrictions, AI training datasets & political biases, 2025 [19]

robots.txt – Impact on Training Data (iv)



Web crawler restrictions, AI training datasets & political biases, 2025 [19]

Internet Standards Discussions

- Several Working Groups and drafts have emerged at the IETF with aims of addressing the governance and technical implications of automated access to web resources:
  - AI Preferences (aipref WG)[20]: developing mechanisms for sites to express preferences about AI use and automated data access.
  - Web Authorization Protocols for Bots (web-bot-auth WG) [21]: exploring authentication and authorisation frameworks for responsible automated web access.
  - Paid Crawling Requirements (draft-nottingham-paid-crawl-reqs) [22]: an Internet-Draft discussing standardised requirements and signals for paid access to crawled data.

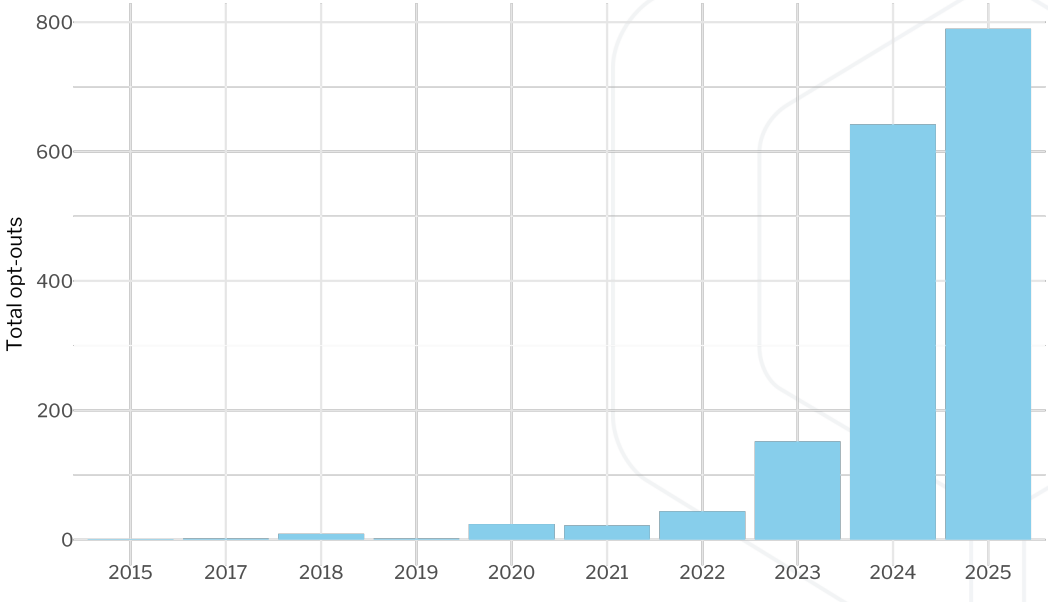
Policy Discussions

- In the policy domain, civil society groups like Open Future, and other non-profit organisations have proposed vocabularies [23] for expressing opt-out preferences related to AI training and text/data mining (TDM).
- Creative Commons has launched CC Signals [24], a framework intended to act as a new social contract for the AI era: it allows data and content stewards to declare machine-reuse preferences (for example, training, inference, attribution) while embedding norms of reciprocity, sustainability, and shared benefit in the development of AI systems.
- CommonsDB [25] provides a registry of open and public datasets, facilitating the discovery and reuse of web-accessible data under open licenses.

Policy Trends

- These initiatives show the growing collective effort to balance open access to public web data and respect for publisher preferences, copyright, and responsible AI development.
- Common Crawl has published the full opt-out list [26] for every legal request we have received to date. These opt-out requests are becoming more frequent year by year.
  - We hope that users of our datasets will use this registry to filter what they have downloaded in the past.

Opt-out Requests per Year



Consent in Crisis

> *"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." [17]*

– Shayne Longpre et al., Consent in Crisis, 2024

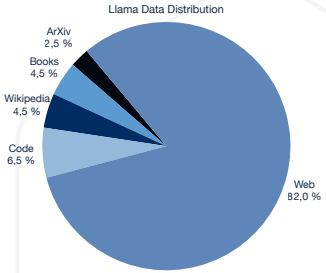
Important Knowledge

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

– Albert Einstein

Prevalence of Web Data in LLMs Today(ish)

- In practice datasets are not balanced
- Web data is always the cheapest and easiest to get
- Web data is diverse, but definitely not balanced or representative of all the language range.
- Web data always contains unwanted content (fiction, bias, propaganda).
- Programming language code (source code) is becoming ubiquitous in the data mix
- Public domain books and encyclopedic data is also common, but availability varies greatly between languages.



