Query Competitiveness and Citation Source Composition in Google's AI Overviews: Evidence From a Repeated-Measures Ecommerce Query Ladder

Abdullah Nouman

WpConsults, Sheridan, WY, United States

Author Note

Abdullah Nouman https://orcid.org/0009-0003-9071-484X

This research was designed, funded, conducted, and approved internally by WpConsults, with no external funding and no external peer review. In place of external review, the method and both falsifiable nulls were fixed before the first capture and both nulls are published as results; every counted row had to meet a stated verification standard, with sampled rows checked by eye against stored screenshots in every round; and the raw data, capture frames and analysis code are deposited with the paper, as set out in the Data Availability Statement. The author owns and is chief executive of WpConsults, a marketing consultancy selling search and AI-visibility services, and the research concerns that same domain. No client data was used, and by design the study did not track the author's own website, so no result here concerns a property the author owns.

Correspondence concerning this article should be addressed to Abdullah Nouman, WpConsults, Sheridan, WY, United States. Email: abdullah@wpconsults.com

Abstract

Practitioner guidance holds that small operators should abandon competitive head terms for the long tail, assuming the AI Overview citation pool opens up as a query becomes less competitive. We tested it on a designed sample: a 21-query ecommerce ladder banded HEAD, MID, and TAIL, containing three matched head-to-tail pairs, captured in four English-language markets across six rounds, signed out, with the sources panel scrolled to exhaustion. One query in one market in one round is a cell; the analyzable sample was 503 usable cells of 504 designed. No Overview-free result page was found among them, every apparent absence resolving to a blocked or faulty instrument, falsifying the pre-declared null that the feature would not fire on long-tail or frontier ecommerce queries. Cited-pool size did not track competitiveness: of three matched pairs, one widened on the tail in all six rounds, one narrowed in all six, one closed at zero, so the pre-declared no-gradient null is published as a result. Citation source composition did move. Scoring 200 of 667 cited domains for domain authority (83.8% of citations) and bracketing the unscored remainder at both extremes, small-site citation share rose from head to tail by 16 to 25 points at every threshold tested. Decomposition under a separate curated classifier, in two rounds, showed mega-platform share approximately flat while big-brand share fell from 40% of head citations to 12 to 14% on the tail. The defensible conclusion is that big brands stop competing on the tail, not that small sites win more.

Keywords: AI Overviews, generative engine optimization, generative search, citation analysis, retrieval, extractability, corroboration, search engine optimization, ecommerce, long tail, repeated measures

Query Competitiveness and Citation Source Composition in Google's AI Overviews: Evidence From a Repeated-Measures Ecommerce Query Ladder

Google's AI Overview places a generated answer above the organic results and attributes it to a set of cited sources. Those sources are listed in an expandable panel beside the answer, carrying a small count label such as "6 sites"; throughout this paper we call that panel the **sources panel** and the label the **source-count chip**, and a capture counts only if the panel was opened and scrolled to its end. For any site owner, the practical question is no longer only whether a page ranks, but whether it enters that cited set. The practice built around that question now travels under the name **generative engine optimization**, and the question this paper asks sits at its center: what decides which sources a generative answer layer retrieves and cites. A widely repeated piece of practitioner advice answers this by direction of travel: do not fight for competitive head terms, go long-tail, because the citation pool is more open down there.

That advice is plausible and, as far as we can establish, untested on a sample designed to test it. Published citation studies in this space are large-sample correlational aggregates, and they are mostly published by companies selling AI-visibility products. Their samples skew toward high-volume commercial queries and toward a single market, which is exactly the region where a competitiveness gradient cannot be observed.

This paper asks a narrower question that a designed sample can answer. Holding vertical, capture protocol, and instrument constant, does the **composition** of the cited set change as query competitiveness falls, and does the **size** of the cited pool change with it? We treat those as two separate measures, because conflating them is how the folk claim survives.

We also ask a free sub-question that the design answers at no extra cost. Does an AI Overview fire at all on near-zero-volume and frontier ecommerce queries? A negative answer would be data in its own right, and we declared in advance that we would publish it either way.

Two commitments shape everything that follows. The method, the measures and the two nulls were written before the first capture, and are reproduced in the Method section as they were written. And where our own earlier readings were wrong, including a size classifier we re-derived three times before freezing it, we report the correction rather than the tidy final number.

Prior Work and the Gap This Paper Takes

We stand on prior work for three points and do not re-derive them. Domain-level citation share concentrates on a small number of large platforms, with community and other user-generated sites at the top of the distribution, and that concentration is unstable: across thirteen weekly snapshots of more than 230,000 prompts, Reddit's share of ChatGPT responses fell from close to 60% to roughly 10% in a matter of weeks (Semrush, 2025). Overview prevalence itself varies sharply when measured at scale: by industry across 118 million tracked keywords (Conductor, 2025), and by query intent and industry across a 10,000-keyword United States sample (Authoritas, 2025). And citation carries measurable consequence: pages cited inside an Overview earn about 120% more organic clicks per impression than uncited pages on the same Overview-present result page, while still underperforming pages on Overview-free result pages, a comparison their study scopes to informational queries (Seer Interactive, 2026).

These are the observations we build on, and they are consistent with our own earlier first-party work in a different vertical (WpConsults Research Lab, 2026b).

We treat these organizations as **sources, not as facts**, and the distinction is load-bearing here. Each sells a product whose value proposition depends on AI visibility being both

measurable and improvable by the buyer, which is a structural reason to read their effect sizes with more care than their directions. None of the four publishes the underlying row-level data, so none of the figures above can be independently re-derived, and the Authoritas whitepaper is gated, so we verified its stated scope rather than its contents. We cite them for the observations they made first, and we disagree with the inference most commonly drawn from them below.

The inference we dispute is the one that runs from "large platforms take a large share of citations" to "the cited pool is closed to small operators." Our data does not support that step in this vertical.

Concentration measured as the top-five platforms' share of citation instances stayed between 23% and 32% across our six rounds, never above one third, and the number of distinct cited hosts per round ranged from 200 to 332 across 503 cells. A pool can be dominated at the top of the distribution and still be wide at the bottom, and the aggregate studies are not designed to show the second thing.

There is a second and more specific gap. Vendor samples are head-term-heavy because head terms are what their customers buy tools to track, so the competitiveness gradient itself is largely unobserved in the published record. That gap, within one vertical, across four markets, read to exhaustion and captured on a repeated-measures schedule, is what this paper takes.

We also position this paper against our own limits. This is one operator, one instrument, one vertical and one 17-day window, so it is not a population estimate of anything. It is a designed field study whose contribution is the design and the published nulls, not the sample size.

Method

The method below was fixed before the first capture. Deviations from the plan are stated in this section.

Design

A repeated-measures observational design over a fixed query ladder. The primary independent variable is **query competitiveness**, operationalized as a band (HEAD, MID, TAIL) assigned before capture from Ubersuggest United States search volume and SEO difficulty. A secondary intent tag (trust, commercial, informational, transactional, operator pain, our lane, frontier) was recorded per query.

The ladder contains **three matched head-to-tail pairs**, each holding topic constant while varying specificity, so the gradient is measurable on the same underlying question and not only in aggregate.

Four markets were captured for every query in every round: the United States as anchor, plus Canada, the United Kingdom and Australia. Markets were never averaged or merged. One row equals one query, in one market, in one round.

