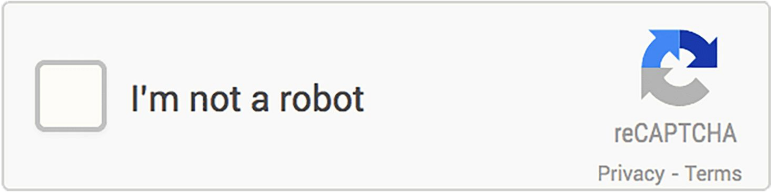
Navigating Ethical Use of Web Scale Data, Ensuring Proper Language Representation and Best Practices

Open Source Ethics & Innovation @ The Turing Way
BMA House, London, November 2024

COMMON CRAWL



□ I'm not a robot

reCAPTCHA

Privacy - Terms

What is Common Crawl?

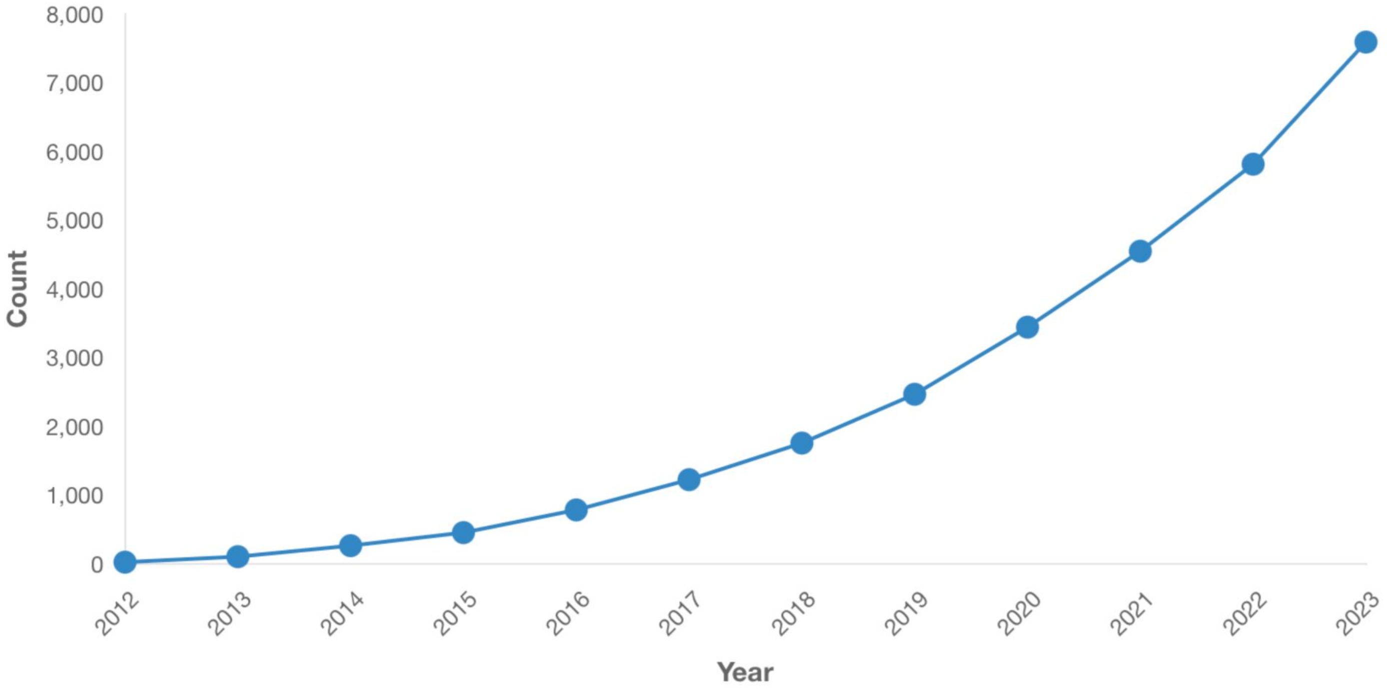
COMMON
CRAWL

What is Common Crawl?

- Over 275 billion pages spanning 17 years
- Free and open corpus since 2007
- Cited in over 10,000 research papers
- 3–5 billion new pages added each month

Cumulative Citations

Plot of Common Crawl citations in Google Scholar until January 2024



https://commoncrawl.org/research-papers
https://huggingface.co/datasets/commoncrawl/citations

Origins

COMMON
CRAWL

Origins

- Started in 2007 by Gil Elbaz
- Intended to be used for research and for search indexes
- In the dawn of large language models, the data has been used as training data
- Over 100 crawl archives released to date
- Usually > 2 billion pages, and > 320 TiB of uncompressed content per crawl
- Over 7 PiB (8 PB) so far!

