Prompt Selection, Denominators and Repeat Noise in AI Citation Tracking: A Recount of Saved Scraper Runs

Ankur Shrestha

Caldrin, United States
ORCID: 0009-0009-6036-5849

12 November 2026

Abstract

Reports of AI visibility often present a citation rate as a trend line. This paper asks how much of a reported movement can come from the counting specification and repetition noise rather than changed answers. It recounts saved runs from one project: 48 informational prompts, three repetitions each, collected twice on 1 November 2026 through a ChatGPT scraper and a Gemini scraper under identical settings. Twenty domain and scraper pairs, the ten most cited domains per scraper in the first sweep, are scored under ten counting rules, and moves are judged against relabeling nulls. For one domain, the repeat sweep gave 27 of 144 runs (18.75%) over all prompts and 23 of 45 runs (51.11%) over the prompts that had cited it in the first sweep. These are run-presence rates on different prompt populations. A share of 27 of 647 citation slots (4.17%) uses a different unit and is not a comparable rate. Across the 20 pairs the selected-prompt rate was a median 4.0 times the default rate (range 2.1 to 12.0). Under the specified selection-only null, 18 of 20 observed selected-prompt rates fell within the central 95% simulation range; this is a compatibility result under an exchangeability assumption, not evidence that selection caused the difference. The direction of the change between the two sweeps depended on the counting rule for 10 of 20 pairs. The largest of ten selected moves reached 7 runs in 52% to 63% of relabelings, against 0% to 26% for a single domain chosen in advance. Between 49% and 68% of the variance in a domain's citation outcome (median over domains) lay between repetitions within a sweep, and 0% to 4% lay between sweeps. The records support a methods demonstration, not a population estimate: one date, one repeat, two scrapers and one project's prompts. The paper closes with a reporting checklist derived from these results.

Keywords: generative search, AI search, answer engines, citation share, measurement, denominators, repeatability, prompt selection

Scope. This is a methods demonstration on saved runs from one project: 48 prompts, one date (1 November 2026), two repeated sweeps, two commercial scrapers, domain-level counts. It estimates nothing about AI search in general and nothing about any named company.

Labels. All analyses beyond the reproduction of earlier counts were chosen after the data were seen and are exploratory. No hypothesis was registered.

1 Introduction

A company that tracks how often AI answer engines cite its domain usually reports one number per period, a share of prompts or runs in which the domain appears, and reads the change from one period to the next as a trend. That reading assumes the number would stay put if nothing about the answers changed. Three features of the measurement make that assumption doubtful. The answers differ from run to run. The prompt set, the repetition count and the denominator can change between reports. And the domains chosen for display are often the ones that stood out in an earlier report, which builds selection into the comparison.

The question here is narrow. Given saved answers whose content is fixed, or repeated under identical settings, how much does a reported citation rate for a domain depend on the counting specification, and how large are the moves that repetition and selection alone produce?

Four sub-questions follow.

- **RQ1.** How many different levels can one set of saved answers report for one domain under defensible counting rules?
- **RQ2.** When prompts are chosen because the domain was cited on them in an earlier sweep, is the higher rate on those prompts compatible with that choice alone?
- **RQ3.** In a repeat of an identical measurement, how large are the sweep-to-sweep moves, how do they compare with the noise implied by repetition and prompt clustering, and how does choosing the largest movers change what looks unusual?
- **RQ4.** What share of the variance in a domain's citation outcome lies between prompts, between sweeps and between repetitions?

What is new and what is not. That AI answers vary between runs is established (Schulte et al., 2026; Sielinski, 2026; Sivakumar, 2026), and bootstrap intervals for citation share have been reported (Sielinski, 2026). That published citation figures disagree because of denominators and prompt universes has been documented in a secondary review (Kunal, 2026). A practitioner source states that a visibility score depends on the prompt set, schedule, engine mix and formula (Parrott, 2026). None of those claims is new here. The contribution is a primary recount that attaches counted magnitudes to them, in a design where the answers are saved and can be rescored, with intervals that resample prompts and not runs: the same answers scored ten ways, a null for the restricted-prompt rate, a comparison of the largest-of-ten move with the move of a domain chosen in advance, and a variance decomposition into prompts, sweeps and repetitions.

Relation to the author's earlier web article. A shorter account of part of this dataset was published on 6 November 2026 on the author's company website (Shrestha, 2026; Appendix A). It reported one matched recount, 27 of 144 runs against 23 of 45 (run-presence rates on different prompt populations), and two domain moves. The present paper reuses those counts, extends them to all 20 domain and scraper pairs, and adds the analyses listed above. The article is a first-party publication of part of the same data. It is not independent confirmation of anything reported here, and no result in this paper rests on it. The article said in words that its relabeling count was for the largest of ten selected domains; this paper quantifies that point by also reporting the probability for a domain chosen in advance, which is much smaller. Sections 5 to 7 contain analyses that were not in the article.

Relation to a separate working paper. The author has a separate exploratory working paper on the overlap between Google results and AI answers for B2B software buying prompts. It uses its own 60 prompts and its own collection, and it does not use the saved runs analysed here.

2 Related work

Repeated runs and uncertainty. Schulte et al. (2026) submitted eight prompts for each of four Swiss-German verticals to four AI search engines, daily for 45 to 46 days and up to ten times within a day, and compared cited-source and brand sets by Jaccard similarity and rank-biased overlap. Source sets overlapped by roughly a third both across consecutive days and across repeats within 24 hours (Jaccard 0.32 to 0.43 within 24 hours), brand sets somewhat more, and similarity varied widely by prompt. The authors conclude that visibility should be characterized from repeated runs, not a single observation. Their uncertainty analysis is a subsampling standard error for per-brand detection rates within engine-and-prompt groups, from which they suggest about seven repeated runs per prompt; they report no intervals for shares aggregated across prompts. Sielinski (2026) treats citation share and prevalence as sample estimators, using 200 generated queries for each of three consumer-product topics sent to three platforms daily for nine days and, for one topic, every ten minutes. Percentile bootstrap intervals from resampling responses within a single sample are several points wide for the most cited domains, so that many apparent differences between domains cannot be distinguished from sampling noise, and weighted rank correlations between consecutive samples of frequently cited domains fall below the author's 0.9 benchmark in many platform-and-topic combinations. That bootstrap does not treat prompts as clusters, and the paper reports no design effects or variance components; its one example of an extreme mover is given after the fact, without a selection null. Sivakumar (2026) put 280 B2B software questions to four AI search products in one collection window and compared Jaccard agreement of the named-vendor sets within an engine, using three runs of a 25-question subsample, with agreement between engines, reporting 95% bootstrap intervals. The within-engine advantage held for Google AI Overviews under every length control but for Claude fell to a difference indistinguishable from zero once answers were truncated to five vendors or an overlap coefficient was used, so Jaccard comparisons across sets of different length can mislead. That study is operator-run and not peer reviewed, and collection was incomplete for two engines (18 ChatGPT and 175 Perplexity answers). It does not examine prompt selection, the selection of largest movers or variance components. Bal (2026), known here from its abstract only, reports a four-week panel on enterprise banking software in which engines agreed on average (mean Jaccard 0.65) while leadership among vendors changed eleven times in 28 days; the abstract does not say how repeat noise bears on those leadership changes.

