

# Open Data for Open Models

Infrastructure, Standards, and Transparency

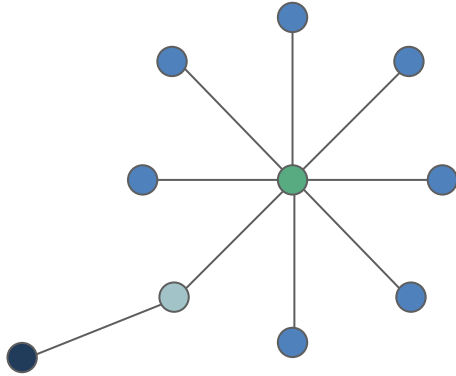
AI4IA Conference 2025

Thom Vaughan, Common Crawl Foundation

# What is Common Crawl?

# What is Common Crawl?

- 501(c)(3) nonprofit, started in 2007 by Gil Elbaz
- Free and open corpus containing > 300 billion pages over > 18 years
- Over 10 PB in WARC [1] and yet more in WAT, WET, and BVGraph [2]
- Hosted on AWS S3 as an Open Data set, enabled by the AWS Open Data Sponsorship Program
- Cited in > 10,000 research papers [3] [4]
- 3–4 billion new pages added each month
- Broad crawl steered by link-based Harmonic Centrality ranks
- Web Graphs showing the structure and connectivity of the web with host and domain ranks

