METEHAN.AI RESEARCH REPORT · PREPRINT COMPANION GUIDE

A Test Protocol for Common Crawl Authority and AI Citation Visibility

Metehan Yesilyurt

AI Search & Information Retrieval Researcher · metehan.ai

https://metehan.ai/pdf/common-crawl-authority-test-protocol.pdf

Abstract. A step by step measurement protocol for testing whether Common Crawl WebGraph authority metrics (HC Rank and PageRank) track with how often a domain gets cited by AI systems. It covers unit definition, sampling frames, instrumentation, decision rules, a worked example, and the failure modes that invalidate most informal CC rank comparisons.

Keywords: Common Crawl, AI citations, measurement protocol, Harmonic Centrality, LLM visibility, test design

Companion to the article The Hidden Authority Signal: Why Your CC Rank May Matter More for AI Visibility. The article covers the topic in full; this document is a separate, practical measurement methodology and test protocol.

1. Scope, Hypotheses, and What Would Falsify Them

I use this protocol when someone asks me whether Common Crawl authority explains their weak AI visibility. It is a measurement procedure, not a summary of anyone else's findings. The goal is to produce a result you can defend, including a result that says the effect is not present in your niche.

Start by writing three hypotheses in a form that can fail. A hypothesis that cannot fail is not worth measuring.

- H1, within-niche gradient: inside one topic niche, domains with stronger HC Rank and PageRank hold a larger share of citations across ChatGPT, Perplexity, and Google AI Overviews.
- H2, incremental value: CC authority predicts citation share after you account for content freshness, topical relevance, and brand demand.
- H3, platform split: the gradient is steeper on some platforms than others, in line with the source preferences reported at aggregate level [5].

Write the falsification conditions before you collect anything. For H1, the condition is a flat or inverted gradient across rank bands that persists across repeated runs. For H2, the condition is that the gradient disappears once brand demand and page freshness are controlled. For H3, the condition is that the same gradient appears on every platform with no meaningful spread.

What this protocol does not do: it does not estimate how much of any model's training data came from your domain, and it does not separate pre-training effects from retrieval effects. Those questions require training corpora and retrieval logs that outside researchers do not have. The protocol measures association inside a niche and gives you a defensible basis for acting on it.

2. Define Your Units and Sampling Frame

Most informal CC rank comparisons fail here. They pick a handful of domains, run one prompt, and then reason about a mechanism. Fix the units first.

- Domain unit: the registrable domain, not the subdomain. Subdomains matter for infrastructure hosts, but for publisher comparisons the registrable domain is the right grain.
- Prompt unit: a single question submitted to a single platform on a single day.
- Citation event: one URL from a tracked domain appearing in the answer body or in the sources list. Count it once per prompt per platform per day, no matter how many times the domain appears.
- Window: the calendar period covered by one run. Keep it short enough that freshness effects do not drift inside a single run.

For domains, pick 30 to 60 inside one niche. Fewer than 30 and the rank correlation is unstable. More than 60 and manual relevance scoring becomes the bottleneck. Stratify by HC Rank using the bands exposed by the free tool, which covers the top 10 million domains [4]:

- Band 1: rank 1 to 10,000.
- Band 2: rank 10,001 to 100,000.
- Band 3: rank 100,001 to 1,000,000.
- Band 4: rank beyond 1,000,000, the long tail discussed in the source analysis [1].

For prompts, write 40 to 100 brand-free questions. Brand-free means no company names, no product names, no domain names. A prompt that contains a brand converts the test into a brand lookup and tells you nothing about authority. Split prompts by intent so you can see whether the gradient depends on the kind of question being asked.

3. Instrumentation: What to Record and Where It Comes From

Two data streams need to be joined: authority on the crawl side, and citations on the answer side.

On the crawl side, pull HC Rank and PageRank for every domain in your list. The tool exposes both metrics across the top 10 million domains, tracks five periods between 2023 and 2025, and lets you compare up to 10 domains at a time [4]. Record the snapshot period next to every value, because the underlying WebGraph releases cover roughly 94 to 163 million domains per crawl period [3] and values move between releases. For context on what these metrics represent, the WebGraph documentation describes the graph structure they are computed over [3].

On the answer side, decide how you collect before you start. Different collection methods produce different citation sets, and the differences are large enough to change conclusions [11]. Whatever method you choose, store the raw answer text and the raw citation list. Storing only a count throws away the evidence you need later.