Why published figures disagree. Kunal (2026) reviews 30 published evidence sources (seven academic, 22 commercial datasets or reports and one reported legal filing), traces headline citation-share figures to their publishers, and codes each for its metric definition, denominator, selection stage (retrieved, cited or mentioned), engine, prompt universe and time window. He attributes much of the disagreement among published figures to different denominators and prompt universes, offers a reconciliation procedure and checklist, and proposes a future mixed-effects panel design to partition variance across engine, prompt, week and repeated run. The

review is secondary: it recounts no raw answers, reports no intervals, variance components or tests, and does not discuss the selection of extreme movers. A practitioner article states the general point that changing a score's inputs moves it when the brand did not change (Parrott, 2026). A variance-components decomposition of classifier sentiment scores for LLM answers about brands (Żatuchin, 2026, version 2) illustrates how a variance share depends on the population fitted: the share attributed to query language fell from 26.5% to 5.1% when English answers were excluded. Its response is a classifier score, not a citation measure, and it reports no intervals.

Statistical background. Selecting the extreme values of a noisy measure on one occasion and re-measuring them on another produces a predictable drift toward the middle, and the largest of several moves is larger than the typical single move. This is standard statistics and is used here without citation. The prompt bootstrap follows Efron (1979).

Position. Among these sources, none rescored one set of saved AI answers under several counting rules, tested a restricted-prompt rate against a selection-only null, or contrasted the largest of several selected moves with the move of a domain chosen in advance. Prompt-clustered resampling, design effects and a split of variance into prompts, sweeps and repetitions likewise do not appear in the full texts read. This describes the sources listed here, not the wider literature, and the paper makes no claim of novelty for the points listed under "What is new and what is not" in Section 1.

3 Data and records

Collections. The records are two sweeps on 1 November 2026. Sweep 1 ran from 14:32:21 to 14:40:19 UTC (run start times). Sweep 2, a repeat of the same prompts with the same collector and settings, ran from 17:20:59 to 18:04:35 UTC. The two sweeps began about 2.8 hours apart. The repeat was run as a same-day controlled test under a written protocol.

Instruments. Two scraper surfaces were used: the DataForSEO LLM Scraper for ChatGPT (with web search forced) and for Gemini, US location, English, standard queue, up to two attempts until sources appeared. They are scraper products. They are not observations of what individual ChatGPT or Gemini users saw, and they carry no personalization and, for ChatGPT, only country-level location.

Prompts. The collection has 61 prompts, each run three times per sweep, which gives 183 runs per scraper per sweep. Forty-eight prompts are informational or problem-seeking and come in eight topical clusters of six (AI visibility measurement artefacts; the Search Equity concept; Google versus AI recommendations; query universe mapping; AI source concentration; signal-based outbound; original research and AI citation; and technical eligibility and llms.txt). The other 13 are entity prompts (5 branded, 5 founder and 3 concept) and are used here only in one denominator rule. The prompts are the author's wording. They are not buyer-sourced and the clusters follow the author's company's own research topics. Every analysis in Sections 5 and 6 uses the 48 informational prompts unless stated otherwise.

Completion and models. All 366 ChatGPT-scraper runs and all 366 Gemini-scraper runs across the two sweeps were completed, with no failed runs for either scraper in either sweep. The ChatGPT scraper reported one model, gpt-6, on all 366 runs. The Gemini scraper reported

3.5 Flash-Lite on 363 runs and 3.6 Flash on 3 runs (1 in sweep 1, 2 in sweep 2). Those three runs are retained in every analysis, as in the earlier article. Two sensitivity checks followed: counting the three as runs with no citations, and dropping the two affected prompts from the 48 in both sweeps. Neither changed the 18 of 20 pairs inside the null, the two intervals that exclude zero or the 7-run largest move. They did move three Gemini items: the top-ten overlap between sweeps fell from 6 to 5 of 10, the Gemini probability of a largest move of 7 runs rose from 52.9% to 58.6% in the first check, and the sweep 1 leader became a tie in the second. Pairs whose direction depends on the rule fell from 10 to 9 in the second check (`verification/VERIFICATION-REPORT.md`).

Cited domains. A run cites a domain if any URL in the answer's source list has that host. A host is lowercased with a leading "www." removed. Subdomains count as separate domains (for example a company's blog subdomain and its main domain). No public-suffix reduction was applied. A completed run whose parsed record lists no source URLs cites no domain and is called ungrounded here; the 48-prompt subset has 5 such runs for ChatGPT in sweep 1 and 8 in sweep 2, and 26 and 27 for Gemini (Section 6.4).

Reference panel. The ten most cited domains per scraper on the 48 prompts in sweep 1 were written to a panel file under the same protocol. For Gemini the tenth place was a three-way tie at 7 runs, resolved in the file. The panel is selected on sweep 1, which matters for every comparison with sweep 2 (Section 6.4). Gemini ranks 5 to 7 are tied at 9 runs, and the ranks shown in the figures follow the order in the panel file. The panel contains 20 domain and scraper pairs. Three are labeled as in the earlier article: Domain A (a Gemini panel domain, rank 4), Domain B (a ChatGPT panel domain, rank 3) and Domain C (the ChatGPT panel domain at rank 1). The supplementary tables name every domain.

4 Measures and definitions

Run, repetition, sweep. A run is one prompt submitted once. A repetition is one of the three runs of a prompt within a sweep. A sweep is the 144 runs (48 prompts, three repetitions) of one scraper on one occasion.

Presence rate. For a domain, the presence rate in a sweep is the number of runs that cite it divided by the number of runs in the stated denominator. The default denominator is the 144 runs of the sweep.

Counting rules. Besides the default, nine alternatives were scored on the same saved answers. (1) Runs that returned at least one source. (2) All 61 prompts in the collection (183 runs). (3) Prompt reach, the share of the 48 prompts on which at least one of three repetitions cites the domain. (4) A majority rule, the share of prompts on which at least two of three do. (5 to 7) A single repetition (first, second or third) over the 48 prompts. (8) Citation-slot share, runs citing the domain divided by the total number of distinct cited domains summed over all runs in the sweep; this is a share of slots, not of runs, so it sits on a different scale. (9) The selected-prompt rule defined next.

Selected-prompt rate. Let K be the prompts on which a domain was cited in at least one of the three sweep 1 runs. The selected-prompt rate is the number of sweep 2 runs on K that cite the domain divided by 3 times the size of K.

Mixture identity. With w the share of the 144 sweep 2 runs that lie on K, the default rate equals w times the selected-prompt rate plus (1 minus w) times the rate on the prompts outside K. This is arithmetic.