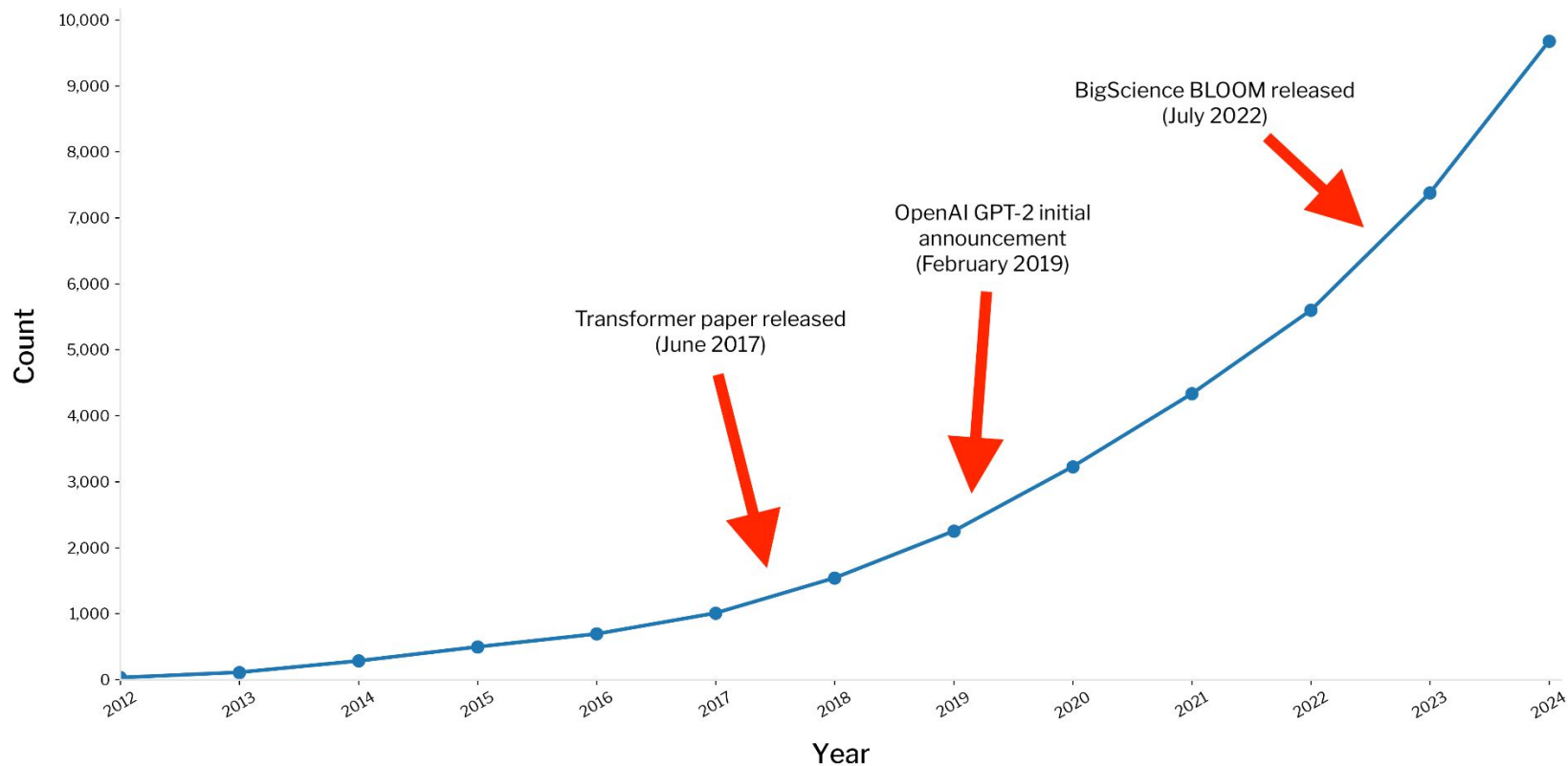


$$H(v) = \sum_{u \neq v} \frac{1}{d(v, u)}$$

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

**Why is it important?**

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



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

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

<https://github.com/commoncrawl/cc-citations>

For comparison, the Hubble Space Telescope is cited in 1,000 papers/year

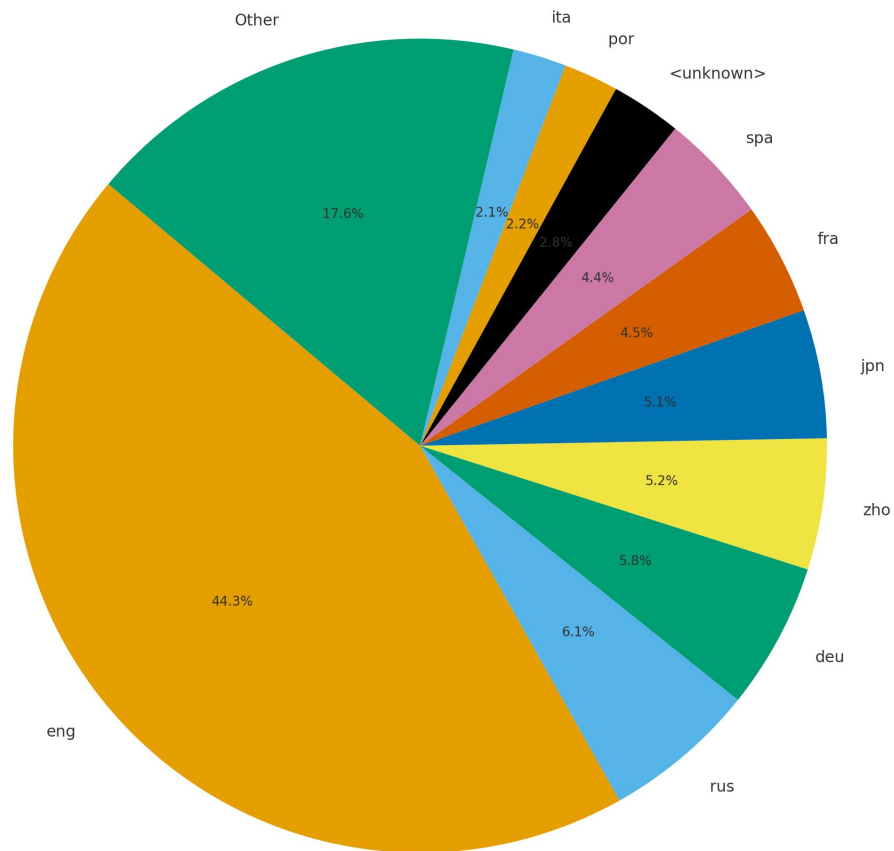
# Common Crawl's Use in AI and Elsewhere