Sample

Twenty-one queries. Volume and difficulty figures are Ubersuggest United States (locId 2840), locked 2026-07-17; the four hyper-specific tail queries and the frontier queries read at roughly zero volume and difficulty about 4.

Table 1

The 21-Query Ecommerce Ladder, Banded by Competitiveness

#	Query	Vol / SD	Intent	Band
1	is temu legit	40,500 / 33	trust	HEAD

#	Query	Vol / SD	Intent	Band
2	is shein legit	22,200 / 34	trust	HEAD
3	is aliexpress safe	8,100 / 44	trust	HEAD
4	how to sell on amazon	22,200 / 40	transactional	HEAD
5	is dropshipping worth it	1,900 / 39	informational	MID
6	is dropshipping dead	390 / 29	informational	MID
7	what is agentic commerce	1,000 / 53	informational	MID
8	best ecommerce platform	1,600 / 58	commercial	HEAD (Pair A head)
9	best ecommerce platform for a handmade candle business	~0 / 4	commercial	TAIL (Pair A tail)
10	shopify vs woocommerce	590 / 34	commercial	MID (Pair B head)
11	shopify vs woocommerce for a small uk clothing brand	~0 / 4	commercial	TAIL (Pair B tail)
12	how to reduce cart abandonment	110 / 38	informational	MID (Pair C head)
13	how to reduce cart abandonment on a one-product store	~0 / 4	informational	TAIL (Pair C tail)
14	how to get products in google shopping for free	0 / 4	transactional	TAIL
15	why is my shopify store not getting sales	10 / 31	operator pain	TAIL
16	is etsy still worth it for sellers	0 / 4	informational	TAIL
17	shopify vs etsy for handmade sellers	0 / 4	commercial	TAIL
18	do i need llms.txt for my shopify store	~0 / 4	our lane	TAIL
19	how to get my products recommended by chatgpt	0 / 4	our lane	TAIL
20	can chatgpt buy products for me	0 / 4	frontier	TAIL
21	will ai overviews hurt ecommerce	0 / 4	our lane	TAIL

Band sizes are therefore 5 HEAD queries, 5 MID queries and 11 TAIL queries, giving 20, 20 and 44 cells per round respectively.

One point of transparency about the pairs. Two of the three pair heads (queries 10 and 12) sit in the MID band rather than HEAD, because banding was assigned on absolute volume

and difficulty rather than on a query's role in a pair. The pair analysis and the band analysis are therefore two different cuts of the same data, and we do not present a pair result as a band result.

Capture Protocol

All captures were taken through one instrument, a scripted real-browser capture tool (named `study`), signed out, with the following conditions required on every counted row: the AI Overview expanded; the sources panel scrolled to exhaustion (`panelSettled: true`); the market confirmed against Google's own page footer (`countryConfirmed: true`); and the organic top 10 read with sponsored results excluded. Per capture we logged AI Overview fired yes or no, the source-count chip, the full cited host list, the organic top 10, the band and the intent tag.

Cadence was twice weekly. After Round 1 tripped Google's anti-bot protection (see Limitations), each capture day was split into two paced slots at 05:20 and 18:20 Asia/Dhaka, with queries 1 to 11 in the morning slot and 12 to 21 in the evening slot; both slots of the same calendar date constitute one round. Inter-request delays were jittered between roughly 8.7 and 11.8 seconds, with 15 to 30 second cooldowns after each query's four-market sweep and a longer cooldown of about two to three minutes after roughly every five queries.

Splitting a round across two same-day slots about 13 hours apart distributes the within-round observation across that window. The split was held identical in every round, so it is constant across the design.

Verification, Usability, and the Treatment of Failures

A cited list was counted **only** from rows flagged `usable: true`, meaning the Overview was expanded and the panel scrolled to exhaustion. A partially read panel is a lower bound, not a count, and no lower bound was ever counted.

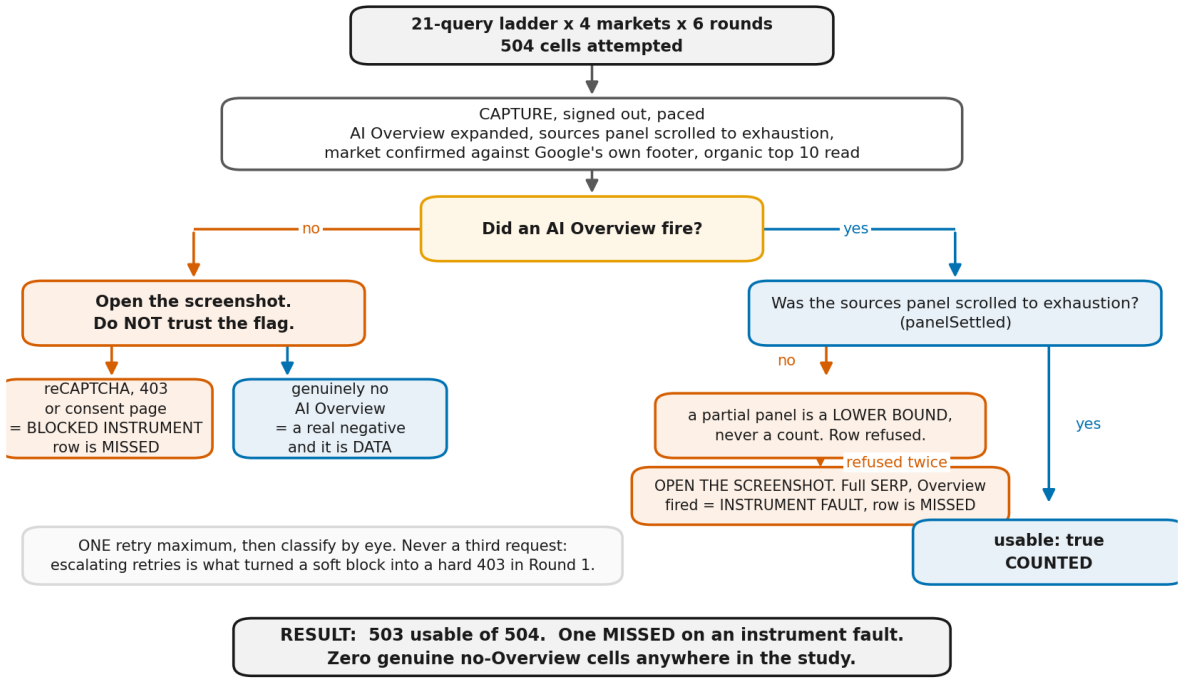
Failure handling followed one hard rule fixed after Round 1: at most one retry per failed cell, then open the screenshot and classify by eye. A reCAPTCHA, a 403 page or a consent interstitial is a **blocked instrument** and the cell is MISSED. A genuinely Overview-free SERP is a real negative and is data.

Nothing was ever requested a third time, because escalating retries is what converted a soft block into a hard 403 in Round 1.

Each round also carried a positive screenshot check: at least one row per slot was opened and matched by eye against its recorded cited set, chip count, sign-out state and footer. The same tool's cited-array parser was observed contradicting its own screenshots in a separate study on 2026-07-29; the Limitations section states which of this paper's results are exposed to that fault.