>7 PiB

COMMON
CRAWL

7 PiB = ~ 7,340,032 GiB

(ish)

COMMON CRAWL

that’s just text.

Data Access

COMMON
CRAWL

Data Access

- We store archives in a public S3 bucket named `commoncrawl`
- `Access via s3:// from inside AWS or https:// from outside`
- `WARC` [1] files, `WET` files, and `WAT` files, with supporting indexes (CDX and Parquet)
- We preserve REP data (`robots.txt`, `HTTP` headers, `<meta>` tags)
- We generate Web Graphs [2] after each crawl from which we get ranks
  - Which you can also access for free…

[1] https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.1/
[2] https://commoncrawl.org/web-graphs

How (Our) Crawling Works

COMMON
CRAWL

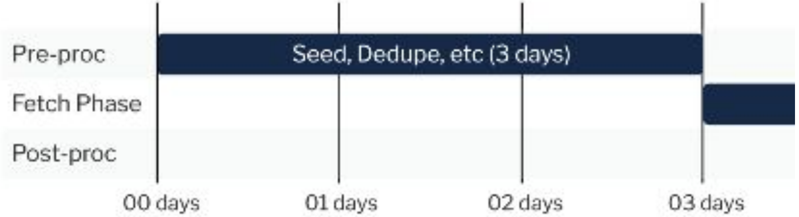
How (Our) Crawling Works

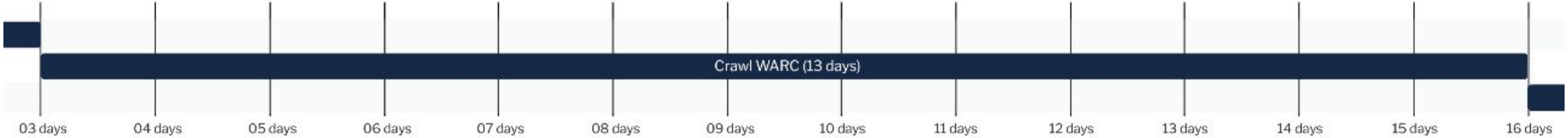
- Our Apache Nutch-based* crawler identifies itself as `CCBot/2.0`
- We run our infrastructure from `us-east-1` in AWS EC2
- First we perform a “seed crawl”, where we get URLs
- Ranking (Harmonic Centrality) is derived from our Web Graphs
- From the list of URLs we request `robots.txt`, and if we are allowed, we request sitemaps, follow links, and take a stratified sample

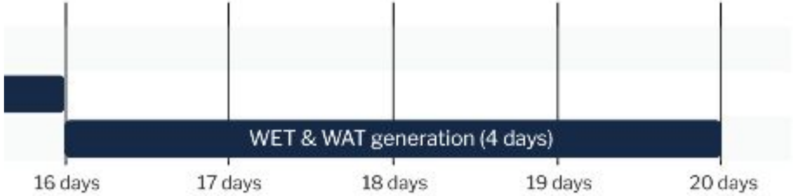
* We use our own fork of Apache Nutch™ which you can check out at https://github.com/commoncrawl/nutch

