The Hidden Authority Signal: Why Your CC Rank May Matter More for AI Visibility

Metehan Yesilyurt

AI Search & Information Retrieval Researcher · metehan.ai

Published: January 07, 2026 • Canonical: https://metehan.ai/blog/cc-rank/

We've been talking about LLMs and web search for a long time now, often finding ourselves in the middle of various debates and discussions.

Recently, I've been exchanging ideas with C-level executives from enterprise companies and publicly traded corporations. One question kept coming up:

What about the training data?

Are there reasons why certain domains get recommended so frequently?

How can we even find the answer to this?

We know that ChatGPT and other LLMs are far behind Google when it comes to link metrics.

I've been working on Common Crawl datasets for about 7 months now. I had to abandon several projects because of the sheer size of the data. But as I started playing with other datasets that Common Crawl publishes, I found an opportunity to enter a somewhat controversial territory.

I'm not the first to bring up Common Crawl. It's been around for years. Its founder is a former Google employee. They crawl the entire web and host gigabytes (let's call it petabytes) of open-source data, sponsored by Amazon. We can verify that many LLMs use it for training (at least the open-source ones).

But does it have any impact on ChatGPT? That question requires going deeper.

I analyzed a dataset of approximately 607 million domains. Even though it's a massive dataset, you can only see limited data for your own domain at index.commoncrawl.org and it takes time.

So I did two things: I compiled the debates that have been happening between major publishers and the AI world over the past few years, and I built an index of 18 million domains. I created a free tool **-open to everyone on the web-** where you can see your website's CC crawl statistics and, hold on tight, the PageRank and Harmonic Centrality scores that CC has generated.

I even ran an expired domain scan. That was fun. Some results (you can debate their reliability, of course) were quite surprising. A few expired domains had decent PageRank and HC Rank positions. I saved those. I'm running experiments. I even found a website with very low Semrush and Ahrefs authority scores that was getting mentioned across thousands of pages in ChatGPT.

I also made some tests with Harmonic Centrality (calculated by HyperBall) and PageRank (by PageRankParallelGaussSeidel) via applying logistic regression. (Just an opinion -> It addresses the pre-training

phase, think like a pre-filter)

Figure

Anyway, let's dive deeper and get started.

INTRODUCTION: THE DEBATE THAT'S HEATING UP

The relationship between Common Crawl and AI has become one of the most contested topics in tech. In November 2025, The Atlantic's investigation exposed uncomfortable truths about how this nonprofit's data reaches AI companies. Meanwhile, researchers are asking: what role does Common Crawl's domain authority data play in shaping what AI systems know and cite?

This article examines the current debate, synthesizes the key research, and introduces a free tool(find the link at the end) to explore one underexamined dimension: Common Crawl's WebGraph authority metrics.[1]

PART 1: THE COMMON CRAWL CONTROVERSY (2023-2025)

The Atlantic Investigation (November 4, 2025)

Alex Reisner's exposé for The Atlantic[2] revealed that Common Crawl has been supplying paywalled content from major publishers to AI companies, despite claiming otherwise.

Key findings from the investigation:

- Archives contain millions of articles from NYT, WSJ, The Economist, The Atlantic, and others
- CC's scraper bypasses JavaScript paywalls by not executing browser code that checks subscription status
- Reisner found that takedown requests from publishers appear unfulfilled, his research indicates content files haven't been modified since 2016
- CC's public search tool returns misleading "no captures" results for domains that requested removal (over 1,000 domains affected)
- In the past year, CCBot has become the most widely blocked crawler by the top 1,000 websites, surpassing even OpenAI's GPTBot

Notable quotes from Rich Skrenta (CC Executive Director):

> *"The robots are people too" suggesting AI should be allowed to "read the books" for free*

> *"You shouldn't have put your content on the internet if you didn't want it to be on the internet"*

Common Crawl's Response (November 4, 2025):

CC published a defense the same day, stating:

> *"The Atlantic makes several false and misleading claims about the Common Crawl Foundation, including the accusation that our organization has 'lied to publishers' about our activities. Our web crawler, known as CCBot, collects data from publicly accessible web pages. We do not go 'behind paywalls,' do not log in to any websites, and do not employ any method designed to evade access restrictions."*

Skrenta also noted that the archive's file format is "meant to be immutable" and "you can't delete anything from it."