LLaMA Data

Source	Proportion
Web	82%
Code	6.5%
Wikipedia	4.5%
Books	4.5%
ArXiv	2.5%

Multilinguality is Hard – Even with Web Data!

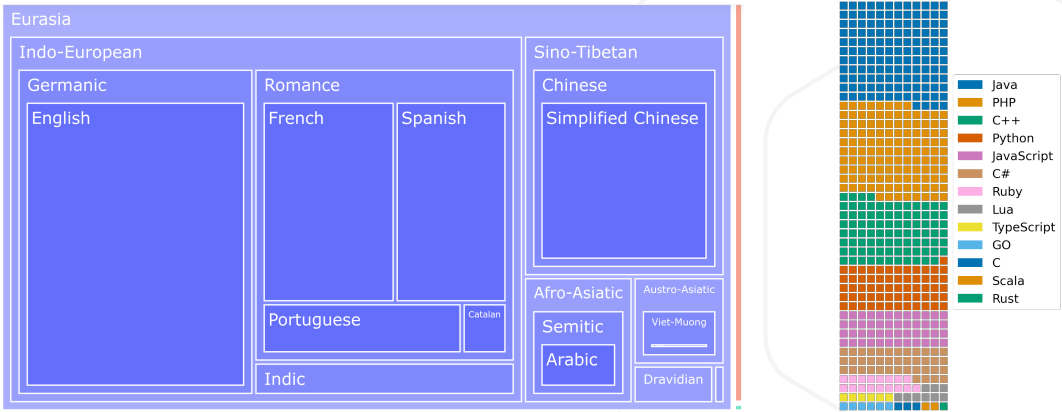


Figure 1: Overview of ROOTS [27] Left: A treemap of natural language representation in number of bytes by language family. Right: A waffle plot of the distribution of programming languages by number of files. One square corresponds approximately to 30,000 files.

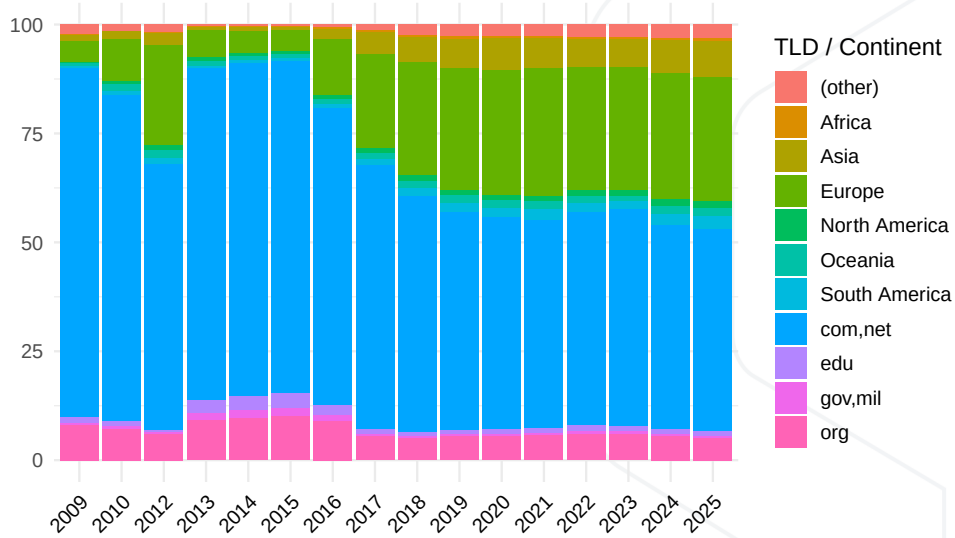
Top-Level Domains and Geographical Coverage