Cited-set Jaccard. For a prompt, the cited set is the union of the domains cited over its three repetitions. The sweep-to-sweep Jaccard is the size of the intersection of the two sweeps' sets divided by the size of their union. The repetition-to-repetition Jaccard is computed for each pair of repetitions within a sweep. Prompts whose sets are empty in both are excluded.

Largest move. For the ten domains selected on sweep 1, the largest move is the maximum over the ten of the absolute difference in runs citing the domain between sweep 2 and sweep 1.

5 Analyses

Everything in this section other than the reproduction of the earlier counts was chosen after the data had been seen and after the earlier article was published. It is exploratory. Seeds are in Appendix B.

5.1 Reproduction. The earlier counts and the earlier analysis script were re-run from the raw responses and compared with the stored results. Panel counts for all 20 pairs were recomputed from the raw responses and compared with the panel file.

5.2 Selection-only null for the selected-prompt rate. For each domain, the six runs of each prompt (three per sweep) were pooled. In each of 5,000 draws, three of the six were dealt at random to a pseudo sweep 1 and three to a pseudo sweep 2, the prompts with at least one pseudo sweep 1 citation were selected, and the pseudo sweep 2 rate on them was computed. The null therefore keeps each prompt's tendency to cite the domain and removes any difference between sweeps. The null is a model-based compatibility check. If the six runs of a prompt are exchangeable across the two sweeps, a selected-prompt rate inside the central 95% of the null is compatible with selection alone. A selection effect is a mathematical property of choosing prompts on an earlier result; the null does not identify what caused the observed difference.

5.3 Counting rules. The ten levels were computed for each of 20 pairs in each sweep. A change between sweeps under a rule counts as an increase or decrease only if it is at least 1 percentage point in size; smaller changes count as none. This threshold was set after seeing the data. The slot-share rule is excluded from the direction comparison because it is on a different scale.

5.4 Uncertainty and design effect. The 48 prompts were resampled with replacement, both sweeps together, 20,000 times for each pair, and the sweep-to-sweep difference recomputed. The standard deviation of the resampled differences is the paired prompt-bootstrap standard error. The naive binomial standard error treats the 144 runs as independent. The design effect is the square of the ratio of the prompt-bootstrap standard error of one sweep's rate to the binomial standard error. The approximate 95% noise threshold used throughout is 1.96 times the paired prompt-bootstrap standard error: the size of a difference that would just fall outside a 95% interval centered on zero. It is not a formal minimum detectable effect calculated for a specified statistical power. Where a power-based comparison is wanted, Section 6.5 also reports 2.8 times the standard error, the difference detected with 80% power under a normal approximation.

5.5 Variance components. For each pair, the 0/1 outcome over 48 prompts, 2 sweeps and

3 repetitions was decomposed with a balanced nested design (sweeps nested within prompts) by the method of moments. Negative estimates were set to zero. This is descriptive: with two sweeps, the sweep component has few degrees of freedom and no interval is reported.

5.6 Largest-move null. The six runs of each prompt were relabeled between pseudo sweeps. Two ways of drawing were used. Whole-run relabeling sends three of a prompt's six runs, each with all its cited domains, to pseudo sweep 1, which preserves co-citation within an answer. Cell-wise relabeling deals each prompt and domain cell separately, which is the procedure in the earlier article. In each draw the ten most cited domains in pseudo sweep 1 were selected afresh (ties broken at random) and the largest absolute move among them recorded. The statistic observed in the data is the largest move over the frozen panel. The same draws give, for each panel domain, the probability that a domain chosen in advance moves by at least a given number of runs. Two draw counts were used: 2,000, as in the earlier article, and 20,000 for precision.

5.7 Leaderboard and cited-set stability. The ten most cited domains were listed in each sweep. Ties at the cutoff were flagged and a tie-inclusive set (every domain with at least the tenth-place count) was also computed. Prompt-level Jaccard values were computed as defined in Section 4.

6 Results

6.1 Reproduction of the earlier counts

The recount reproduced. Domain C (ChatGPT scraper) was cited in 28 of 144 runs in sweep 1 and 27 of 144 in sweep 2. Domain B was cited in 22 and then 29 of 144, and Domain A (Gemini scraper) in 10 and then 3 of 144. In sweep 1 Domain C appeared in 10 of 48 first repetitions, 7 of 48 second repetitions and 11 of 48 third repetitions. Panel counts for all 20 pairs recomputed from the raw responses matched the panel file. The largest absolute sweep-to-sweep difference was 4.86 percentage points for each scraper, and the median absolute difference over the ten panel domains was 2.09 (2.085 unrounded; ChatGPT) and 2.78 (Gemini). Prompt-bootstrap intervals excluded zero for 0 of 10 ChatGPT domains and 2 of 10 Gemini domains; two further ChatGPT intervals end at 0.0, and the two Gemini intervals end one and two runs (0.69 and 1.39 points) from zero, so these counts are not sharp. The earlier article's relabeling counts (1,209 of 2,000 for ChatGPT and 1,071 of 2,000 for Gemini) were reproduced in size with the new seeds: 1,205 and 1,074 of 2,000 under the cell-wise draw. The Gemini scraper reported a different model on 1 of 183 runs in sweep 1 and 2 of 183 in sweep 2.

6.2 One domain, ten reported levels (RQ1)

The same saved answers gave Domain C a repeat-sweep level of 14.58% to 31.25% in rows 1 to 8 of Table 1, which are related measures of incidence: run-presence rates count occurrence across runs, and the prompt-reach and majority rules count occurrence across prompts. They are not interchangeable estimates of one quantity, and they are tabulated together to show how far the reported level moves with the rule. Two further levels use other bases. 51.11% (23 of 45 runs) is a run-presence rate on a different prompt population, the 15 prompts cited in sweep 1, and 4.17% (27 of 647) is a share of citation slots, a different unit from a rate per run. Neither is directly comparable with the default rule's 18.75% (27 of 144 runs). Table 1 and Figure 1 list

all ten.

Table 1. Domain C, repeat sweep (sweep 2), ChatGPT scraper: the same saved answers under ten counting rules. Rows 1 to 3 and 6 to 8 are run-presence rates (citation occurrence across runs); rows 4 and 5 are prompt-level rates (occurrence across prompts); row 9 uses citation slots as its unit; row 10 uses a different prompt population. The rows are related incidence measures, not interchangeable estimates of one quantity.