Figure 1

How a Blocked Instrument Is Kept From Becoming a Finding



Note. N = 504 designed cells. The single MISSED cell (`how to get products in google shopping for free`, United States, Round 6) took the right-hand path: the Overview fired on both reads but the sources panel refused to open, so after the one permitted retry the frame was opened by eye. The 734 KB screenshot shows a full signed-out US SERP with the Overview fired and expanded, no reCAPTCHA, no 403, no consent interstitial: an instrument fault, not a block and not a no-Overview. Counting it as an absent Overview would have manufactured the study's only negative, which is precisely the error the eye-check step exists to prevent.

Pre-Declared Measures

Eleven measures were declared before capture: (1) platform citation share with an explicit user-generated-content read; (2) platform share by market; (3) consistency and churn across rounds; (4) citation versus organic rank in both directions; (5) cited-domain size buckets and small-site share head to tail; (6) the popularity gradient including the three matched-pair

deltas; (7) the intent split; (8) the stable core of most consistently cited domains; (9) AI Overview fire rate; (10) cited-pool size and concentration; (11) a yes or no flag for the presence of a Shopping or product module on the SERP.

Our own site's presence was deliberately not tracked in this study, so nothing in this paper concerns the author's property.

Pre-Declared Nulls

Two nulls were written before the data arrived, with a commitment to publish either as a result.

We use "null" here in the sense of a pre-declared outcome we committed to publish whichever way it fell. The first is also a null hypothesis in the strict sense, since it states no effect of competitiveness. The second is a prediction of absence, so publishing it had it held would have been a null result rather than the rejection of a null hypothesis.

Null 1, "no gradient": the cited pool is the same on head and tail.

The null as written did not name a measure. We report it against both pre-declared measures, cited-pool size and cited-source composition, and it falls differently on each. Splitting it this way is a choice made after the data, and a reader who prefers to read the null as a single statement should treat it as not surviving. Our reason for the split is that size and composition were pre-declared as separate measures in the study design, and the Introduction states the distinction as a design commitment.

Null 2, "no fire": the AI Overview does not fire on long-tail or frontier ecommerce queries.

Both are reported in the Results section. One survived on one measure and one was falsified, and we report both outcomes at the same length.

Size Classification and the Sensitivity Analysis

Measure 5 requires sorting every cited domain into mega-platform, big brand or media, and small independent.

The classifier was re-derived by hand at Rounds 2 and 3, and the head-to-tail number swung between versions.

At Round 4 the rule was corrected (about 24 consumer security and review brands, including Norton, Avast, ESET, McAfee, Kaspersky, Bitdefender, NordVPN, Surfshark, Trustpilot, Which, Consumer Reports and NerdWallet, were moved from "small independent" to "big brand"), applied retroactively to all rounds, and **frozen in a script** so it could not drift again. Those domains cluster almost entirely on the three trust queries, which all sit in the HEAD band, so their placement moved the head number by more than 20 points on its own.

Under the immediately preceding rule the gradient was absent, and this must be stated plainly. Classifier v3 gave HEAD/MID/TAIL small-site shares of 63/53/64 in Round 1, 61/54/66 in Round 2 and 61/54/60 in Round 3, head-to-tail lifts of roughly +1, +5 and -1. These figures are quoted from the study log; the v3 rule was never frozen in code, so they cannot be recomputed from the published scripts. The Round 3 log calls that result essentially flat. The v4 rule was derived at Round 4, applied retroactively, and the same three rounds became +23, +23 and +20. A rule adopted after observing an absence, which then produces the effect, is the kind of decision a reader is entitled to weigh directly. Our case for it is that the reassignment was independently vindicated: Norton scores 91, Trustpilot 93, Avast 90, Which 87, Bitdefender 86 and NordVPN 84 on the authority measure, so treating them as large brands was a correction. The reader can weigh that case against the v3 series above.

A curated list moved by hand is not a publishable basis for a size claim, however carefully curated. So before any size number was written, every cited host was scored against **real domain-authority values** and the curated list was replaced by a numeric threshold.

The authority figures are Ubersuggest domain authority scores, read through the same connected API that supplied the volume and difficulty figures in Table 1, via its backlink-overview endpoint. The score runs 1 to 100, and across the 200 domains we scored it spans that full range with a median of 48.5. It is a vendor estimate of a domain's overall strength derived from its backlink profile, not a property of the domain, and we treat it as we treat the vendor literature: it is used because it is reproducible and independent of our judgment, not because it is true. Any authority metric would be a decision, which is why three thresholds are reported rather than one.

Coverage is partial and we state it exactly. **200 of 667 distinct cited domains were scored, carrying 3,602 of 4,298 citations (83.8%).** The lane's citations are heavily concentrated, so 200 domains buy most of the mass; the unscored remainder is 16.2% of citations spread across 467 domains, nearly all cited only once or twice.

The two rules do not use the same unit of analysis, and that gap is material. The curated classifier operates on FULL HOSTS and forces subdomains with community, forum, support, help, answers, discussion, board or blog prefixes to small before any lookup, so `community.shopify.com` is small under it. The blog prefix catches corporate blogs too: `blog.avast.com` and `blog.google` count as small independents under the frozen rule, which is a misfiling, but one that runs against our claim (scoring them as large would lower the HEAD small share and widen the lift), so it is disclosed rather than repaired post hoc. The authority file is keyed on ROOT DOMAINS, so the same host folds into `shopify.com` at authority 95 and is

large under the numeric rule. Running both over all six rounds, the two rules disagree on **516 of 3,602 scored citation instances, 14.3%**, concentrated on `shopify.com` (119 instances), `fullstory.com` (37) and `printify.com` (28). Community and support sources also skew toward operator-pain and how-to queries, which sit MID and TAIL, so the disagreement is not evenly distributed across bands. The two rules agree on the direction of the head-to-tail result, which is what the sensitivity analysis is for; they are not, however, independent checks on one another.

Partial coverage is made publishable by a **two-way sensitivity analysis**. We compute the band split twice, once assuming every unscored domain is small and once assuming every unscored domain is large, so the truth is bracketed between the two.

If the conclusion holds at both extremes, the unscored remainder cannot overturn it, and we report the table rather than a single number.

Deviations From the Preregistered Plan

The study was designed for eight rounds and closed after **six**, on the principal investigator's instruction on 2026-08-02. This is a real limitation, not a rounding of the design, and it bears specifically on the churn and pool-contraction readings, which have the fewest observations. It is restated under Limitations.

Measure 11, the Shopping-module flag, was **never emitted by the capture harness in any of the six rounds**. The full row-key set was dumped and checked at close and contains no key matching shop, product or module.

Measure 11 is therefore reported as **NOT COLLECTED**. It is not a null, it was never observed, and nothing in this paper should be read as evidence about Shopping modules.

Figures

All figures use the Okabe and Ito (2008) colorblind-safe palette, with a second visual channel (dash style, marker shape, or hatch) in every multi-series chart, because those hues are engineered for colorblind distinguishability rather than for luminance separation and do merge in grayscale. Ordered categories are shaded on a single sequential ramp rather than given unrelated hues. Every figure caption states its N, and figures with different Ns are never merged onto shared axes.

Results

Every figure below carries its N. Percentages computed on 21 cells per market per round are directional and are labeled as such.

The AI Overview Fires Everywhere in This Vertical

No Overview-free result page exists anywhere in this dataset. The rate was 84 of 84 in each of Rounds 1 through 5 and 83 of 83 in Round 6, in all four markets, and it includes every roughly zero-volume tail query and every AI-shopping frontier query (the Round 6 tail and frontier subset alone was 43 of 43).