year %	com	org	ru	net	de	uk	jp	edu	fr	it	pl	nl	br	au	cz
(all)	52.38	6.42	4.02	3.79	3.26	2.06	1.65	1.45	1.36	1.35	1.23	1.11	0.97	0.82	0.81
2009	71.20	8.10	0.05	9.19	0.05	4.04	0.05	1.28	0.02	0.04	0.19	0.01	0.34	0.60	0.01
2010	68.93	7.14	0.46	6.03	1.51	3.12	0.50	1.31	0.44	0.57	0.48	0.44	0.79	1.02	0.17
2012	55.86	6.02	1.71	5.43	4.75	3.45	1.14	0.61	1.30	1.30	1.79	1.43	0.76	0.98	0.76
2013	73.08	9.40	0.06	3.27	1.12	2.00	0.16	2.91	0.38	0.34	0.02	0.25	0.22	0.63	0.10
2014	73.25	9.75	0.11	3.20	0.81	1.74	0.13	3.36	0.30	0.26	0.16	0.16	0.28	0.51	0.06
2015	73.31	10.25	0.11	3.16	0.76	1.67	0.14	3.34	0.28	0.25	0.14	0.17	0.26	0.51	0.05
2016	64.55	8.96	2.56	3.66	1.97	1.79	0.81	2.48	0.63	0.65	0.49	0.51	0.43	0.60	0.52
2017	56.54	5.64	5.30	4.22	2.83	2.03	2.09	1.02	1.15	1.08	0.99	0.73	0.81	0.70	0.85
2018	50.06	5.28	6.06	4.41	3.49	2.22	2.16	0.75	1.42	1.28	1.37	1.02	0.99	0.78	0.93
2019	46.21	5.68	5.09	3.98	3.88	2.38	1.99	0.89	1.67	1.63	1.51	1.38	1.19	0.92	0.99
2020	44.97	5.68	4.85	3.71	4.03	2.30	1.89	1.08	1.75	1.70	1.52	1.46	1.31	0.98	1.02
2021	44.33	5.71	4.79	3.59	4.15	2.38	1.89	1.28	1.79	1.77	1.59	1.52	1.40	1.05	1.05
2022	45.38	6.16	4.14	3.56	4.21	1.75	1.78	1.49	1.85	1.88	1.66	1.64	1.05	0.80	1.03
2023	46.53	6.06	4.27	3.49	4.30	0.77	1.66	1.31	1.91	2.06	1.64	1.76	0.49	0.41	1.08
2024	43.65	5.51	4.40	3.32	4.03	2.09	2.08	1.14	1.75	1.92	1.75	1.54	1.41	1.05	1.03
2025	43.32	5.13	4.24	3.30	4.15	2.19	2.23	1.06	1.70	1.76	1.66	1.49	1.54	1.16	1.00

Top-Level Domains and Geographical Coverage

% year	(other)	com net	org	edu	gov mil	North America	South America	Oceania	Africa	Asia	Europe
2009	2.37	80.38	8.10	1.28	0.38	0.11	0.49	0.64	0.03	1.48	4.75
2010	1.49	74.96	7.14	1.31	0.51	0.77	0.98	1.41	0.13	1.78	9.53
2012	1.81	61.29	6.02	0.61	0.27	1.31	1.19	1.74	0.23	2.77	22.76
2013	0.40	76.35	9.40	2.91	1.41	1.02	0.42	1.08	0.19	0.85	5.97
2014	0.42	76.45	9.75	3.36	1.69	0.96	0.53	0.87	0.18	0.83	4.96
2015	0.37	76.47	10.25	3.34	1.68	0.84	0.51	0.85	0.17	0.81	4.69
2016	0.74	68.20	8.96	2.48	1.27	0.95	0.81	1.09	0.24	2.48	12.77
2017	1.38	60.76	5.64	1.02	0.42	1.15	1.38	1.33	0.30	5.13	21.49
2018	2.47	54.47	5.28	0.75	0.28	1.40	1.72	1.51	0.40	5.87	25.83
2019	2.80	50.19	5.68	0.89	0.31	1.36	2.04	1.65	0.48	6.62	27.99
2020	2.76	48.68	5.68	1.08	0.35	1.36	2.18	1.66	0.48	7.16	28.61
2021	2.81	47.91	5.71	1.28	0.41	1.45	2.34	1.72	0.50	6.68	29.17
2022	2.88	48.93	6.16	1.49	0.48	1.51	2.08	1.40	0.48	6.37	28.21
2023	2.99	50.02	6.06	1.31	0.47	1.57	1.66	0.94	0.43	6.36	28.21
2024	3.05	46.96	5.51	1.14	0.44	1.59	2.63	1.66	0.60	7.68	28.76
2025	3.20	46.62	5.13	1.06	0.44	1.65	2.79	1.80	0.63	8.31	28.36