Two covariates are easy to forget. First, the age of the specific page that got cited, not the age of the domain [12]. Second, whether the domain is being co-mentioned in forums, which is a separate visibility channel from being cited [13].

One caveat for historical work. Common Crawl's archive format has been described by its executive director as meant to be immutable, and removal requests have been contested publicly [8][14]. If your study window overlaps

periods when removal requests were filed, note that in the writeup rather than ignoring it.

4. The Measurement Table

Record these fields for every domain in every run. The table is the schema; keep it stable across runs so the comparison stays valid.

Metric	Where it comes from	Unit	Cadence	Role in the test
HC Rank	CC WebGraph via the free tool [4]	Ordinal rank	Monthly	Primary independent variable
PageRank	CC WebGraph via the free tool [4]	Ordinal rank	Monthly	Secondary independent variable
Citation share	Your prompt panel [11]	Percent of citation events	Per run	Primary dependent variable
Citation rate per prompt	Your prompt panel	Count per 100 prompts	Per run	Secondary dependent variable
Cited page age	Crawl of the cited URLs	Days	Per run	Covariate [12]
Topical relevance	Human rubric, 0 to 3	Score	Per domain-prompt pair	Covariate
Forum co-mention count	Forum search [13]	Count	Monthly	Covariate
Brand demand proxy	Search volume or branded query share	Index	Monthly	Covariate

5. Test Execution: Ordered Steps

1. Freeze the domain list and the prompt list in a versioned file. Any change after this point starts a new run.
2. Pull HC Rank and PageRank for every domain and record the snapshot period alongside each value [4].
3. Run the full prompt panel on each platform on the same day, in a logged-out session, with no personalization signals.
4. Capture raw answers and citation lists before any processing.
5. Tag every citation event with domain, prompt, platform, and date.
6. Score topical relevance for each domain-prompt pair using the same rubric across all domains.
7. Compute citation share per domain: citation events for that domain divided by total citation events in the run.
8. Assign each domain to its HC Rank band and compare citation share across bands.
9. Run a rank correlation between HC Rank and citation share. With fewer than 50 domains, rank methods are more honest than comparing band means.
10. Repeat the entire run at least three times, spaced at least two weeks apart, and report the spread across runs rather than the best single run.

Log every deviation from the plan. If you added a prompt mid-run, or if a platform returned an error for part of the panel, that belongs in the record. The deviations are often more informative than the headline number.

6. Decision Rules

These are the rules I apply when reading the output. They are conventions I use, not findings from anyone's study.

What you observe	What it licenses you to say	What to do next
Citation share rises monotonically across HC Rank bands and repeats across three runs	H1 is supported in this niche	Use CC rank as a screening variable when prioritizing content and link work [1]
No gradient, and cited page age explains most of the variance	H1 is not supported here	Shift budget to refresh cadence [12]
Gradient holds on one platform and not another	H3 is supported	Report platform by platform, following the split seen in aggregate citation studies [5]
Band 4 domains are cited at the same rate as Band 1 domains	No long-tail penalty in this niche	Do not treat a rank above 1,000,000 as a blocker [1]
Gradient disappears after controlling for brand demand	H2 is not supported	Measure brand demand before blaming crawl authority
Gradient is present but weak and the sample is under 30 domains	Inconclusive	Add domains before acting

7. Worked Example: A Mid-Size Publisher

Suppose I run this for a B2B publisher. The niche has 45 domains and 60 brand-free prompts across three platforms, so one run produces 180 prompt-platform opportunities.

The publisher's domain sits in Band 3. Two competitors sit in Band 2 and Band 4. I compute citation share for each: citation events for the domain divided by total citation events across all 45 domains in the run. Then I compare that share against the band.

For scale, keep the aggregate benchmark in view. Across 680 million citations, Wikipedia accounted for 7.8% of ChatGPT citations, Reddit for 6.6% of Perplexity citations, and Reddit for 2.2% of Google AI Overviews citations [5]. A niche domain will never approach those shares. What matters is ordering inside the niche, not the absolute number. Two other studies are useful as sanity checks on the shape of the distribution: a 150,000 citation analysis that put Reddit at 40.1% and Wikipedia at 26.3% [6], and a 5.17 million citation analysis across 907,003 unique domains [7].