COMMON CRAWL

Typical Crawl Workflow Timeline



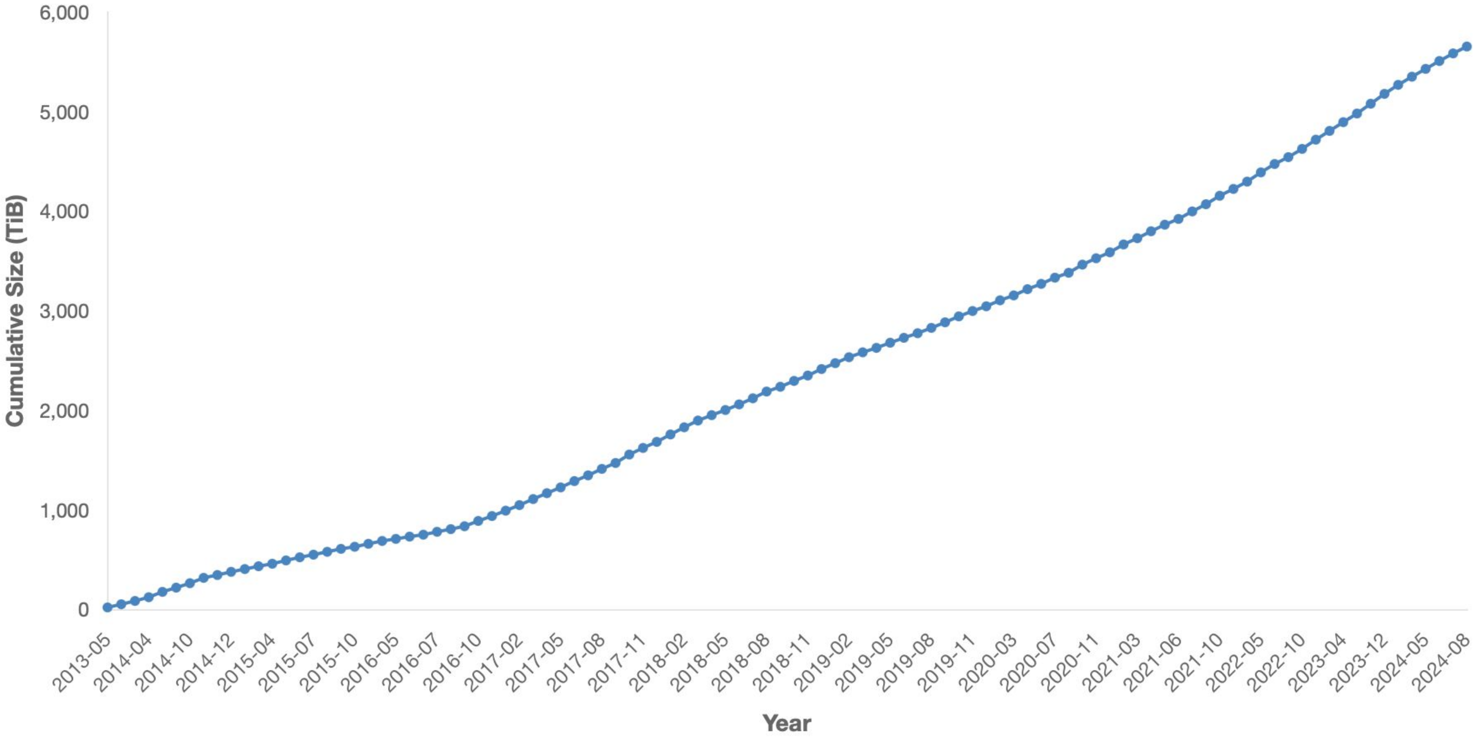




COMMON CRAWL

Cumulative Growth: WARC

Plot showing growth of WARC data in Common Crawl's public data archives



https://commoncrawl.github.io/cc-crawl-statistics/

amazon
ATHENA

I am the Greek goddess of
database queries

```
CREATE EXTERNAL TABLE IF NOT EXISTS commoncrawl_index       -- let’s create a new table with the following columns:
(
  url_surtkey                 STRING,                       -- Sort-friendly URI Reordering Transform
  url                         STRING,                       -- the URL (duh) including protocol (http or https)
  url_host_name               STRING,                       -- the hostname, including subdomain(s)
  url_host_tld                STRING,                       -- the top-level domain such as `.org`
  url_host_registered_domain  STRING,                       -- the registered domain name
  url_host_private_domain     STRING,                       -- private domain such as `example.com`
  url_host_public_suffix      STRING,                       -- public suffix of the domain such as `.co.uk` or `.edu`
  url_protocol                STRING,                       -- the transfer protocol used, (http or https)
  url_port                    INT,                          -- the port used, typically 80 for http or 443 for https
  url_path                    STRING,                       -- the stuff after the hostname, like `/cool/stuff.html`
  url_query                   STRING,                       -- the URL query such as `?foo=bar`
  fetch_time                  TIMESTAMP,                    -- when the page was fetched, `ISO 8601`
  fetch_status                SMALLINT,                     -- the HTTP status returned, like 200 or 404 etc
  content_digest              STRING,                       -- the SHA-1 hash of the content
  content_mime_type           STRING,                       -- the media type, such as `text/html` or `application/pdf`
  content_mime_detected       STRING,                       -- the _detected_ mime type (in case it differs)
  content_charset             STRING,                       -- like `UTF-8` or `ISO-8859-1` and so on
  content_languages           STRING,                       -- ISO 639-3 of the detected lang(s) (up to three)
  warc_filename               STRING,                       -- the S3 path e.g. `/over/here/foo.warc.gz`
  warc_record_offset          INT,                          -- the offset within the WARC file (bytes)
  warc_record_length          INT                           -- the content length (bytes)
)
PARTITIONED BY (crawl STRING, subset STRING)                -- group by crawl ID and subset
STORED AS PARQUET                                           -- columnar format
LOCATION 's3://commoncrawl/cc-index/table/cc-main/warc/';   -- just WARC stuff in `s3://commoncrawl/`

MSCK REPAIR TABLE commoncrawl_index;                        -- then add the partitions to the metastore
```