- Common Crawl datasets are instrumental in advancing machine learning and artificial intelligence , particularly in natural language processing and web content analysis. [1]
- Common Crawl is the largest freely available web-crawled dataset and a cornerstone of pre-training data for LLMs, with over 80% of GPT-3's tokens derived from it. [2]
- OSCAR [3], mC4 [4], The Pile [5], and many other foundational datasets are all either *directly derived* from Common Crawl or include large portions of data that originate from it.

# Challenges with languages



Distribution of Web Pages by Primary Language (Common Crawl CC-MAIN-2025-33)



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

# 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

The screenshot shows a web browser window with the URL `dynabench.org/tasks/text-language-identification`. The page features a green header with the Dynabench logo and navigation links. The main content area has a colorful banner with the title "Text Language Identification" and "Common Crawl's Lang ID". It also displays "1 ROUNDS" and "1120 EXAMPLES". Below the banner are tabs for "Leaderboard" and "Overview", and a "Create Examples" button. The "Overview" tab is active, showing a "Description" section with the title "Common Crawl - MLCommons Language Identification task". The "Instructions" section follows, explaining the goal of the task and the role of annotators. At the bottom, there is a small example of the task: "TEXT LANGUAGE IDENTIFICATION" and "Label the text with the languages you think it is written in".

Dynabench

dynabench.org/tasks/text-language-identification

MLC Dyna Bench About Communities

Search

Others

Text Language Identification

Common Crawl's Lang ID

1 ROUNDS

1120 EXAMPLES

Leaderboard Overview Create Examples

Description

### Common Crawl - MLCommons Language Identification task

#### Instructions

Welcome to the Language Identification task from Common Crawl and MLCommons!

The main goal of our task is to produce a new LangID dataset solely based on Common Crawl's data that covers as many languages as possible, with the aim of improving our LangID model so that we can discover more content for your language.

In this task, annotators will be first give a prompt in which they select a language that they are proficient on. The bar is a search field so that the annotator can easily find the language they are looking for:

TEXT LANGUAGE IDENTIFICATION  
Label the text with the languages you think it is written in

5 examples created

# Language Annotations

The screenshot shows the GitHub repository page for `commoncrawl/web-languages`. The repository is public and has 43 stars, 44 forks, and 8 watchers. The main branch is selected. The repository description states: "Crowd-sourced lists of urls to help Common Crawl crawl under-resourced languages. See <https://github.com/commoncrawl/web-languages-code> for the code". The repository is categorized under `crawling`, `language-detection`, and `dataset`. The README section is visible, titled "Web Languages Project", and contains the following text: "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



## COLM Workshop

# Open Innovations

# The definition of Open Source

- The Open Source Definition (OSD) from the OSI sets a precedent for clear, permissive licensing. [1]
- Open Data for LLMs needs similar standards to ensure free use, redistribution, and modification of datasets.
- Most datasets used in AI training lack OSD-style guarantees. [2]
- The OSAID builds on OSD principles to include data, models, and training code, explicitly recognising open data as essential to building reproducible and transparent AI systems. [3]
- Without open data, models aren't truly open. Source code is required for open software, so access to open data is necessary for models to be meaningfully open.



# Collaborators' Work in Signals

- Creative Commons proposes a framework to help content stewards express how they want their works used in AI training: CC Signals [1]
- Groups at the IETF (AIPREF [2] and WEBBOTAUTH [3]) are also working towards formal standards for how websites signal preferences about AI-driven access, and also how automated agents identify and behave accordingly.
- Open Future proposes [4] a move beyond the usual copyright-infringement debates, to think of generative AI as a cultural and social technology that is reshaping how societies access, produce, and value information.

# Conclusions

# Conclusions

- Openness is more than access.
- Web-scale datasets deserve web-scale governance.
- We're building trust at the infrastructure layer.
- Openness must include everyone.



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>