Top-Level Domains and Geographical Coverage



Language Coverage

	2018	2019	2020	2021	2022	2023	2024	2025
<other>	6.47	7.31	7.88	7.59	7.51	7.84	7.90	7.91
<unknown>	3.32	2.62	2.35	2.53	2.80	2.84	3.00	2.91
ara	0.76	0.59	0.56	0.60	0.63	0.64	0.66	0.67
ces	1.03	1.04	1.05	1.06	1.03	1.09	1.04	1.02
deu	5.15	5.47	5.57	5.63	5.60	5.79	5.36	5.57
eng	**43.96**	**43.84**	**43.20**	**44.98**	**46.64**	**45.70**	**44.40**	**44.07**
fas	0.66	0.59	0.57	0.63	0.63	0.67	0.71	0.72
fra	4.53	4.56	4.53	4.46	4.50	4.64	4.31	4.31
ind	0.75	0.74	0.76	0.83	0.75	0.82	1.01	1.09
ita	2.06	2.31	2.38	2.42	2.48	2.69	2.53	2.31
jpn	5.47	4.81	4.78	4.66	4.68	4.83	5.03	5.09
kor	0.59	0.69	0.76	0.66	0.65	0.67	0.71	0.77
nld	1.50	1.74	1.79	1.80	1.93	2.09	1.85	1.81
pol	1.68	1.70	1.68	1.62	1.62	1.70	1.80	1.75
por	1.99	2.07	2.16	2.15	1.78	1.27	2.14	2.26
rus	**9.27**	**7.41**	**7.11**	**7.07**	**5.90**	**5.84**	**6.02**	**5.97**
spa	4.18	4.16	4.25	4.33	4.36	4.58	4.56	4.46
tur	0.90	0.87	0.91	1.00	0.86	0.84	1.18	1.15
vie	0.75	0.74	0.80	0.92	0.92	1.04	1.01	1.03
zho	4.98	6.76	6.91	5.07	4.73	4.41	4.75	5.12

Percentage identified by CLD2 Source: [28]

Language Coverage

- 45% of the content is English
- About 5% – Chinese, French, German, Japanese, Russian, Spanish
- Other languages: 2% or below
- Is this balanced?

Improving Language Coverage and Balance

Initiatives to improve the language balance

- The Web Languages project
  (`https://github.com/commoncrawl/web-languages`)
  - Crowd-sourced effort to improve crawling of low-resource languages
  - Language, Community, Culture
  - Show the crawler high-quality links to “steer” into the right direction
  - Hopefully, this will influence the harmonic centrality ranks over time