Quality Signals

COMMON
CRAWL

Quality Signals 🥴

- Harmonic Centrality ranking [1]
- Diversity of languages [2]
- Measure of deduplication
- Recency
- Measure of spam reduction
- Completeness of metadata
- Token density* [3]

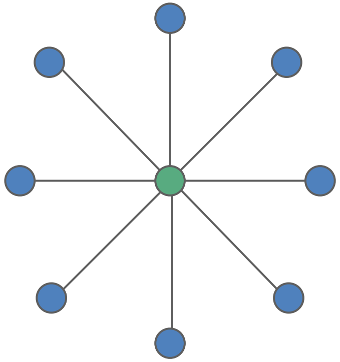
[1] https://www.youtube.com/watch?v=cnGJtGP4gL4
[2] https://commoncrawl.github.io/cc-crawl-statistics/plots/languages
[3] https://cybernetist.com/2024/10/21/you-should-probably-pay-attention-to-tokenizers/

* token density is a downstream metric, not something we compute

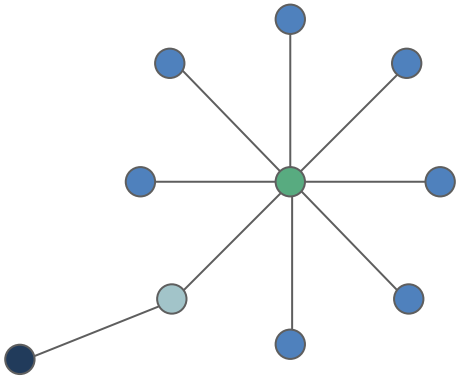
COMMON
CRAWL

Harmonic Centrality

COMMON
CRAWL



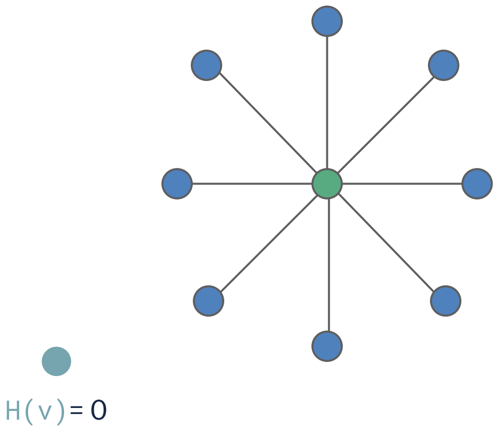
Centrality measures are generally scores given to vertices based on how well connected they are, or how “important” they are within a network.



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

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

COMMON
CRAWL



The **harmonic centrality** of a vertex is the mean inverse distance to all other vertices. The inverse distance to an unreachable vertex is considered to be zero.

Languages

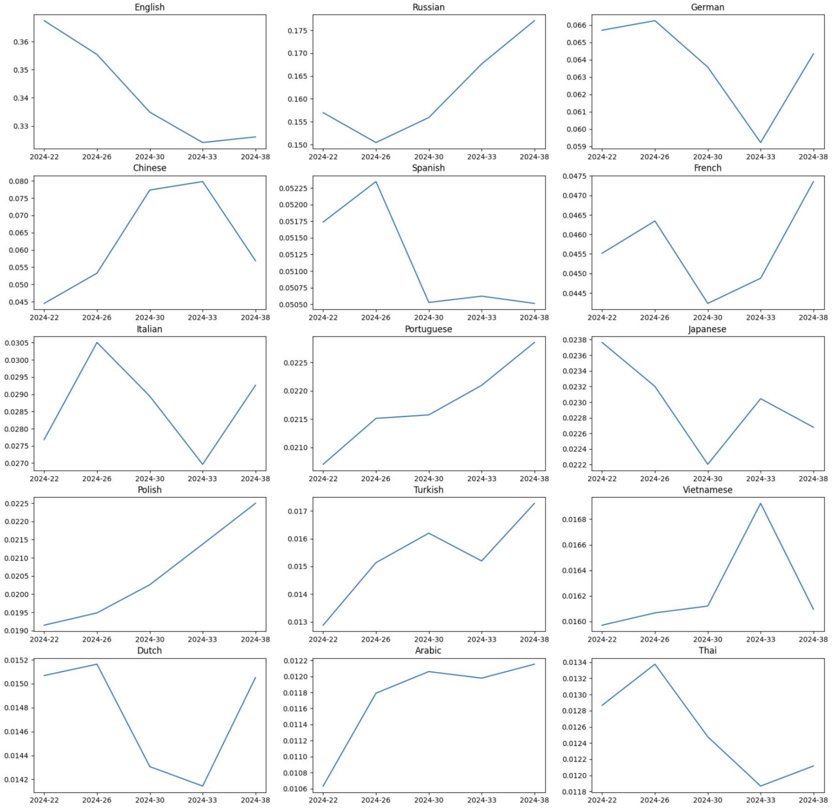
Top ten languages from the three most recent crawls