Pre-declared **Null 2 is falsified**.

One framing point matters more than the number. A row qualifies as usable only if the sources panel was scrolled to exhaustion, and a sources panel cannot exist on a page with no Overview, so **a usable row entails a fired Overview by construction**. The figure of 503 of 503 is therefore a consistency check on the pipeline, not an estimate of a rate. The claim that carries weight is the one stated over the **504 designed cells**: every apparent absence was chased to a cause, and each resolved to a blocked or faulty instrument rather than an Overview-free page.

One cell requires care and is the reason the denominator is 503 rather than 504. In Round 6, `how to get products in google shopping for free` in the United States returned `usable: false` on two separate reads: the Overview fired both times, the footer bound to the United States both times, the organic top 10 read, and the chip reported "6 sites" on the first read and "5 sites" on the second, but the sources panel never opened either time and the cited list came back empty. The harness correctly refused to record a lower bound.

We opened the 734 KB screenshot rather than infer. It shows a full, genuine, signed-out United States SERP with the Overview expanded and an inline sources card, with no reCAPTCHA, no 403, no consent interstitial and no near-blank frame.

That cell is therefore an **instrument fault**, logged as MISSED. It is explicitly **not a block and not a no-Overview**, and counting it as an absent Overview would have manufactured the only negative in the study.

The same query and market also required a retry in Round 5, which makes it the one cell in the ladder with a repeat instrument weakness, and we name it rather than smooth it.

The Pool Is Already Wide Open, Including at the Head

Table 2

Cited-Pool Width and Concentration by Round

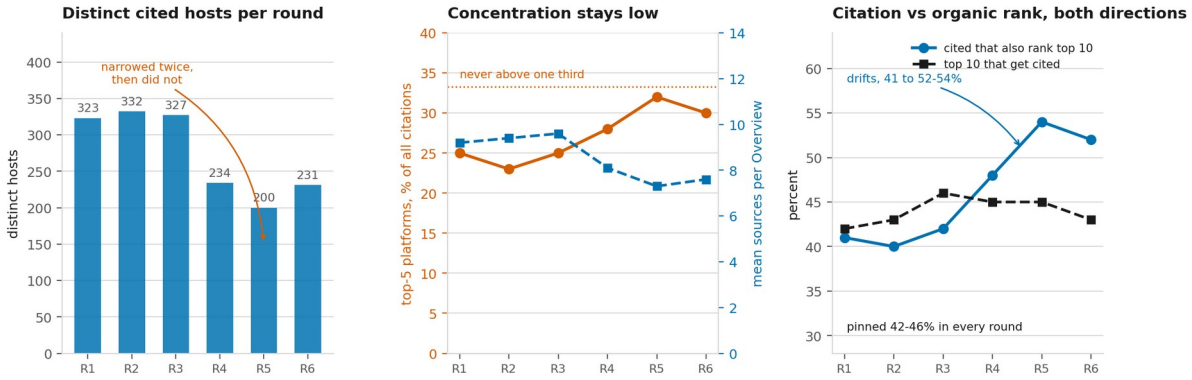
Round	Date	Distinct cited hosts	Mean cited per Overview	Top-5 share	Range
1	2026-07-17	323	9.2	25%	3 to 25
2	2026-07-19	332	9.4	23%	3 to 22
3	2026-07-22	327	9.6	25%	4 to 20
4	2026-07-26	234	8.1	28%	3 to 18
5	2026-07-29	200	7.3	32%	2 to 14

Round	Date	Distinct cited hosts	Mean cited per Overview	Top-5 share	Range
6	2026-08-02	231	7.6	30%	2 to 15, median 8

Note. N per round is 84 cells for Rounds 1 to 5 and 83 for Round 6; 503 cells and 4,298 citation instances in total. Top-5 share is the five most-cited ROOT platforms' share of citation instances in that round, recomputed from the raw rows under this one definition; the study log's original column mixed two computations (Rounds 1 to 3 ad hoc at host grain, Rounds 4 to 6 from a script line whose numerator counted cells rather than instances) and is corrected in the study record. The distinct-host column counts full hosts, of which there are 695 across all rounds; the 667 figure used throughout the domain-authority analysis counts root domains, into which those hosts fold. The Round 1 to 3 ranges were not recorded in the round log and are computed from the deposited raw rows.

Figure 2

The Pool Is Wide at Every Band, and One of the Two Rank Measures Moved



Note. N = 84 cells per round for Rounds 1 to 5 and 83 for Round 6, 503 in total. LEFT: the Round 4 and 5 narrowing did not continue into Round 6, so the study closes calling pool size a fluctuation rather than a trend, and the earlier reading is downgraded on the record. RIGHT: the share of cited sources that also rank in the organic top

10 climbs from 41% to 52-54%, while the reverse direction stays pinned at 42-46% in every round. Both directions are reported in Table 6.

Concentration is low in every round and at every band. A small operator is not structurally locked out at the head, which is the assumption the study was built to test and the first thing the data declines to support.

We also record and then withdraw an intermediate reading. At Rounds 4 and 5 we observed two consecutive contractions in pool size and flagged them as a possible trend. Round 6 did not continue the narrowing, so the study closes calling this **fluctuation within a band, not a trend**, and the earlier reading is downgraded here on the record rather than quietly dropped.

Citation Source Composition by Band

This is the headline measure, and it is reported as a table of bounds rather than as a number.

Table 3

Small-Site Share of Citations by Band Under a Two-Way Domain-Authority Sensitivity Analysis

DA threshold for "large"	Unscored assumed SMALL	Head-to-tail lift	Unscored assumed LARGE	Head-to-tail lift
DA >= 50	HEAD 23% / MID 39% / TAIL 47%	**+24 pts**	HEAD 10% / MID 28% / TAIL 27%	**+16 pts**
DA >= 60	HEAD 28% / MID 46% / TAIL 53%	**+25 pts**	HEAD 16% / MID 35% / TAIL 33%	**+17 pts**
DA >= 70	HEAD 33% / MID 49% / TAIL 57%	**+24 pts**	HEAD 20% / MID 37% / TAIL 37%	**+17 pts**

Note. Band shares and lifts are each rounded independently from unrounded values, so a lift will not always equal the difference of the two rounded cells beside it. The DA >= 50 worst-case row is the clearest case: 27 minus 10 reads as 17, while the unrounded difference rounds to 16.

The head-to-tail lift in small-site citation share is **positive at every threshold and under both extreme assumptions, ranging +16 to +25 percentage points**. The lift moves by at most one point across the three thresholds within each assumption, so the result does not depend on where the authority cut is drawn.