Counting rule	Measure; unit; prompt population	Numerator / denominator	Level
1. All 48 prompts, all runs (default)	run presence; runs; 48 prompts	27 / 144	18.75%
2. Runs that returned any source	run presence; runs with a source list; 48 prompts	27 / 136	19.85%
3. All 61 prompts in the collection	run presence; runs; 61 prompts	27 / 183	14.75%
4. Prompt reach (at least 1 of 3 runs)	prompt presence; prompts; 48 prompts	15 / 48	31.25%
5. Majority of three (at least 2 of 3 runs)	prompt presence; prompts; 48 prompts	7 / 48	14.58%
6. First repetition only	run presence; one run per prompt; 48 prompts	7 / 48	14.58%
7. Second repetition only	run presence; one run per prompt; 48 prompts	9 / 48	18.75%
8. Third repetition only	run presence; one run per prompt; 48 prompts	11 / 48	22.92%
9. Share of all citation slots	slot share; slots; 48 prompts	27 / 647	4.17%
10. Selected prompts only	run presence; runs; 15 prompts cited in sweep 1	23 / 45	51.11%

Across all 20 pairs, the range of sweep 2 levels over those eight rules had a median width of 11.46 percentage points for each scraper (ChatGPT range 4.17 to 17.48; Gemini range 2.09 to 22.92). The highest of those levels was a median of 1.86 times the default rate.

6.3 Selected prompts and the selection-only null (RQ2)

The selected-prompt rate was above the default rate for all 20 pairs. Under the specified selection-only null, 18 of 20 observed selected-prompt rates fell within the central 95% simulation range. The mixture identity, which is arithmetic, shows why the two rates differ. It holds for all 20 pairs. For Domain C the 15 prompts cited in sweep 1 carry 45 of the 144 sweep 2 runs, a weight of 0.3125, and the domain is cited in 23 of those 45 (51.11%). On the other 33 prompts it is cited in 4 of 99 runs (4.04%). So 27 of 144 is a mixture of two very different rates, and the restricted view reports only the first.

The ratio of the selected-prompt rate to the default rate had a median of 3.98, with a range from 2.06 to 12.00. The selection-only null (Section 5.2) centers on a selected-prompt rate of 51.17% for Domain C, with a central 95% range from 40.74% to 64.10%; the observed 51.11% lies near that center, which is largely mechanical because the null pools the observed runs (Figure 2). The two pairs outside the range were Gemini panel domains whose observed rate was below it: Domain A (3 of 18 runs, 16.67%, against a null range of 20.00% to 66.67%) and the domain at Gemini rank 7 (4 of 12, 33.33%, against 41.67% to 83.33%). These are the same two pairs whose bootstrap intervals exclude zero (Section 6.5).

The 18-of-20 result is a model-based compatibility result under exchangeability of a prompt's

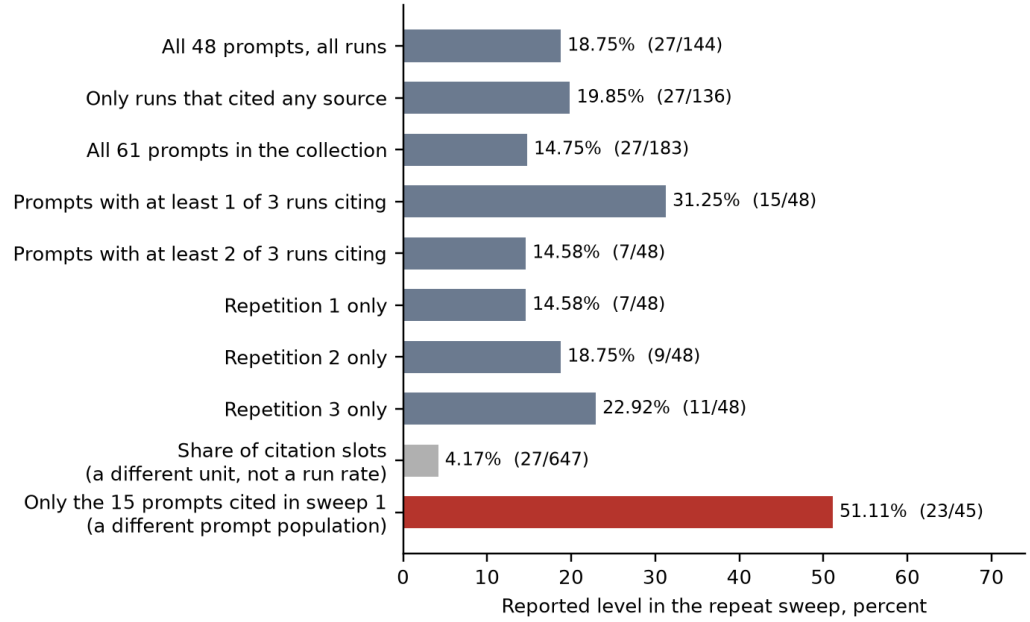

Figure 1. Reported level for Domain C in the repeat sweep under ten counting rules applied to the same saved ChatGPT-scraper answers. The blue bars are related incidence measures (run presence or prompt presence), not interchangeable estimates. The grey bar is a share of citation slots, a different unit, and the red bar is a rate on a selected prompt population (the 15 prompts cited in sweep 1). Neither is comparable with the blue bars. Source: supplementary tables T2 and T1.

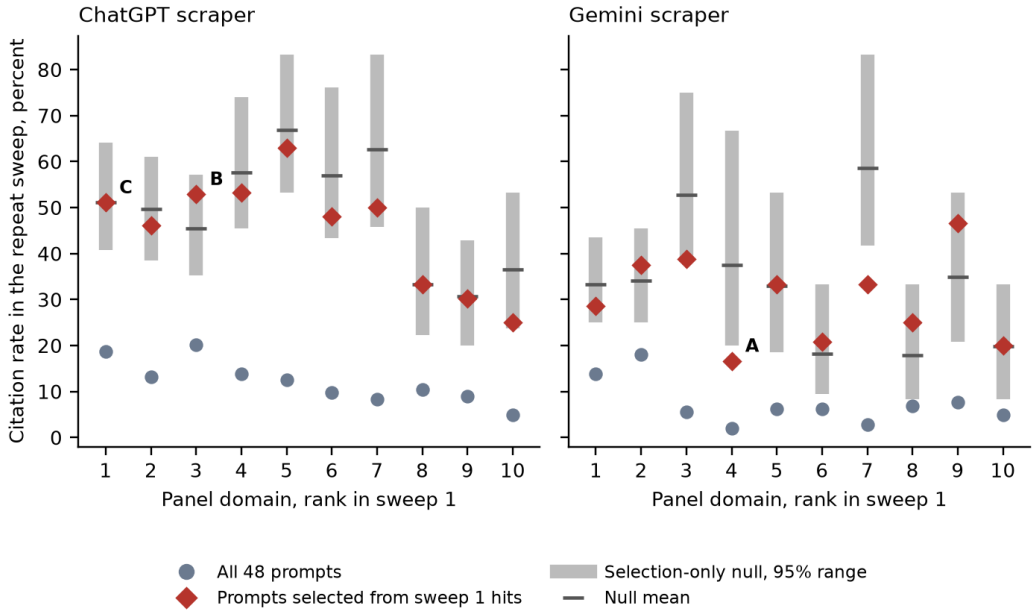

Figure 2. Repeat-sweep citation rate for each panel domain: over all 48 prompts (circles), over the prompts selected from sweep 1 citations (diamonds), and the range the selection-only null gives for the selected-prompt rate (bar, central 95%, mean marked). Letters mark Domains A, B and C. Source: supplementary table T1.

six runs, and the 20 pairs share the same 48 prompts, so they are not 20 independent tests. Selecting prompts on an earlier result raises the rate on those prompts as a matter of arithmetic, and the null shows that the size of the observed rise is within what that mechanism produces under the stated assumption. It does not establish that selection caused the difference, and it does not test whether a domain's citations persist on the selected prompts. The null keeps each prompt's tendency to cite the domain and removes any difference between sweeps; the selected prompts are the ones where the domain tends to be cited, in any sweep.