- Better text extracts and language identification

The Web-Languages Project

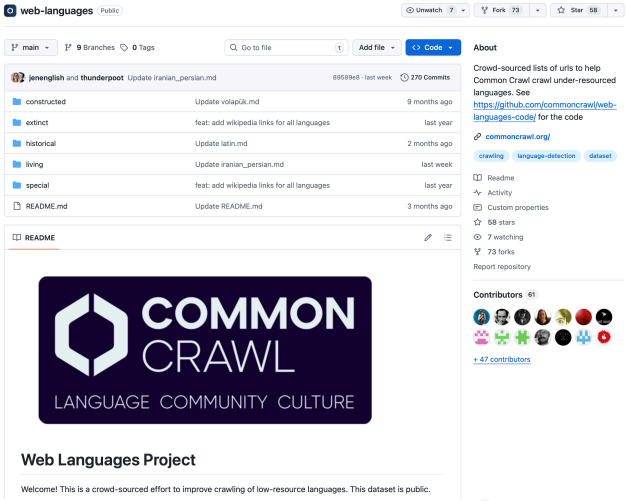


Figure 2: https://github.com/commoncrawl/web-languages

The LangID Project

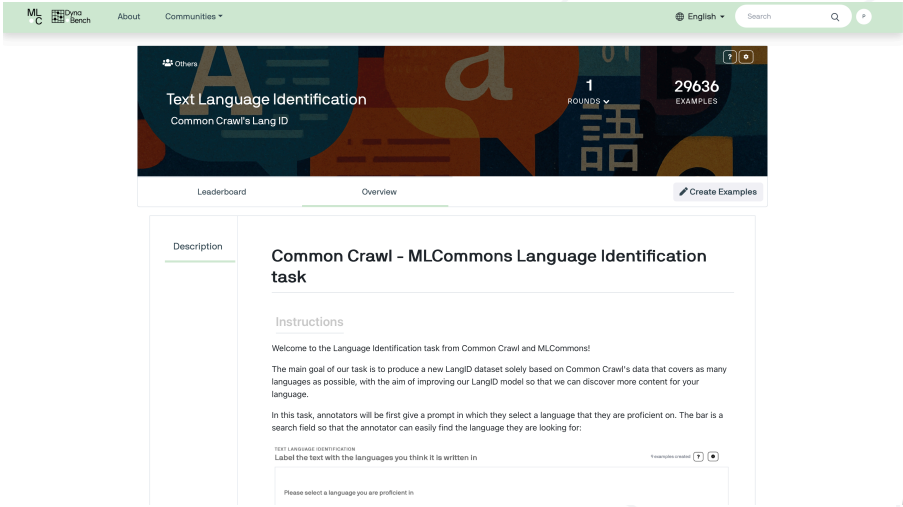


Figure 3: https://dynabench.org/tasks/text-language-identification

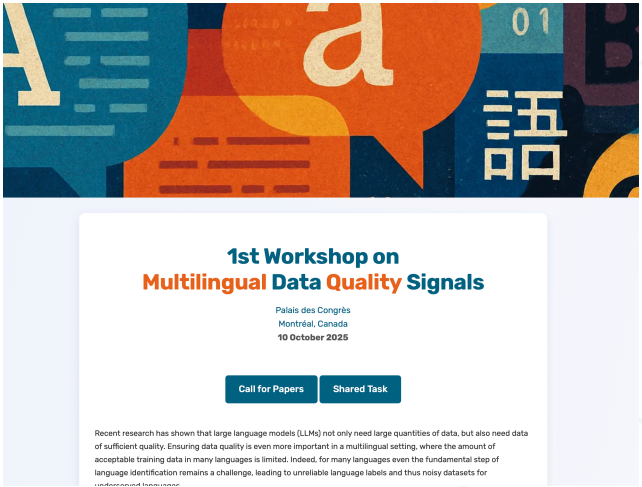


Figure 4: https://wmdqs.org

Common Crawl – A Brief Introduction

Crawler Politeness and robots.txt

Legal & Policy

Web Data and Language Coverage

Summary

Summary

- We are currently classifying 160 languages, we want to support many more
  - Human-curated seed lists to cover every language, community, and culture
  - Collect human-labeled samples to build a more robust and diverse language labeling algorithm
- Policy: IETF working groups, UK/EU/US Policy papers, expand collaborations with civic groups, academic institutions, and NSF

Questions?

https://github.com/commoncrawl/presentations

References i

[1] Amazon Web Services. ***Open Data Sponsorship Program.*** https://aws.amazon.com/opendata/open-data-sponsorship-program/.

[2] Amazon Web Services. ***Registry of Open Data on AWS.*** https://registry.opendata.aws/.

[3] Martijn Koster et al. ***Robots Exclusion Protocol.*** Tech. rep. 9309. Sept. 2022. 12 pp. https://www.rfc-editor.org/info/rfc9309.

[4] Martijn Koster. ***A Standard for Robot Exclusion.*** 1996. https://www.robotstxt.org/meta.html.

[5] Gary Illyes. ***Robots Exclusion Protocol Extension for URI Level Control.*** Internet-Draft draft-illyes-repext-02. Work in Progress. Internet Engineering Task Force, Oct. 2024. 6 pp. https://datatracker.ietf.org/doc/draft-illyes-repext/02/.

References ii

[6] MHM Schellekens. **“Are internet robots adequately regulated?”** In: *Computer Law & Security Review* 29.6 (2013), pp. 666–675. https://www.sciencedirect.com/science/article/pii/S0267364913001659.