If the Band 2 competitor leads, the Band 3 publisher sits in the middle, and the Band 4 domain trails, the ordering is consistent with H1. If the Band 4 domain leads, H1 fails for this niche and I stop treating CC rank as an explanation for the gap.

The arithmetic is deliberately simple. One citation event out of 180 opportunities is a citation rate of roughly 0.56%. That is the resolution you are working at, which is why a single run cannot settle anything and why three runs is a floor rather than a target.

8. Failure Modes and Falsification Checks

- Brand contamination: any prompt containing a company or product name converts the test into a brand lookup and inflates citation rates for well-known domains.
- Personalization: logged-in sessions, location signals, and prior chat history change retrieval. Use clean sessions and record the setup.
- Single-run conclusions: cited sources churn between periods, and freshness effects are large enough to dominate a single snapshot [12].
- Treating rank as cardinal: rank 10 and rank 11 are indistinguishable for this purpose. Rank 10 and rank 10,000 are not.
- Infrastructure distortion: CDN and infrastructure hosts sit at the top of the WebGraph because they are embedded across millions of sites [1]. Exclude them from publisher comparisons.
- Universal-condition error: 30 of 47 LLMs analyzed between 2019 and 2023, or 64%, used at least one filtered version of Common Crawl [2], and GPT-3 drew more than 80% of its tokens from filtered CC data [10]. Presence in CC is close to a baseline condition, so it cannot by itself explain why one domain is cited over another.
- Dataset-composition blind spot: the Washington Post analysis of the C4 dataset found 15 million websites in it, with a mix of high-circulation news domains and low-trust sources [9]. A domain's presence in a filtered crawl says little about how it is weighted.
- Causation overreach: the domains that top the WebGraph are also among the most cited, but that co-occurrence does not establish a mechanism [1]. Treat the gradient as a screening signal until brand demand and freshness are ruled out.
- Removal-history blind spot: if your window overlaps contested removal requests, note it explicitly [8][14].

Report the negative result when you get one. A niche where CC rank shows no gradient is a useful finding, because it tells you to spend on freshness and relevance instead of chasing an authority metric you cannot move quickly. The protocol earns its keep by ruling things out as much as by confirming them.

REFERENCES & SOURCES

[1] **Metehan Yesilyurt, The Hidden Authority Signal: Why Your CC Rank May Matter for AI Visibility**, https://metehan.ai/blog/cc-rank/

[2] **Mozilla Foundation, Training Data for the Price of a Sandwich**, https://www.mozillafoundation.org/en/research/library/generative-ai-training-data/

[3] **Common Crawl WebGraph documentation**, https://commoncrawl.org/web-graphs

[4] **CC Rank Checker tool**, https://webgraph.metehan.ai

[5] **Profound, AI Platform Citation Patterns**, https://www.tryprofound.com/blog/ai-platform-citation-patterns

[6] **Semrush, LLM citation analysis**, https://www.semrush.com/blog/ai-mode-comparison-study/

[7] **Search Atlas, Industry Patterns in LLM Responses**, https://searchatlas.com/blog/domain-industry-analysis-in-llm-responses/

[8] **Alex Reisner, The Atlantic, Common Crawl investigation**, https://www.theatlantic.com/technology/2025/11/common-crawl-ai-training-data/684567/

[9] **Washington Post, Inside the secret list of websites that make AI like ChatGPT sound smart**, https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

[10] **Brown et al., Language Models are Few-Shot Learners**, https://arxiv.org/abs/2005.14165

[11] **Metehan.ai, How AI visibility tools actually collect their data**, https://metehan.ai/articles/how-ai-visibility-tools-collect-data/

[12] **Metehan.ai, The recency bias reshaping AI search**, https://metehan.ai/blog/i-found-it-in-the-code-science-proved-it-in-the-lab-the-recency-bias-thats-reshaping-ai-search/

[13] **Metehan.ai, Co-mentions in forums and AI visibility impact**, https://metehan.ai/articles/co-mentions-in-forums-ai-visibility-impact-reddit-quora-and-beyond/

[14] **Common Crawl Foundation, Setting the record straight**, https://commoncrawl.org/blog/setting-the-record-straight-common-crawls-commitment-to-transparency-fair-use-and-the-public-good