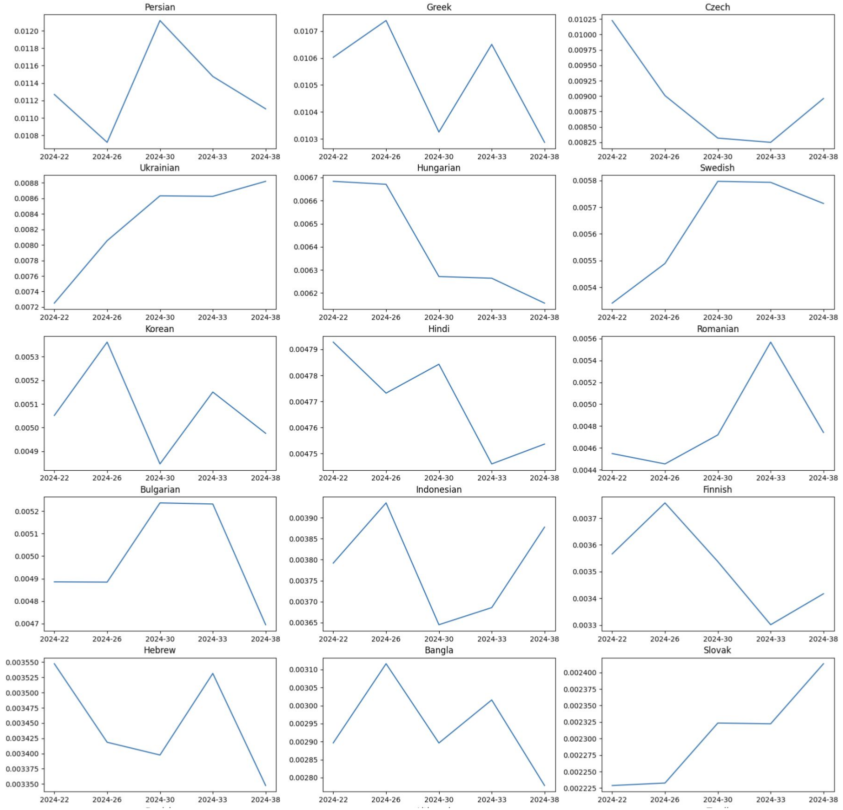
Language	2024-38	2024-42	2024-46
eng	44.1210	43.4241	42.9924
rus	6.1556	6.0444	6.2093
deu	5.4471	5.3038	5.3362
jpn	5.1119	5.0419	5.1463
zho	4.6266	4.8129	5.1252
spa	4.4769	4.5387	4.6154
fra	4.4292	4.3960	4.3541
<unknown>	2.6706	3.2780	3.0185
ita	2.5224	2.5282	2.4952
por	2.2141	2.3146	2.3292