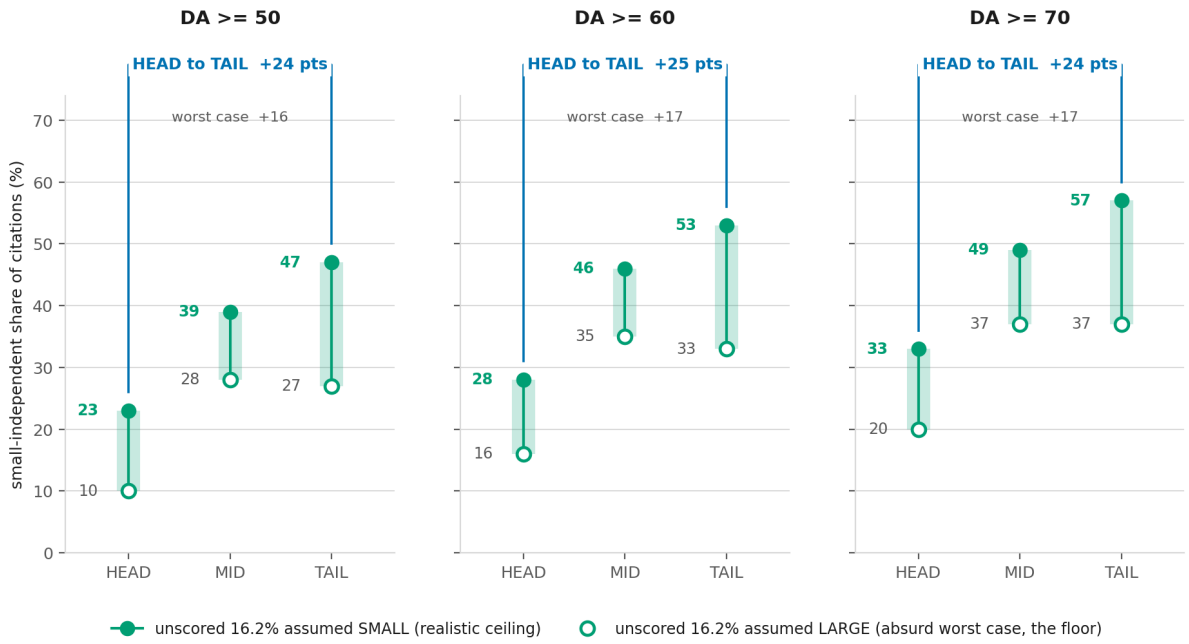
The gradient is therefore **not an artifact of the hand-curated list**, which was the single objection capable of killing this claim. It is, however, sensitive to the unit of analysis: the numeric rule brackets the lift at +16 to +25 while the curated rule places it at +20 to +28 across the six rounds, for the reason set out under Size Classification.

One caveat travels with this table and must not be dropped. In the absurd worst case, where every unscored domain is treated as a large brand, TAIL sits at or below MID at every threshold (28 versus 27, 35 versus 33, 37 versus 37). The clean monotonic HEAD < MID < TAIL ordering therefore **flattens between MID and TAIL** at the worst-case bound, even though head-to-tail stays clearly positive at every threshold.

We therefore claim HEAD versus TAIL only. This paper does not claim a smooth three-step gradient, and any reader who takes one from the realistic column is reading past the bound.

Figure 3

Small-Site Citation Share by Query Band, Reported as a Bracket Rather Than a Number



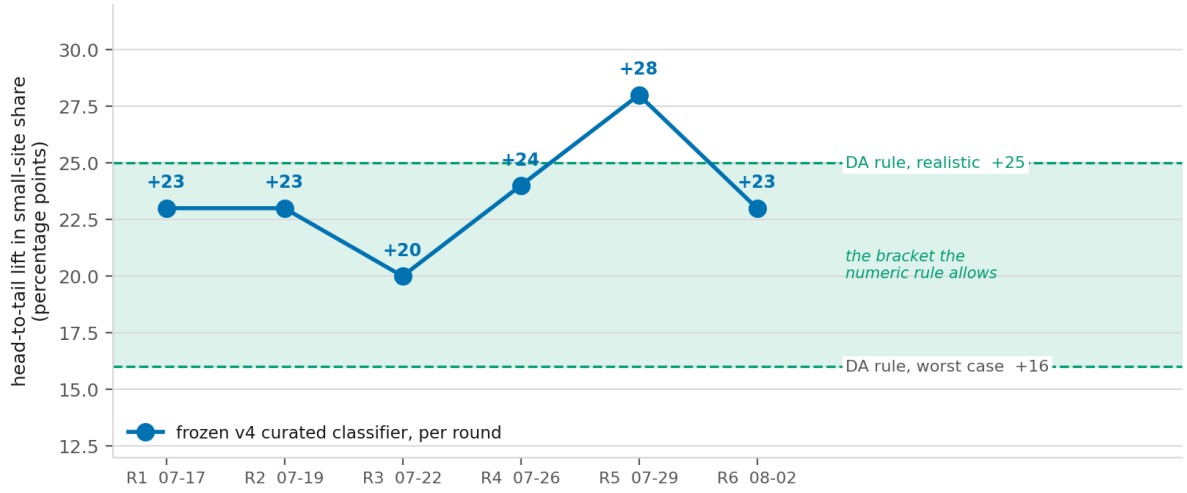
Note. N = 503 usable cells (21 queries x 4 markets x 6 rounds, minus 1 MISSED). 4,298 citations across 667 distinct domains; 200 domains scored for authority, covering 3,602 citations (83.8%). The true value lies somewhere inside each bar. Head to tail is positive at every threshold and under both assumptions, +16 to +25 points. MID is plotted but deliberately NOT connected to the end bands: at the floor TAIL sits at or below MID at every threshold (28 vs 27, 35 vs 33, 37 vs 37), so the paper claims HEAD versus TAIL only and never a smooth gradient. Reproduce with da-sensitivity.py.

The result is the most consistently reproduced in the study. Small-site share by band across all six rounds was 38/54/61, 41/53/64, 40/51/60, 34/52/58, 28/48/56 and 34/49/57 (HEAD/MID/TAIL), giving head-to-tail lifts of **23, 23, 20, 24, 28 and 23 points, positive in all six rounds**, with no exception. These are six repeated measurements of the same 21 queries by one operator on one instrument over 17 days, not six independent replications, and consecutive rounds are correlated by construction.

Two classification rules, one of them reported as a bracket, agree on the direction and rough magnitude: the frozen curated classifier gives a +23 point lift at Round 6 (34% / 49% / 57%), and the numeric authority rule brackets the same quantity between +25 and +17 at DA >= 60. We do not describe these as three independent methods. The two ends of the bracket are the two extremes of a single sensitivity analysis on a single scoring pass, and the two genuinely distinct rules are not independent of each other either, for the reason set out under Size Classification.

Figure 4

Two Classification Rules, One Direction and a Rule-Dependent Magnitude



Note. N = 503 cells. The curated classifier gives a head-to-tail lift of +23, +23, +20, +24, +28 and +23 points, reproduced in all six rounds of the same repeated-measures ladder rather than in six independent samples. The numeric authority rule at 83.8% coverage brackets the same quantity between +16 and +25. The curated per-round lifts run +20 to +28, so the two overlap across most of their range while the curated rule runs slightly hotter at its top end. The direction is common to both, which is the claim; the exact magnitude is rule-dependent, which is why a bracket is reported rather than a single number.

The Mechanism: The Tail Is Vacated, Not Won

Decomposing the band split into all three buckets shows where the movement comes from, and it is not where the folk claim puts it.

Table 4

Three-Bucket Citation Composition by Band, Rounds 5 and 6

Band	Round 5 (mega / big / small)	Round 6 (mega / big / small)
HEAD	32% / 40% / 28% (20 cells, n=147 citations)	25% / 40% / 34% (20 cells, n=157 citations)
MID	31% / 21% / 48% (20 cells, n=159)	30% / 21% / 49% (20 cells, n=157)
TAIL	33% / 12% / 56% (44 cells, n=308)	29% / 14% / 57% (43 cells, n=315)

Note. Shares are rounded independently and do not always total 100.