6.4 Direction of the change depends on the rule (RQ1)

For 10 of the 20 pairs, at least one of the eight related rules in rows 1 to 8 of Table 1 showed an increase of 1 percentage point or more between sweeps and at least one showed a decrease of 1 point or more. Domain B illustrates: the default rate rose by 4.86 points (22 to 29 of 144), but the prompt-reach rule fell by 2.09 points. Domain C illustrates the repetition-only rules: moving from sweep 1 to sweep 2 the first repetition's rate fell by 6.25 points, the second rose by 4.17 and the third did not change. Across the 20 pairs the three single-repetition changes spanned a median of 5.21 points (maximum 16.67), and for 8 of 20 pairs the single-repetition changes included both a rise and a fall of at least 1 point.

Denominators moved between the sweeps as well. The number of runs that cited any source fell from 139 to 136 of 144 for the ChatGPT scraper and from 118 to 117 for the Gemini scraper. Runs with no source list were 5 and 8 of 144 (ChatGPT) and 26 and 27 of 144 (Gemini). The mean number of distinct domains per run fell from 4.88 to 4.49 for ChatGPT and from 6.08 to 5.96 for Gemini, so the total number of citation slots fell by 8.0% for ChatGPT (703 to 647) and 1.9% for Gemini (875 to 858). A share of slots changes with that total even when a domain's runs do not.

6.5 Size of same-day moves relative to noise (RQ3)

Table 2. Three labeled domains, same prompts and settings, same day (144 runs per sweep).

Domain	Scraper	Sweep 1	Sweep 2	Change (pp)	Prompt-bootstrap 95% interval (pp)	Chance a domain chosen in advance moves at least this much
A	Gemini	10 / 144	3 / 144	−4.86	−9.72 to −1.39	2.6%
B	ChatGPT	22 / 144	29 / 144	+4.86	−1.39 to 11.81	26.0%
C	ChatGPT	28 / 144	27 / 144	−0.69	−7.64 to 6.25	100%

Domain C's 100% is arithmetic: it is cited in 55 runs over the two sweeps, an odd total that cannot split evenly, so every relabeling moves it by at least one run. Seven runs is 4.86 percentage points. Domain A fell by 70% and Domain B rose by about 32% in relative terms, but each is a move of seven runs.

The prompt-bootstrap standard error of a sweep-to-sweep difference was between 1.64 and 3.93 percentage points, giving an approximate 95% noise threshold of 3.22 to 7.71 points, with a median of 4.83 (ChatGPT) and 4.37 (Gemini). A true difference of that size would be detected about half the time. At 80% power the difference would need to be about 2.8 times the standard

error, a median of 6.90 points for ChatGPT and 6.24 for Gemini, and no panel domain's observed move reached its own 80% value. The design effect, the inflation of the variance of a sweep's rate relative to a binomial treatment of independent runs, had a median of 1.97 for ChatGPT (range 1.18 to 2.43) and 1.64 for Gemini (range 0.88 to 2.53). A confidence interval that treats the 144 runs as independent is therefore too narrow for most pairs.

The standardized differences (difference divided by the paired bootstrap standard error) had a standard deviation across the ten panel domains of 0.98 for ChatGPT and 1.17 for Gemini. Each is within sampling error of 1 (about 0.23 with ten domains), but the ten differences share prompts and are not independent, and selection on sweep 1 gives them a negative mean, so this is a rough match to the bootstrap standard error and not a demonstration of pure noise. Two of the 20 pairs had a bootstrap interval that excluded zero, both Gemini, and both ended within two runs of zero. For a domain chosen in advance, the two smallest probabilities of a move that large were 2.6% and 5.1%, against about 1 expected below 5% by chance among 20 pairs; the pairs share prompts and the panel domains were picked from sweep 1, so this count is descriptive. The sweep differences are consistent with repetition noise; differences much smaller than the 80% values above could not have been reliably detected.

Selection of the largest movers. The largest move among the ten selected domains was 7 runs for each scraper. Under whole-run relabeling a move that large or larger appeared in 62.5% of 20,000 draws (12,507) for ChatGPT and 52.2% (10,430) for Gemini. Cell-wise relabeling gave 62.3% and 53.5% over 20,000 draws. With 2,000 draws the whole-run figures were 62.1% and 52.4%. Because the observed statistic is the maximum of ten selected domains, it is a statistic that chance alone makes large.

A single domain chosen in advance behaves differently. The chance that such a domain moves by 7 runs or more ranged from 1.5% to 26.0% across the ten ChatGPT panel domains and from 0.0% to 25.8% across the ten Gemini panel domains (Figure 3). The earlier article's relabeling count applies to the first statistic, the largest of ten selected domains, and is not a threshold for one domain.

The mean signed move across the ten panel domains was -1.7 runs for ChatGPT and -1.0 for Gemini, which matches the drift toward the middle expected when domains are chosen for high sweep 1 counts. The relabeling null's mean was -0.33 runs for ChatGPT and -1.29 for Gemini; the observed values were not unusual under it (16% and 62% of null draws at or below them).

6.6 Where the variance sits (RQ4)

Most of the variance in a domain's citation outcome lay between prompts and between repetitions, and little between the two sweeps. For the 0/1 citation outcome of the ten ChatGPT panel domains, the median share of variance between prompts was 49%, between sweeps within a prompt 0%, and between repetitions within a sweep 49%. For the ten Gemini panel domains the medians were 28%, 4% and 68%. The sweep component was estimated at zero for seven of the ten ChatGPT domains and four of the ten Gemini domains (Figure 5). In these records, a repeat of the identical measurement added little systematic difference beyond what repetitions within one sweep already showed, and most of the variance came from repetition noise and from which prompts a domain is cited on.