[7] Peter Henderson et al. ***Foundation Models and Fair Use.*** 2023. arXiv: 2303.15715 [cs.CY]. https://arxiv.org/abs/2303.15715.

[8] Enze Liu et al. ***Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers.*** 2024. arXiv: 2411.15091 [cs.HC]. https://arxiv.org/abs/2411.15091.

[9] Victor Le Pochat et al. **“Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation”.** In: *Proceedings of the 26th Annual Network and Distributed System Security Symposium*. NDSS 2019. Feb. 2019. https://tranco-list.eu/.

References iii

[10] ***Data Sets Containing Robots.txt Files and Non-200 Responses – Common Crawl.***
https://commoncrawl.org/2016/09/robotstxt-and-404-redirect-data-sets/.

[11] Sebastian Nagel and Thom Vaughan. ***Robots.txt and crawler politeness in the age of generative AI.*** English. Poster. Apr. 2025.
https://digital.library.unt.edu/ark:/67531/metadc2472442/ (visited on 08/12/2025).

[12] ***Robots.txt Experiments and Metrics.*** 2025.
https://github.com/commoncrawl/robotstxt-experiments (visited on 10/19/2025).

[13] Y. Sun et al. **"Determining bias to search engines from robots.txt".** In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2007*. 2007, pp. 149–155.

References iv

[14] Santanu Kolay et al. **“A larger scale study of robots.txt”.** In: *Proceedings of the 17th international conference on World Wide Web*. 2008, pp. 1171–1172. https://dl.acm.org/doi/abs/10.1145/1367497.1367711.

[15] Greg Elmer. **“Exclusionary rules? The politics of protocols”.** In: *Routledge handbook of internet politics* (2008), pp. 376–383.

[16] Greg Elmer. **“The spam book: On viruses, porn and other anomalies from the dark side of digital culture”.** In: ed. by Jussi Parikka and Tony D. Sampson. Creskill, New Jersey: Hampton Press, 2009. Chap. Robots.txt: The politics of search engine exclusion, pp. 217–227.

[17] Shayne Longpre et al. ***Consent in Crisis: The Rapid Decline of the AI Data Commons.*** 2024. arXiv: 2407.14933 [cs.CL]. https://arxiv.org/abs/2407.14933.

References v

[18] Dongyang Fan et al. ***Can Performant LLMs Be Ethical? Quantifying the Impact of Web Crawling Opt-Outs.*** 2025. arXiv: 2504.06219 [cs.CL]. https://arxiv.org/abs/2504.06219.

[19] Paul Bouchaud and Pedro Ramaciotti. **"Web Crawler Restrictions, AI Training Datasets & Political Biases".** working paper or preprint. Oct. 2025. https://hal.science/hal-05302425/.

[20] Internet Engineering Task Force. ***IETF AI Preferences WG.*** 2024. https://datatracker.ietf.org/wg/aipref/about/.

[21] Internet Engineering Task Force. ***IETF web-bot-auth WG.*** 2025. https://datatracker.ietf.org/wg/webbotauth/about/.

[22] Mark Nottingham. ***Requirements for Paid Web Crawling.*** 2025. https://datatracker.ietf.org/doc/draft-nottingham-paid-crawl-reqs/.

[23] Paul Keller. ***A vocabulary for opting out of AI training and other forms of TDM.*** 2025. https://openfuture.eu/publication/a-vocabulary-for-opting-out-of-ai-training-and-other-forms-of-tdm/.

[24] Creative Commons. ***CC Signals: A New Social Contract for the Age of AI.*** 2025. https://creativecommons.org/ai-and-the-commons/cc-signals/.

[25] CommonsDB. ***A registry for Public Domain and openly licensed works.*** 2025. https://www.commonsdb.org/.

[26] Common Crawl. ***Common Crawl Foundation Opt-Out Registry.*** 2025. https://commoncrawl.org/blog/common-crawl-foundation-opt-out-registry.

References vii

[27] Hugo Laurençon et al. **“The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset”.** In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 31809–31826. https://proceedings.neurips.cc/paper_files/paper/2022/file/ce9e92e3de2372a4b93353eb7f3dc0bd-Paper-Datasets_and_Benchmarks.pdf.

[28] ***Statistics of Common Crawl Monthly Archives*.** https://commoncrawl.github.io/cc-crawl-statistics/.