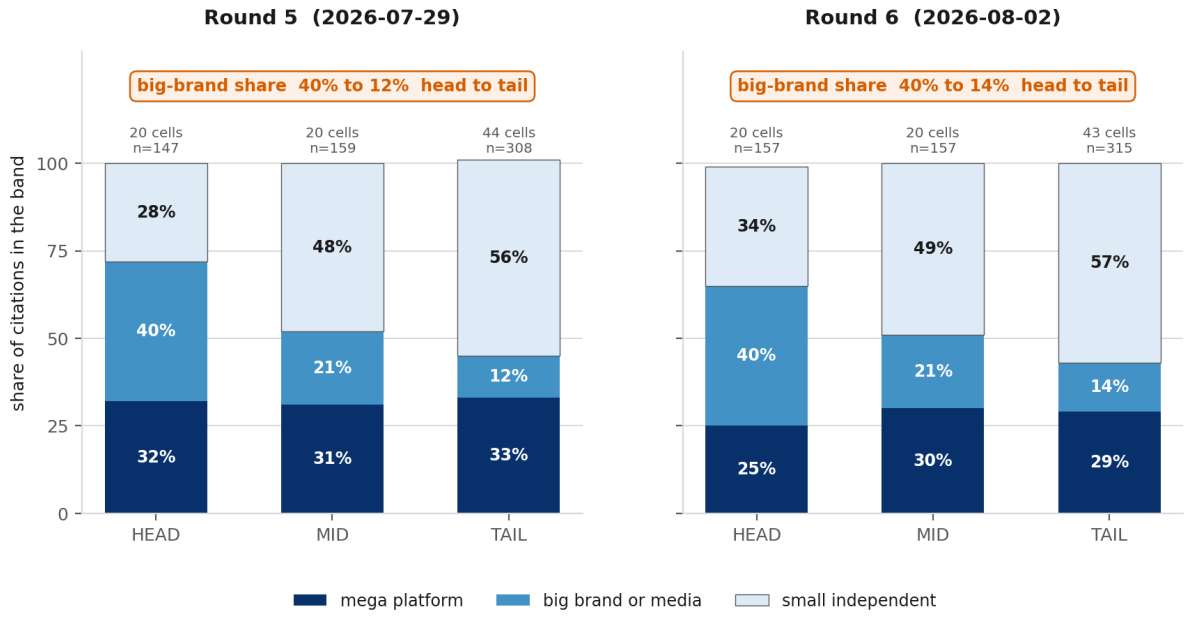
Mega-platform share is close to flat across bands in both rounds (31 to 33% in Round 5, 25 to 30% in Round 6). Big-brand share falls from **40% of head citations to 12 to 14% on the tail** in both. The entire lift in small-site share is that collapse.

So the accurate sentence is **"big brands stop competing on the tail, and what is left is smaller,"** not "small sites win more on the tail." The tail is vacated, not won, and the distinction changes the practical advice completely.

We state the N of the mechanism claim separately from the N of the gradient claim, because they are different. The small-share gradient reproduces in **all six rounds** of the same repeated-measures ladder. The three-bucket decomposition that identifies the mechanism was recorded in full in **two rounds (5 and 6)**, so the mechanism is the weaker of the two claims and should be read as such.

Figure 5

The Tail Is Vacated, Not Won: Mega Share Is Flat, Big-Brand Share Collapses



Note. N = 84 cells in Round 5 and 83 in Round 6. These are the only two rounds whose full three-bucket split was recorded, so the MECHANISM rests on 2 rounds while the small-share gradient itself reproduces in 6 of 6. Mega-platform share moves 31-33% (R5) and 25-30% (R6) across bands, essentially flat. Big-brand share falls from 40% of head citations to 12-14% on the tail. Buckets come from the frozen v4 curated classifier, which is why Figure 3 rather than this figure carries the headline claim.

Pool Size Does Not Track Competitiveness

The three matched pairs were the design's strongest instrument, because each holds topic constant while varying specificity. They disagree with each other, stably, and that disagreement is the result.

Head-to-tail delta in mean cited-pool size, averaged across the four markets:

Table 5

Head-to-Tail Change in Mean Cited-Pool Size for Three Matched Query Pairs

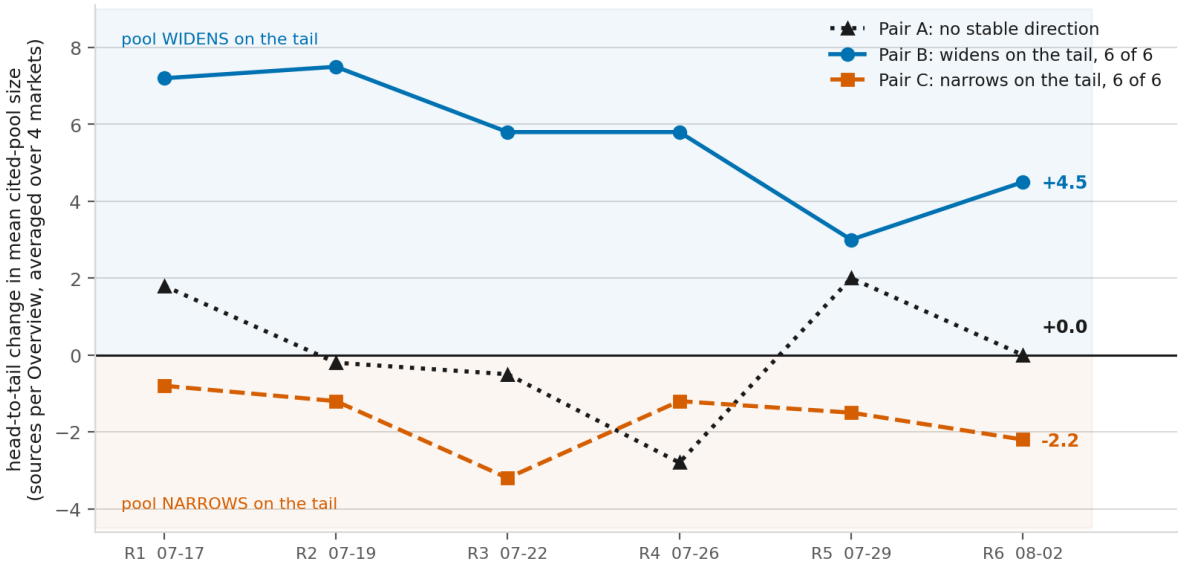
Pair	R1	R2	R3	R4	R5	R6	Verdict
A: head `best ecommerce platform`, tail `...handmade candle business`	+1.8	-0.2	-0.5	-2.8	+2.0	0.0	no stable direction
B: head `shopify vs woocommerce`, tail `...small uk clothing brand`	+7.2	+7.5	+5.8	+5.8	+3.0	+4.5	widens, 6 of 6
C: head `how to reduce cart abandonment`, tail `...one-product store`	-0.8	-1.2	-3.2	-1.2	-1.5	-2.2	narrows, 6 of 6

Pair B widened on the tail in **six rounds of six**, though the effect is smaller in the back half than the front. Pair C narrowed on the tail in **six rounds of six**, the most consistent single behavior in the study. Pair A is positive in Rounds 1 and 5, negative in Rounds 2, 3 and 4, and closes at exactly zero, so it has no stable direction and we report it as noise.