COMMON
CRAWL

https://commoncrawl.github.io/cc-crawl-statistics/plots/languages

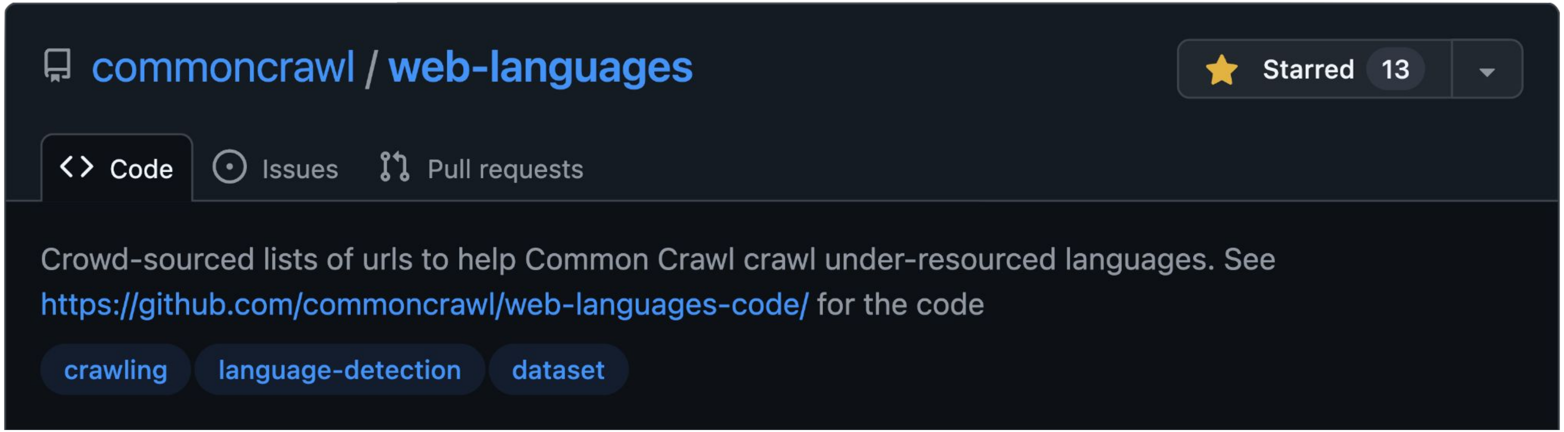
Language Distribution of Recent Crawls





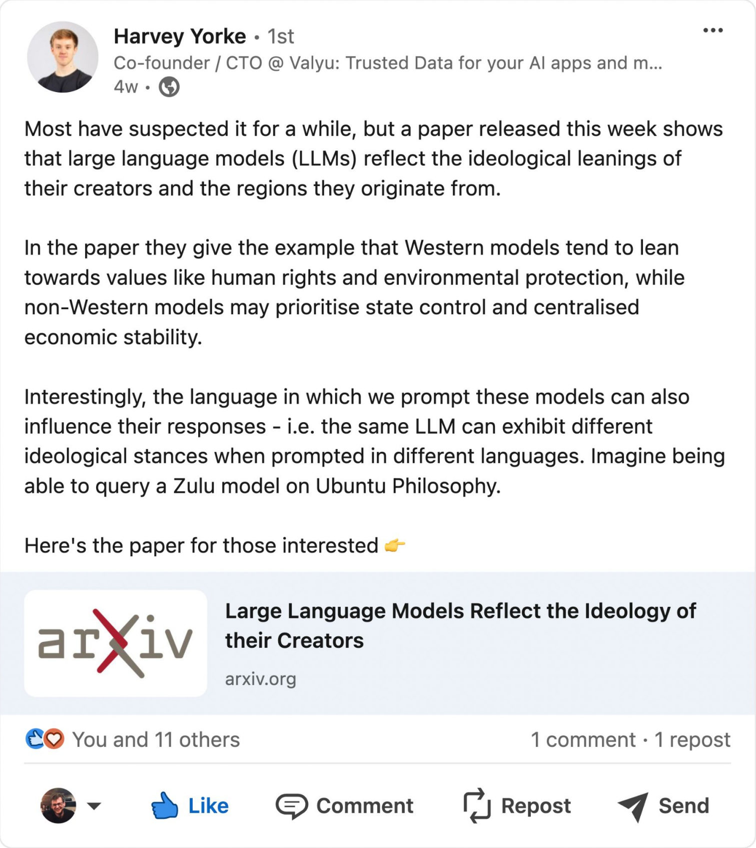
Web Languages Project

COMMON
CRAWL



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

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



Harvey Yorke • 1st
Co-founder / CTO @ Valyu: Trusted Data for your AI apps and m...
4w •

Most have suspected it for a while, but a paper released this week shows that large language models (LLMs) reflect the ideological leanings of their creators and the regions they originate from.

In the paper they give the example that Western models tend to lean towards values like human rights and environmental protection, while non-Western models may prioritise state control and centralised economic stability.

Interestingly, the language in which we prompt these models can also influence their responses - i.e. the same LLM can exhibit different ideological stances when prompted in different languages. Imagine being able to query a Zulu model on Ubuntu Philosophy.

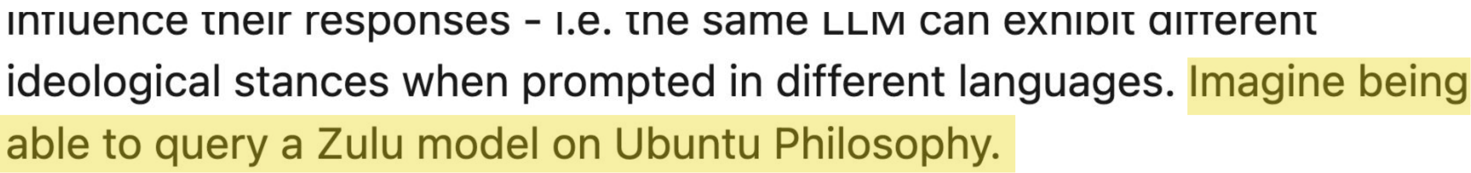
Here's the paper for those interested 👉

Large Language Models Reflect the Ideology of their Creators
arxiv.org

You and 11 others   1 comment • 1 repost

COMMON CRAWL

https://arxiv.org/abs/2410.18417



influence their responses - i.e. the same LLM can exhibit different ideological stances when prompted in different languages. Imagine being able to query a Zulu model on Ubuntu Philosophy.