Timeline:

- July 2023: The New York Times requested content removal
- July 2024: Danish Rights Alliance initiated removal requests
- November 4, 2025: The Atlantic publishes investigation

Mozilla Foundation Report (February 6, 2024)

"Training Data for the Price of a Sandwich" provided the most comprehensive analysis of CC's role in LLM development.

Key statistics:

- 64% of 47 LLMs analyzed (2019-2023) used at least one filtered version of Common Crawl (30 out of 47 models)
- GPT-3: Over 80% of tokens came from filtered CC data (Brown et al., 2020)
- CC archive: 9.5+ petabytes, containing billions of web pages
- As of 2024, cited in more than 10,000 academic papers

Critical findings on methodology:

- CC uses **Harmonic Centrality** to determine crawl priority, domains with higher scores are crawled more frequently
- Higher-scoring domains appear more often in training data
- "Digitally marginalized communities less likely to be included" due to this approach
- Popular filtered versions (C4, RefinedWeb, Pile-CC) rely on "simplistic automated filtering" techniques
- CC deliberately doesn't remove hate speech. It wants data useful for researchers studying it

Key quote from CC's main crawl engineer:

> *"Often it is claimed that Common Crawl contains the entire web, but that's absolutely not true. Based on what I know about how many URLs exist, it's very, very small."*

The Washington Post Analysis (April 19, 2023)

WaPo analyzed Google's C4 dataset (filtered Common Crawl) to reveal what's actually in LLM training data.

Findings:

- 15 million websites in C4
- Top sources by token count include: patents.google.com, nytimes.com (#4), latimes.com (#6), theguardian.com (#7), huffpost.com (#9)
- Also includes low-trust sources: RT.com (#65), Breitbart (#159), vdare.com (#993)
- Personal blogs, voter registration databases, and copyrighted content present
- "Experts say many companies do not document the contents of their training data"

Financial Ties (2023)

- OpenAI donated $250,000 to Common Crawl
- Anthropic donated $250,000 to Common Crawl
- NVIDIA listed as "collaborator" on CC's website
- Amazon Web Services sponsors CC's data hosting through Open Data Sponsorship Program

PART 2: WHAT WE KNOW ABOUT LLM CITATION PATTERNS

The Data (2024-2025)

Multiple studies have analyzed which domains LLMs cite most frequently:

Semrush (June 2025) - 150,000+ citations analyzed:

Domain	Citation Frequency
Reddit	40.1%
Wikipedia	26.3%
Google	23%
YouTube	23%

Note: Reddit's dominance likely influenced by Google-Reddit $60M API licensing deal in early 2024.

Profound (August 2024 - June 2025) - 680 million citations analyzed:

Platform	Top Source	Share of Total Citations
ChatGPT	Wikipedia	7.8%
Perplexity	Reddit	6.6%
Google AI Overviews	Reddit	2.2%

Additional findings:

- .com domains represent 80.41% of all citations
- .org sites are second at 11.29%

Search Atlas (August-September 2025) - 5.17 million citations, 907,003 unique domains:

- Commercial domains dominate across all platforms
- Academic and government sources remain underrepresented
- "LLM citations reflect the structure of the public web rather than institutional authority"

What Influences Citations

Research has identified several factors affecting LLM citation selection:

Confirmed factors:

- Content freshness and recency (significant impact)
- Semantic relevance to query
- Structured data and formatting
- Cross-platform presence and brand mentions

Possible contributing factors:

- Frequency of domain appearance in training data
- Authority signals embedded in parametric knowledge
- Platform-specific retrieval preferences

Platform differences:

- ChatGPT: Heavy Wikipedia preference, authoritative knowledge bases
- Perplexity: Strong Reddit emphasis, community-driven content, real-time sources
- Google AI Overviews: Mix of organic SERP signals + forum content (Reddit 21%, YouTube 18.8%, Quora 14.3% among top 10)

Also SERanking published a great study about citation patterns.

PART 3: THE UNDEREXPLORED QUESTION

Common Crawl's WebGraph

What's often overlooked in this debate: Common Crawl doesn't just archive web pages. It also publishes **WebGraph data,**comprehensive domain authority metrics including:

- **Harmonic Centrality (HC)**: Measures how "close" a domain is to all other domains in the link graph
- **PageRank**: Measures authority based on quality and quantity of incoming links
- **Domain rankings**: 607 million domains ranked by these metrics

This data is released every month (covering appr. 94-163 million domains per crawl period) and represents one of the largest publicly available authority datasets.