Three matched pairs, three different behaviors, each replicated across six rounds. Cited-pool **size** on the tail is therefore query-specific, and any paper claiming "the long tail has a bigger citation pool" would be selecting a pair. Pre-declared **Null 1 is published as a result on the pool-size measure**, at the same time as the composition measure moves in the other direction.

Figure 6

The Pre-Declared Null, Published: Cited-Pool Size Does Not Track Competitiveness



Note. N = 3 matched head-to-tail query pairs, each captured in 4 markets in each of 6 rounds (144 cells). Each pair holds the topic constant and varies only specificity. Pair B widens on the tail in 6 rounds of 6, Pair C narrows in 6 of 6, and Pair A is positive in rounds 1 and 5, negative in 2, 3 and 4, and closes at exactly zero. Three pairs, three individually stable but mutually contradictory behaviors. A study running any ONE of these pairs would have published a confident and wrong finding.

That split is the honest shape of the answer, and we resist resolving it into a single verdict. Composition changes; size does not.

Pair-level small-site share at close (Round 6, frozen curated rule) reinforces that the two measures are independent: Pair A 40% to 53%, Pair B 67% to 47%, Pair C 50% to 70%. Pair B's pool widens on the tail while its small-site share falls, which is the clearest single demonstration in the dataset that pool size and pool composition are not the same variable.

Instability, and Where It Concentrates

Round-to-round overlap was measured as mean per-cell Jaccard similarity of cited host sets. A pair is formed only where a cell exists in both rounds, so the closing read is computed over 83 cells rather than 84, and the query lost to the MISSED cell contributes three markets there rather than four. Five reads are available across six rounds: **0.43, 0.47, 0.48, 0.44 and 0.59**, giving a mean near **0.48** with one high outlier at close.

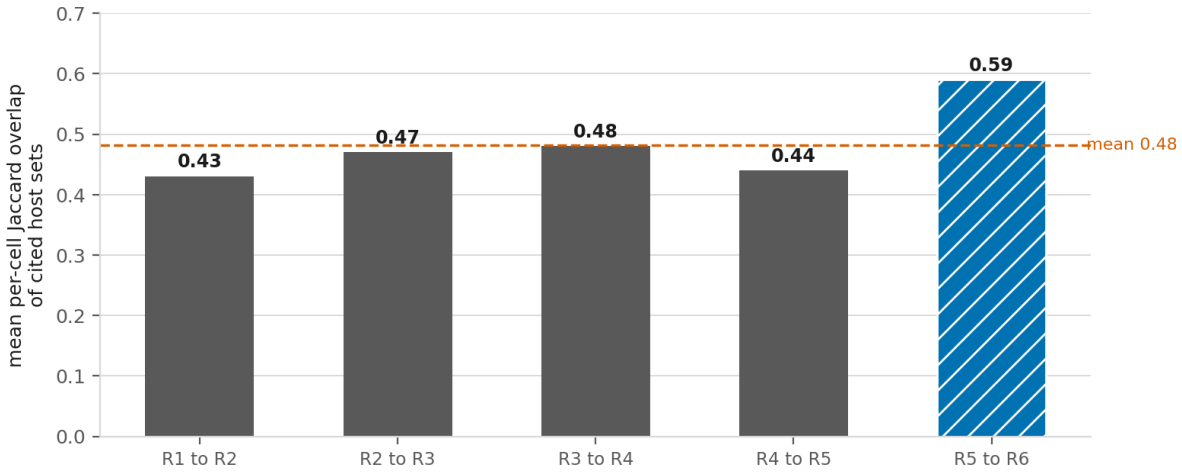
Roughly half the cited set turns over between rounds. In the final read, 151 of 280 union hosts (54%) appeared in both rounds; in the R4 to R5 read it was 125 of 309 (40%).

The band-level claim depends on the measure, and we state both rather than pick the favorable one. **Comparing each band's union of cited hosts between rounds, TAIL is the least stable band in all five reads** (union Jaccard 0.23, 0.35, 0.39, 0.29, 0.42). Under the per-cell measure this section is built on, averaged within a band, TAIL is least stable in three of the five reads and HEAD in the other two (Rounds 2 to 3 and 3 to 4). The individually least stable cells are tail cells in every read, so the tendency is real, but "without exception" belongs only to the union measure. Least stable cells at close were `how to get products in google shopping for free` (0.30), `how to sell on amazon` (0.30), the UK clothing-brand tail (0.36) and `will ai overviews hurt ecommerce` (0.37). Most stable were `how to reduce cart abandonment` (1.00, identical cited sets in all four markets, the only perfect cell in the study), `is dropshipping dead` (0.86) and `best ecommerce platform` (0.81).

One observation cuts against a tidy reading. In Round 4 the head anchor `how to sell on amazon` churned like a tail query (0.38), so "head terms are sticky" is a tendency in this data, not a rule.

Figure 7

Roughly Half the Cited Set Turns Over Between Rounds



Note. N = 5 round-to-round reads across 6 rounds, 503 cells. The final read of 0.59 is the highest in the study and sits well above the other four, which is one reason the six-round close rather than the designed eight is a real limitation on this measure specifically. By band-level union of cited sets, TAIL is the least stable band in all five reads (0.23, 0.35, 0.39, 0.29, 0.42); by mean per-cell Jaccard within a band, TAIL is least stable in three of five reads and HEAD in two, so the flat claim holds only under the union measure and the text states both.

Citation Versus Organic Rank, in Both Directions

Both directions were computed every round, and they behave differently. Reporting only one of them is how the tidy "about 40%" story survives.

Table 6

Overlap Between Cited Sources and the Organic Top 10, Both Directions, by Round

Round	Cited that also rank top 10	Top 10 that get cited
1	41% (317/773)	42% (317/748)
2	40% (313/792)	43% (313/734)
3	42% (340/807)	46% (340/740)

Round	Cited that also rank top 10	Top 10 that get cited
4	48% (331/683)	45% (331/736)
5	54% (329/614)	45% (329/739)
6	52% (327/629)	43% (327/752)

The cited-that-rank figure **drifts upward** across the study, from 41% to a close of 52 to 54%, moving roughly in step with the narrowing pool: fewer cited slots, and the ones that remain go more often to pages that already rank. The reverse direction stays **pinned between 42% and 46% in every single round**.

The practical implication is unchanged by the drift and is worth stating plainly. Roughly half of cited sources did not rank in the organic top 10 for the same query in the same capture, so ranking is one input to citation and not the gate. We note the limit of our own instrument here: our organic capture records only the top 10, so a cited site absent from that array may be ranking at 11 or beyond, or nowhere at all, and this study cannot distinguish those cases.

Market Dependence

Cited sets differ across the four markets for the same query captured within the same slot. This is a repeat of an effect we have observed before in a different vertical (WpConsults Research Lab, 2026b), and this study adds a designed sample and a demonstration of how easily it is over-read.

At close (Round 6, 21 cells per market, directional), YouTube led in all four markets (US 18, CA 20, UK 20, AU 17). Reddit skewed non-US (US 9, CA 13, UK 10, AU 13). Shopify.com was even (US 9, CA 6, UK 10, AU 9).

The Quora result is the useful one. Quora skewed away from Australia in Round 4, away from the United Kingdom in Round 5, and away from Canada in Round 6 (US 5, CA 2, UK 3, AU 4). **Three consecutive rounds produced three different market skews on the same**

platform, which is the cleanest demonstration in the dataset that a single round's market skew is churn rather than a market fact.