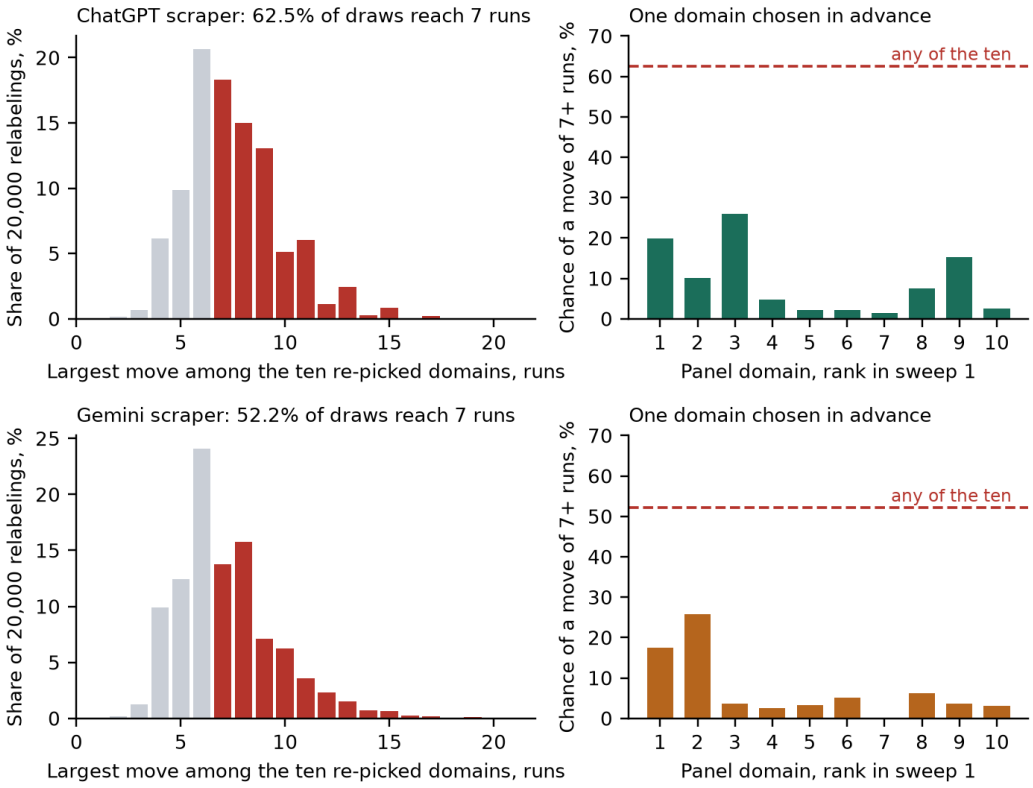

Figure 3. Left: distribution of the largest move among ten domains re-selected in each of 20,000 whole-run relabelings; draws at or above the observed 7 runs are shaded red. Right: for each panel domain, the chance in the same draws that a domain chosen in advance moves by 7 runs or more; the dashed line is the left-panel share for the largest of ten. Source: supplementary tables T6 and T6b.

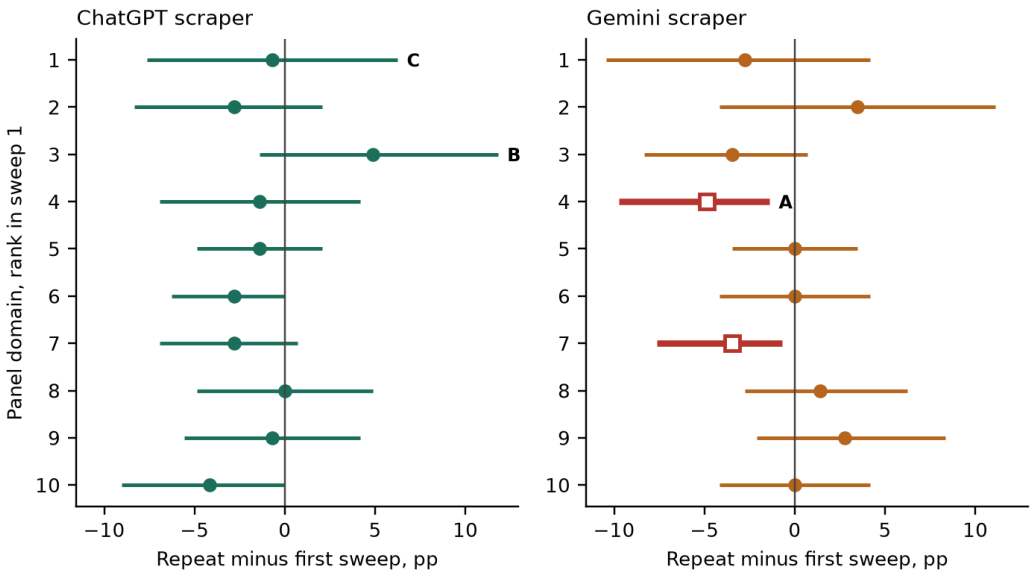

Figure 4. Change in citation rate between the first and the repeat sweep for each panel domain, with the paired prompt-bootstrap 95% interval. Red marks the two intervals that exclude zero. Letters mark Domains A, B and C. Source: supplementary table T4.

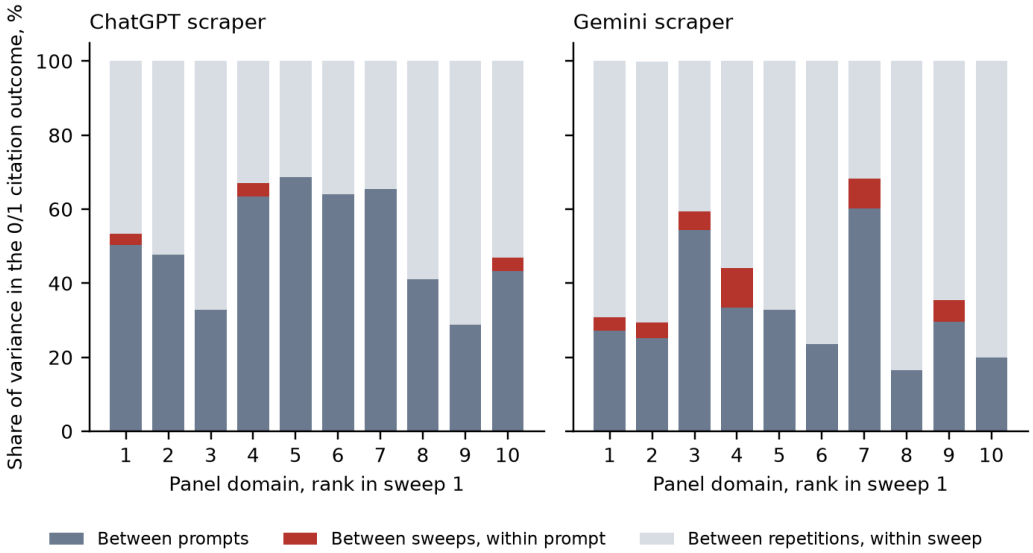

Figure 5. Share of variance in the 0/1 citation outcome by source, for each panel domain (method-of-moments estimates; negative estimates set to zero). Source: supplementary table T4b.