Figure

The Open Questions

1. Training Data Composition

If Common Crawl prioritizes crawling high-HC domains, these domains appear more frequently in training data. Does this create a baseline familiarity in LLMs with certain sources?

2. Correlation vs. Causation

The domains that rank highest in CC's WebGraph (Facebook, Google, YouTube(is exploding), Wikipedia) are also among the most-cited by LLMs. Is this because:

- They're genuinely authoritative (likely primary factor)
- They were overrepresented in training data (possible contributor)
- They perform well on real-time retrieval signals (confirmed factor)
- All of the above in some combination

3. The Long Tail Problem

Mozilla noted that CC's crawling process underrepresents "digitally marginalized communities." If 100M+ domains exist in CC's "long tail" (rank >1M), how often do these domains appear in LLM responses(offline), regardless of content quality?

4. Authority Threshold

Is there a minimum domain authority level below which LLMs rarely cite a source, even with excellent content and freshness?

PART 4: A TOOL FOR EXPLORATION

Introducing CC Rank Checker

To help explore these questions, I built a free tool that makes Common Crawl's WebGraph data accessible:

https://webgraph.metehan.ai[3]

Figure

Features:

- Check any domain's HC Rank and PageRank across top 10 million domains.
- View rank history across 5 time periods (2023-2025) // Planning to add more, my Mac is dying while creating chunks.
- Track authority changes over time
- Compare up to 10 domains simultaneously
- Browse the Top 1000 domains[3]

What the Data Shows

Top 20 Domains (October-November-December 2025):

Rank	Domain	HC Rank	PageRank
1	facebook.com	#1	#3
2	googleapis.com	#2	#2
3	google.com	#3	#1
4	instagram.com	#4	#5
5	googletagmanager.com	#5	#4
6	youtube.com	#6	#8
7	twitter.com	#7	#10
8	gstatic.com	#8	#7
9	linkedin.com	#9	#12
10	gmpg.org	#10	#9
11	cloudflare.com	#11	#6
12	gravatar.com	#12	#14
13	wordpress.org	#13	#13
14	wikipedia.org	#14	#37
15	apple.com	#15	#19

Interesting observations:

- Wikipedia is #14 in HC but only #37 in PageRank, yet it's ChatGPT's one of the most-cited sources.
- CDN and infrastructure domains (gstatic, jsdelivr, cloudflare) rank extremely high due to being embedded across millions of sites
- Social platforms dominate the top 50

Potential Research Applications

1. **Benchmark your domain** against competitors' CC authority
2. **Track authority trends** over time
3. **Correlate CC rank with citation frequency** in your niche
4. **Identify authority gaps** that content alone may not solve

PART 5: WHAT THIS MEANS FOR AEO

The Balanced View

LLM citation selection is clearly multi-factorial:

Confirmed factors:

- Content quality and relevance
- Freshness and recency[4] (significant impact, studies show 40-60% of cited sources change monthly)
- Structured formatting
- Real-time retrieval performance
- Platform-specific preferences
- Brand search volume and entity recognition

Possible contributing factors:

- Historical training data presence
- Embedded authority associations
- WebGraph-derived signals (direct or indirect)

Practical Takeaways

1. **Don't ignore authority**: While content and freshness matter significantly, domain-level signals likely play some role in the overall equation
2. **Track multiple metrics**: CC Rank is one data point among many, not a silver bullet, but potentially useful for benchmarking
3. **Understand platform differences**: Wikipedia dominates ChatGPT; Reddit dominates Perplexity and Google AI Overviews
4. **The long tail question**: If your domain is in CC's long tail (>1M rank), it's worth investigating whether this correlates with citation challenges
5. **More research needed**: The relationship between CC authority metrics and LLM citations deserves rigorous empirical study

CONCLUSION

The Common Crawl debate isn't just about copyright and paywalls. It's about understanding the foundation of AI's knowledge and potentially, its tendencies toward certain sources.

We know that:

- Most major LLMs were trained on CC data (64% of models studied, 80%+ of GPT-3 tokens)
- CC prioritizes high-authority domains in its crawling via Harmonic Centrality
- These same domains tend to be cited most frequently by LLMs

We don't yet know:

- How much training data composition directly influences citation selection vs. real-time signals
- Whether CC authority metrics have independent predictive value for LLM visibility
- How these factors interact with confirmed signals like freshness and relevance

The CC Rank Checker is a small contribution to making this data accessible. The bigger questions require more research, more data, and more transparency from AI companies about their training data composition.