https://arxiv.org/abs/2410.18417

Ethical Considerations

COMMON
CRAWL

COMMON
CRAWL

“With great power comes great responsibility”

Uncle Ben

Ethical Considerations

- Environmental impact / resource consumption
- Privacy (GDPR, Data Protection laws, &c)
- Types of data (“Open”, “Public”, “Obtainable”, and “Private”)
- Robots Exclusion Protocol compliance
- More opt-out vs opt-in (preference signals)
- IAB / IETF recommendations
- Long-term impacts of exclusion

IAB/IETF Recommendations

COMMON
CRAWL

IAB/IETF Recommendations

The Common Crawl Foundation is in the process of authoring an Internet Draft proposing the AI-CONTROL vocabulary.

The draft outlines a structured approach for content publishers to express detailed preferences regarding the use of their content in AI model training, encompassing permissions, intended purposes, data retention, and associated usage conditions.

- **Objective**
  - A proposed vocabulary to help content publishers specify preferences for the use of their content in AI model training
- **Content Preferences**
  - Vocabulary that allows publishers to define permissions, purpose, data retention, granularity, and other preferences
- **Metadata Application**
  - Preferences can be attached to content as metadata, to allow interoperability across different Internet protocols
- **Scope**
  - Applies only to expressing preferences, enforcement is outside its remit, depending on mutual agreements or legal means
- **Elements**
  - Key elements include permissions, purpose (e.g., text generation), temporal restrictions, content type, derivative use, metadata persistence, and geographic restrictions

	Term	Values	Description	Example
Permission	`allow_training`	Bool	Basic indicator of whether content can be used for AI training	`allow_training: false`
	`restricted_training`	String: `public, non-commercial, internal`	Specifies permitted training contexts e.g. public models or internal usage	`restricted_training: non-commercial`
Purpose	`purpose`	String: `text-gen, classification, summarisation, embedding`, etc	Defines acceptable applications for training e.g. fine-tuning, classification, summarisation, etc	`purpose: classification, summarisation`
Time-Based Restrictions	`effective_date`	Date string, **ISO 8601**	Start date of when permissions take effect	`effective_date: 2024-10-30T15:52:55.440238`
	`expiration_date`	Date string, **ISO 8601**	Date after which permissions no longer apply	`expiration_date: 2024-10-30T15:52:55.440238`
Granularity	`scope`	`global, content-specific, conditional`	Defines whether the preferences apply universally, to specific content, or under certain conditions	`scope: content-specific`
Content Type	`content_type`	`text, image, video, audio`	Specifies the type(s) of content the preference applies to	`content_type: text, image`
Derivative Content	`allow_derivatives`	Bool	Indicates whether derivative works (summaries, paraphrasing) are allowed based on content	`allow_derivatives: true`
	`derivative_type`	String: `summary, paraphrase, translation`	Lists permissible types is `allow_derivatives` is `true`	`derivative_type: summary, paraphrase`
Data Retention	`retention_period`	Duration string, **ISO 8601**	Specifies how long content may be retained after use (e.g. after training)	`P3Y6M4DT12H30M5S` representing three years, six months, four days, twelve hours, thirty minutes, and five seconds
Metadata Persistence Required	`metadata_must_persist`	Bool	Whether preferences must persist with derived data, boolean for either `required` or `optional`	`metadata_must_persist: true`
Notification	`notification`	Bool	Whether publishers must be notified when content is used	`notification: true`
Precedence	`precedence`	`high, medium, low`	Sets priority when preferences conflict with other layered preferences	`precedence: high`
Geographic Restrictions	`geo_limitations`	Location codes, **ISO 3166**	Specifies geographic regions where training permissions apply	`geo_limitations: EU, US`