6.7 Leaderboards and cited sets

For the ChatGPT scraper, 9 of the ten most cited domains in sweep 1 were also in the top ten of sweep 2, and the leading domain changed: Domain C led sweep 1 with 28 runs, and a different domain led sweep 2 with 29 runs against 27 for Domain C. For the Gemini scraper 6 of 10 were shared under the panel file's order for the tied tenth place (5 to 8 under other resolutions), and the leader changed: the sweep 1 leader had 24 runs and a different domain led sweep 2 with 26. The tenth place was tied in ChatGPT sweep 2 and in both Gemini sweeps, so top-ten membership at the cutoff is arbitrary; counting every domain with at least the tenth-place count, the sweeps' sets had Jaccard overlaps of 0.75 (ChatGPT) and 0.50 (Gemini). The Spearman correlation of the two sweeps' run counts over the union of the two top-ten lists was 0.89 for ChatGPT and 0.11 for Gemini.

At prompt level, the mean sweep-to-sweep Jaccard of cited sets was 0.358 for the ChatGPT scraper (48 prompts) and 0.232 for the Gemini scraper (43 prompts with a cited set in at least one sweep). The mean repetition-to-repetition Jaccard within a sweep was 0.270 and 0.159. Sweep-to-sweep Jaccard was below 0.5 on 36 of 48 and 41 of 43 prompts. Prompts with larger cited-set unions had lower Jaccard values (Spearman correlation of -0.37 for each scraper), so a Jaccard figure is not comparable between sets of different size, which agrees in direction with the dependence of Jaccard similarity on set length that Sivakumar (2026) reports.

7 Reporting checklist

These are proposals. Each follows from a result above and none is a standard.

1. **Publish the numerator and the denominator, not only the percentage.** One domain read 14.6% to 31.3% under related run-presence and prompt-level rules in the same repeat sweep, 51.1% on the prompts that had cited it earlier, and 4.2% as a share of citation slots, a different unit (Table 1).
2. **State the counting rule beside every figure.** Run-presence, prompt-reach and majority rules gave different levels, and for 10 of 20 pairs different directions of change (Section 6.4).
3. **Say whether the prompt set was chosen using earlier results.** The selected-prompt rate was above the default rate for all 20 pairs, by a median factor of 4.0, and under the specified selection-only null 18 of 20 observed rates fell within the central 95% simulation range (Section 6.3).
4. **When the prompt set must change, rescore the old sweep under both sets and show the step.** The mixture identity shows what the change contributes with no new answer.
5. **Report the repetition count and the repetitions used.** Single-repetition changes for one domain spanned a median of 5.21 points (Section 6.4).
6. **Separate runs that returned no source list from runs that returned a list without the domain.** Ungrounded runs were 5 to 27 of 144 in the study subset, and counts moved between sweeps (Section 6.4).
7. **Resample prompts, not runs, for intervals.** Treating 144 runs as independent understated the variance by a median factor of 1.64 to 1.97 (Section 6.5).

8. **Compare a move with a noise threshold measured on the same design.** The approximate 95% noise threshold at 144 runs per sweep had a median of about 4.4 to 4.8 percentage points, and observed moves were mostly below it (Section 6.5).
9. **State whether the domains shown were picked after seeing the first result, and judge their moves against a null that repeats the picking.** The largest of ten selected moves reached 7 runs in 52% to 63% of relabelings, against 0% to 26% for one domain chosen in advance (Section 6.5).
10. **Log the reported model and the collector settings for every run.** The Gemini scraper reported a different model on 3 of 366 runs (Section 3).
11. **Report cited-set overlap with the set sizes.** Jaccard values fell as union size rose (Section 6.7).
12. **Keep the old series when the specification changes, and label the new one.**

8 Limitations

The records come from one project's prompts. Forty-eight informational prompts on AI visibility, search equity and related topics are not a sample of buyer questions or of any population of prompts, and the domains cited here are the ones these prompts happen to attract.

There is one date, one repeat and two scrapers. The two sweeps were 2 to 3.5 hours apart. Equal state of the world is assumed, not shown. The variance component between sweeps rests on two sweeps and cannot separate real change during the day from repetition noise. A failure to detect a sweep difference is not evidence that none occurred.

The scrapers are not the consumer products. They differ from the apps in personalization and location, and they were reached through one vendor. Nothing here covers Perplexity, Claude, Copilot or Google AI Overviews as products.

Domains are hosts with "www." removed, not registrable domains. Reducing to registrable domains would merge some subdomains and change some counts.

The selection-only null assumes the six runs of a prompt are exchangeable across the two sweeps. That is the assumption of no sweep difference, which is the null being tested against. Under it a prompt's tendency to cite a domain is kept fixed, so it does not model a change in that tendency.

The panel was selected on sweep 1 and has ties at its cutoff for Gemini. Prompt-bootstrap intervals use 48 prompts; they are unlikely to be well calibrated for domains cited on few prompts. Twenty intervals and many rules were examined without multiplicity adjustment. The one-percentage-point threshold in Section 5.3 was set after the data were seen.

The checklist items are proposals derived from one design. The paper makes no claim about how often any tracker or vendor changes its prompt set, no claim about any named vendor, and no causal claim about what changes AI citations.

9 Conclusion

In saved scraper runs from one project, one domain's repeat-sweep level ranged from 14.6% to 31.3% across eight related run-presence and prompt-level rules. Restricting to the prompts

that had already cited it gave 51.1% (23 of 45 runs), a rate on a different prompt population, and a share of citation slots gave 4.2% (27 of 647), a different unit; neither is a comparable visibility rate. The restricted rate was a median 4.0 times the default rate. Under the specified selection-only null, 18 of 20 observed selected-prompt rates fell within the central 95% simulation range, which is a compatibility result under an exchangeability assumption and not evidence about cause. In a same-day repeat of the identical measurement, the direction of change for a domain depended on the rule for half of the pairs, the largest of ten selected moves reached 7 runs in about half or more of relabelings, and most of the variance lay between repetitions and between prompts, not between sweeps. These results describe one date and one design and are a methods demonstration. They support a short list of reporting practices: publish the counts, state the counting rule and the selection history of the prompts and domains, and judge movements against a noise floor measured on the same design.

10 Statements

Data availability. The analysis scripts, the aggregate output tables, the five figures, the input hash files, the number ledger and the manuscript source are in a data package with a SHA-256 list. The package is deposited on Zenodo under CC BY 4.0 (https://doi.org/10.5281/zenodo.24163756). The raw scraper responses are third-party output and are not included, pending a rights check, so the scripts cannot be re-run from the package alone; the output tables they produced are included and can be checked against the numbers in this paper. The supplementary tables name the cited domains. The text of the paper labels three of them Domain A, B and C.

Competing interests. The author is the founder of Caldrin, a company that sells search and AI-visibility services. Caldrin carried out and paid for the data collection and the analysis. Domains cited in the saved answers include sites of companies that sell tools in the same field; they are named only in the supplementary tables and no claim is made about any of them.

Funding. No external funding.