REFERENCES

Investigations & Reports

- Reisner, A. (November 4, 2025). "Common Crawl Is Doing the AI Industry's Dirty Work." The Atlantic.[2]
- Common Crawl Foundation. (November 4, 2025). Response statement.[5]
- Baack, S. & Mozilla Insights (February 6, 2024). "Training Data for the Price of a Sandwich: Common Crawl's Impact on Generative AI." Mozilla Foundation.[6]
- Baack, S. (June 2024). "A Critical Analysis of the Largest Source for Generative AI Training Data." ACM FAccT 2024.[7]
- Schaul, K., Chen, S.Y., & Tiku, N. (April 19, 2023). "Inside the secret list of websites that make AI like ChatGPT sound smart." Washington Post.[8]

LLM Citation Studies

- Semrush (June 2025). LLM Citation Analysis - 150,000+ citations across 5,000 keywords[9]
- Profound (August 2024-June 2025). AI Platform Citation Patterns - 680 million citations[10]
- Search Atlas (August-September 2025). Industry Patterns in LLM Responses - 5.17 million citations, 907,003 domains[11]

Academic Papers

- Brown, T. et al. (2020). "Language Models are Few-Shot Learners." arXiv:2005.14165[12]
- Raffel, C. et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." arXiv:1910.10683[13]

Common Crawl

- WebGraph Documentation: commoncrawl.org/web-graphs[14]
- WebGraph releases: 94-163 million domains per crawl period

TOOL

Check your domain's Common Crawl authority:

🔗 https://webgraph.metehan.ai[3]

- 18M+ million domains indexed
- 5 time periods (2023-2025)
- HC Rank + PageRank metrics
- Free to use

For the methodology side, here's how AI visibility tools actually collect their data[15].

Related: co-mentions in forums and their AI visibility impact[16].

CC rank is one of the cleaner predictors of AI citations across ChatGPT, Perplexity and Gemini. Pages with strong crawl-time authority signals earn disproportionate AEO and GEO share. Treat CC rank as an answer engine optimization input, not a curiosity.

REFERENCES & SOURCES

[1] **Common Crawl's WebGraph authority metrics.** — https://commoncrawl.org/web-graphs

[2] **Alex Reisner's exposé for The Atlantic** — https://www.theatlantic.com/technology/2025/11/common-crawl-ai-training-data/684567/

[3] **https://webgraph.metehan.ai** — https://webgraph.metehan.ai

[4] **Freshness and recency** — https://metehan.ai/blog/i-found-it-in-the-code-science-proved-it-in-the-lab-the-recency-bias-thats-reshaping-ai-search/

[5] **Common Crawl Foundation. (November 4, 2025). Response statement.** — https://commoncrawl.org/blog/setting-the-record-straight-common-crawls-commitment-to-transparency-fair-use-and-the-public-good

[6] **Baack, S. & Mozilla Insights (February 6, 2024). "Training Data for the Price of a Sand...** — https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/

[7] **Baack, S. (June 2024). "A Critical Analysis of the Largest Source for Generative AI Tra...** — https://dl.acm.org/doi/10.1145/3630106.3659033

[8] **Schaul, K., Chen, S.Y., & Tiku, N. (April 19, 2023). "Inside the secret list of website...** — https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

[9] **Semrush (June 2025). LLM Citation Analysis - 150,000+ citations across 5,000 keywords** — https://www.semrush.com/blog/ai-mode-comparison-study/

[10] **Profound (August 2024-June 2025). AI Platform Citation Patterns - 680 million citations** — https://www.tryprofound.com/blog/ai-platform-citation-patterns

[11] **Search Atlas (August-September 2025). Industry Patterns in LLM Responses - 5.17 million...** — https://searchatlas.com/blog/domain-industry-analysis-in-llm-responses/

[12] **Brown, T. et al. (2020). "Language Models are Few-Shot Learners." arXiv:2005.14165** — https://arxiv.org/abs/2005.14165

[13] **Raffel, C. et al. (2020). "Exploring the Limits of Transfer Learning with a Unified Tex...** — https://arxiv.org/abs/1910.10683

[14] **commoncrawl.org/web-graphs** — http://commoncrawl.org/web-graphs

[15] **how AI visibility tools actually collect their data** — https://metehan.ai/articles/how-ai-visibility-tools-collect-data/

[16] **co-mentions in forums and their AI visibility impact** — https://metehan.ai/articles/co-mentions-in-forums-ai-visibility-impact-reddit-quora-and-beyond/