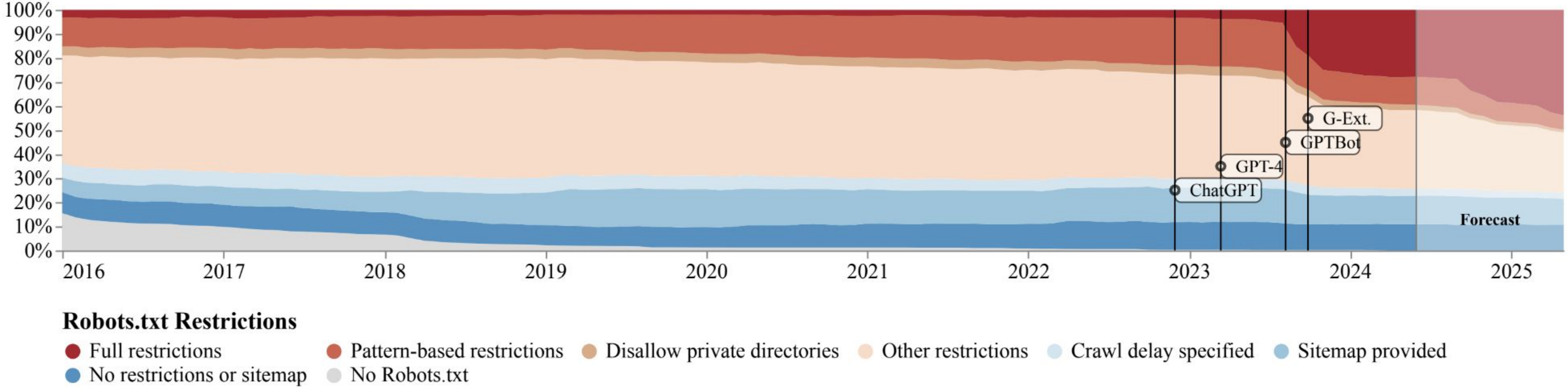
COMMON CRAWL

> *“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

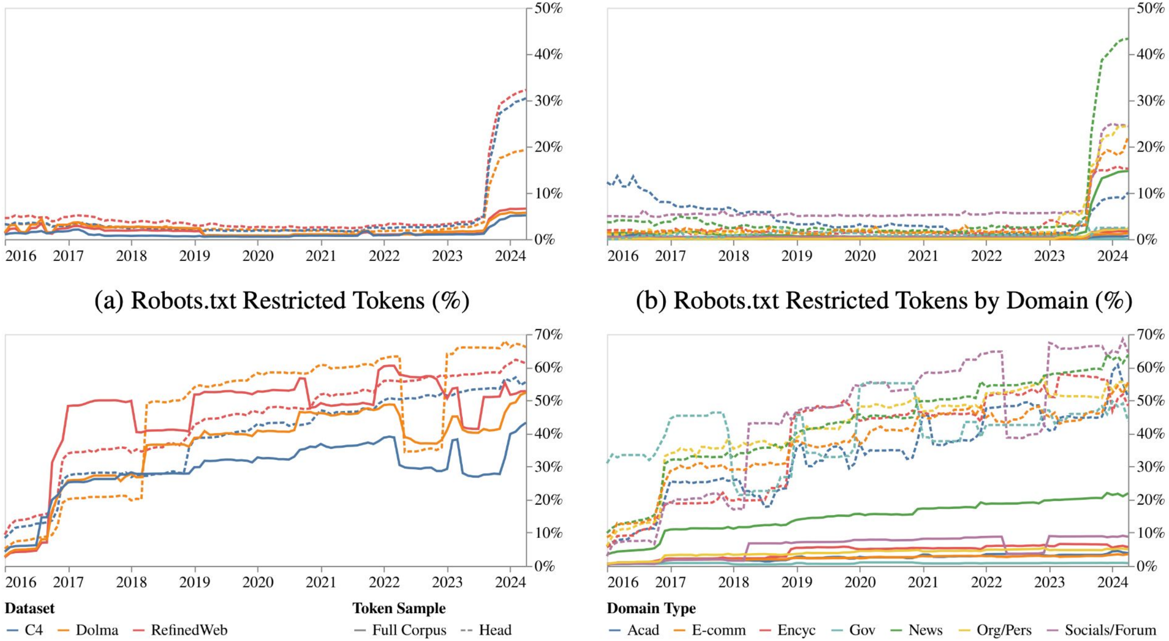
https://arxiv.org/abs/2407.14933



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

Shayne Longpre et al., 2024

https://arxiv.org/abs/2407.14933



(a) Robots.txt Restricted Tokens (%)

(b) Robots.txt Restricted Tokens by Domain (%)

(c) Terms of Service Restricted Tokens (%)

(d) Terms of Service Restricted Tokens by Domain (%)

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

Shayne Longpre et al., 2024

https://arxiv.org/abs/2407.14933

COMMON
CRAWL

> *“The only thing that 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