Use of AI tools. An AI system, Claude (Anthropic), was used under the author's direction to write and run the analysis scripts reported here, to build the figures and consistency checks, to search for and read the literature, and to draft and edit the text. The new analysis code and the text of this paper were produced by the model claude-sonnet-5-5 on 12 November 2026. The data collection scripts and the earlier reproduction script were produced with Claude in earlier sessions of the same project; their model identifiers are not recorded in the files. The AI system is not an author. The author is responsible for the content.

Ethics. No human participants and no personal data. The data are answers returned by commercial scraper products.

Author contributions. Ankur Shrestha: conceptualization, data curation, and supervision of AI-assisted analysis and writing.

References

Bal, A. (2026). *Who The Machine Names: A Four-Week Panel Audit of AI Citation in Enterprise Banking Software.* SSRN. https://doi.org/10.2139/ssrn.7268163 (abstract only).

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1), 1-26.

https://doi.org/10.1214/aos/1176344552

Kunal, R. (2026). *Who Does AI Cite? Reconciling Conflicting Evidence on the Sources Behind Generative Search Answers (2024-2026).* SSRN. https://doi.org/10.2139/ssrn.7528060 (full text read in the author's version 1.0 deposit, 25 September 2026, https://doi.org/10.5281/zenodo.22979108).

Parrott, J. (2026, April 13; updated 2026-11-11). *AI Visibility Score Accuracy: Why One Number Changes Between Runs.* AuthorityTech blog (vendor article; not peer reviewed). https://authoritytech.io/blog/ai-visibility-score-accuracy-measurement-reality-2026 (accessed 2026-11-12).

Schulte, J., Bleeker, M., & Kaufmann, P. (2026). *Don't Measure Once: Measuring Visibility in AI Search (GEO).* arXiv:2604.07585 (v1, 8 April 2026). https://arxiv.org/abs/2604.07585

Shrestha, A. (2026, November 6). *Your AI Visibility Trend Line Might Be an Artefact.* Caldrin (company website; web article). https://caldrin.co/research/ai-visibility-trend-line-artefact (accessed 2026-11-12).

Sielinski, R. (2026). *Quantifying Uncertainty in AI Visibility: A Statistical Framework for Generative Search Measurement.* arXiv:2603.08924 (v1 9 March 2026; v3 26 August 2026, the version cited). https://arxiv.org/abs/2603.08924

Sivakumar, S. (2026). *Divergence Survives a Length Control on Google AI Overviews and Vanishes on Claude. Measuring Vendor-Set Agreement Across Four AI Search Engines.* SSRN. https://doi.org/10.2139/ssrn.7234418 (version 3.0, 4 August 2026, read in the author's deposit, https://doi.org/10.5281/zenodo.21789120; operator-run, not peer reviewed).

Żatuchin, D. (2026). *Query Language Sensitivity in Automated Sentiment Scores of LLM Responses.* arXiv:2607.13304 (version 2, 1 November 2026; version 1 of 14 July 2026 appeared under an earlier title and its results were withdrawn in version 2). https://arxiv.org/abs/2607.13304

A Prior publication disclosure

A shorter account of part of this dataset appeared on 6 November 2026 on the author's company website as a web article, "Your AI Visibility Trend Line Might Be an Artefact", by Ankur Shrestha (Shrestha, 2026; https://caldrin.co/research/ai-visibility-trend-line-artefact). It is a first-party publication of part of the same data, so it is not independent validation of the results here. It contains the counts for Domains A, B and C, the 27-of-144 and 23-of-45 recount, a relabeling count for the largest of ten selected moves, and a reporting section in prose. SHA-256 hashes of the article source file and of its page text as extracted on 12 November 2026 are in `paper/ARTICLE-HASHES.txt` in the data package.

What is new in this paper: the extension from three labeled domains to all 20 domain and scraper pairs; the ten-rule table and the direction-of-change comparison; the selection-only null for the selected-prompt rate and its comparison across 20 pairs; the mixture identity applied to all pairs; whole-run relabeling and the probability of a move for a domain chosen in advance; prompt-clustered intervals and design effects; the variance decomposition; the leaderboard and cited-set stability results; and the checklist with its links to results. What is carried over from the article: the reproduced counts in Section 6.1 and the definition of the selected-prompt recount.

B Reproducibility

Inputs. The saved raw scraper responses for the two sweeps (366 raw response files per sweep, 183 per scraper, 732 in all), the two run indexes, the prompt file, and the reference-domain panel file. SHA-256 hashes of every input file are in `analysis/INPUT-`

`HASHES.csv`. The tree hashes of the raw response sets are in `analysis/INPUT-TREE-HASHES.json` (sweep 1: d8c62ff82f83f0e52604f46b4caffeb9a6676d0c7c88cd9436a4e320c5db957c; sweep 2: 99fafbf9c55461588beb30eb56cb04338873ae57ef7a9bc13fd3adb173cdc166). Raw responses are third-party scraper output and are released only if a rights check allows; see Data availability.

Scripts (Python 3.12, NumPy 2.5; figures use Matplotlib 3.11):

- `analysis/p2_hashes.py`: hashes the inputs.
- `analysis/p2_analyses.py`: all tables and the headline summary. Run as `python3 -I p2_analyses.py`. Deterministic: a second run gave identical output.
- `analysis/p2_figures.py`: Figures 1 to 5 from the tables.
- The earlier reproduction script `a08_reproduce.py` (with its two helper modules, in the data package under `a08_reproduction/`) was re-run from the raw files into a separate folder and its results compared with the stored results; both outputs were identical.

Seeds. All random streams are NumPy default generators. Base seed 20261112. Selection-only null for the selected-prompt rate: 20261113. Paired prompt bootstrap: 20261116 (20,000 resamples). Largest-move null, whole-run relabeling: ChatGPT 20261212 (2,000 draws) and 20261213 (20,000); Gemini 20261312 and 20261313. Cell-wise relabeling: ChatGPT 20261222 and 20261223; Gemini 20261322 and 20261323.

Second implementation. A second script, `verification/indep_check.py`, was written separately by an AI agent from the text of this paper, without importing the analysis script, and re-derived every result from the saved raw responses with its own seed. It agreed with the tables on 1,756 of 1,759 compared values; the three flags were a comparison-format artefact and one count at the 5% boundary (`verification/VERIFICATION-REPORT.md`). Its comparison led to a rerun of the paired bootstrap at 20,000 resamples in place of 2,000, which moved five of 40 interval limits by one run. It is a second AI-written implementation, not a human or independent-laboratory replication.

Output tables (`analysis/`): T0 data description and collection windows; T1 matched recount and selection-only null; T1b mixture identity; T2 alternative counting rules; T4 bootstrap and design effects (the field `crit_diff_pp` in T4 and HEADLINE.json is the approximate 95% noise threshold of Section 5.4, 1.96 times the paired prompt-bootstrap standard error, and `power80_diff_pp` is 2.8 times that standard error; the field names predate the wording of the paper and the calculations are unchanged); T4b variance components; T5 and T5b leaderboard and prompt-level Jaccard; T6 and T6b largest-move null and single-domain null; HEADLINE.json for the summary values quoted in the text.