Any AI-visibility figure reported without a market and without repeated measures is therefore under-specified, including figures reported by vendors and including the ones in this section.

A measurement artifact is recorded for reuse of our data. The labeling script folds every `*.co.uk` host into a single "co.uk" label and every `*.com.au` host into "com.au". These are many distinct small UK and Australian sites rather than one platform, and the artifact reinforces rather than manufactures the small-site reading.

The Stable Core

A small set of hosts is cited across most of the ladder in every round. At close (Round 6, number of the 21 queries citing the host at least once): YouTube 20, Reddit 16, shopify.com 12, Quora 8, printful.com 6, Facebook 5, wise.com 5, Forbes 5, LinkedIn 5, Norton 3, NordVPN 3, Avast 3.

YouTube appeared in 18 to 21 of the 21 queries in every round (21, 18, 18, 19, 20, 20 across Rounds 1 to 6), and was cited in 74% to 93% of all cells in a given round. Reddit was cited in 51% to 58% of cells in every round. The core composition is unchanged from Round 1 to Round 6.

The range reported here follows the per-round log, which is the primary record. A documentation discrepancy in the study's summary block, and how it was settled, is described under Limitations.

Intent, With a Caveat That Limits It

Under the frozen curated classifier, small-site citation share by intent band in the three rounds scored under that one rule was as follows. Each round covers every cell captured that round (84, 84 and 83 cells for Rounds 4, 5 and 6), distributed across seven intent tags, so a single tag can rest on as few as four cells per market.

Table 7

Small-Site Share of Citations by Query Intent, Rounds 4 Through 6

Intent	R4	R5	R6
trust	32%	24%	28%
transactional	43%	37%	54%
frontier	50%	40%	50%
informational	53%	44%	45%
commercial	51%	55%	53%
operator pain	61%	50%	50%
our lane (AI search topics)	65%	64%	68%

The "our lane" band finished first or second in all six rounds and under every classifier version, making it the sturdiest intent result the study has. The trust band was the most brand-locked in each of the three rounds scored under the frozen rule, which is unsurprising given that consumer security and review brands cluster there.

The caveat is binding. These intent figures rest on the curated classifier and were **not** covered by the domain-authority sensitivity pass, which was computed over bands rather than intent tags. They are therefore weaker than the band results in Table 3 and should not be quoted at the same confidence.

Cross-Lane Triangulation, Reported Side by Side

We hold a second, independently sampled lane in a different vertical (SEO, WordPress and AI-search practitioner queries), captured with the same instrument and the same verification rules but on a different query set. It is a **separate sample with its own N** and is never merged with the ecommerce lane into a single percentage, for the same reason we never average across markets.

Table 8

Convergent Observations Across Two Independently Sampled Lanes

Observation	Ecommerce lane (this study)	SEO lane (separate sample)
Citation versus organic rank	41% to 54% of cited also rank top 10; 42% to 46% of top 10 get cited. N = 503 cells, 6 rounds	39% of cited sites rank top 10. N = 99 cited entries across 15 queries
UGC and video over-cited	YouTube 74% to 93% of cells per round; Reddit 51% to 58%. N = 503 cells	Quora and YouTube lead the tally; Quora 51 rows, YouTube 42, Reddit 23. N = 89 verified rows. Later captures in that lane are excluded because they are not yet separable by vertical from a combined count
Market dependence of the cited set	Cited sets differ by market on the same query in the same slot; three different Quora skews in three rounds. N = 503 cells	Same effect observed on a balanced same-query view. N = 6 captures per market on one query
Churn	Mean per-cell Jaccard about 0.48 across five reads. N = 503 cells	A four-round vendor core vanished and stayed gone for six rounds

Two independent samples pointing the same way is convergent validity and is stronger than either alone. It is not a larger sample, and the reader should note that the two lanes land at different values on the citation-versus-rank measure (39% in the SEO lane, drifting 41% to 54% in the ecommerce lane). We report the divergence rather than the average, and treat it as a possible vertical effect worth its own study.

Discussion

The practical advice this study set out to test is half right for the wrong reason, and the reason matters more than the direction.

Moving to the long tail does change who you are competing with, and the change is large. Under the numeric authority rule, small-site share of citations rises by 16 to 25 points from head to tail depending on the threshold and the treatment of unscored domains. Under the separate curated classifier, and in the two rounds where the full three-bucket split was recorded, big-brand share falls from 40% at the head to 12 to 14% on the tail. Those two figures come from two different rules and are not one analysis, but they point the same way: a small operator publishing on specific questions is entering a materially different competitive set.

But the mechanism is not that the system opens a door for smaller sites. Mega-platform share stays roughly flat across bands, so the shift is not a redistribution toward the small; it is the withdrawal of a category that was never going to write the specific page in the first place. **The tail is vacated, not won**, and advice built on "the algorithm favors small sites down there" is building on a mechanism our data does not show.

What this study cannot say is which property of a page earns it a place in the retrieved set. Two candidate mechanisms are live in the practitioner literature and in our own prior work: **extractability**, meaning whether a page states a self-contained answer to the exact question in a liftable passage, and **corroboration**, meaning whether the same claim is available from enough independent sources for the answer layer to treat it as consensus. This study manipulated neither, so it can motivate both and settle neither. Separating them requires holding rank constant while varying extractability on a property the researcher controls, which is a different design and a separate paper.

Practical Implications

Four conclusions follow for a publisher working in this vertical, and they are gathered here because they are the part of this paper a practitioner is most likely to act on.

The competitive set on the tail is materially different, so moving there is a real change of opponent rather than a change of tactic. But the mechanism is vacancy, not favor, so nothing about the tail should be expected to reward a weaker page; it simply contains fewer strong ones.

The pool is already wide open at the head (top-five platforms never above one third of citations, 200 to 332 distinct hosts per round across 503 cells), so retreating to the tail to escape a closed door is retreating from a door that was not closed.

The tail is the least stable band in the churn reads, so what is won there has to be maintained rather than won once. And there is no query specific enough to sit below the feature: the Overview fired on every one of the 504 designed cells we could resolve, including every roughly zero-volume and frontier query, so the strategic question is not whether to engage with it but where.

The pool-size null deserves the same emphasis as the composition result, which is why it appears in the abstract. Three matched pairs produced three different, individually stable behaviors across six rounds each. Any study that had run one pair would have published a confident and wrong finding in whichever direction its pair happened to point, and that is a general warning about single-query evidence in this literature rather than a complaint about anyone in particular.

On the falsified null, the direction of the surprise is worth naming. We expected the feature to thin out on near-zero-volume and frontier commercial queries, and it did not thin out at

all. Across 504 designed cells, including every frontier query about AI shopping agents, not one Overview-free result page was found.

Finally, this study is a caution about our own class of evidence. We changed our size classifier three times before freezing it, and the head number moved by more than 20 points on the placement of about 24 domains.

Any single-classifier size claim in this field, including one published by a vendor with a larger sample, should be read as a classification decision until it is shown to survive a sensitivity analysis. That is why this paper publishes a table of bounds instead of a number.

Limitations

These are stated at full length, and none of them is qualified away.

Six rounds, not the designed eight. The study was designed for eight rounds and closed after six on the principal investigator's instruction on 2026-08-02. This bears most directly on the churn measure, which has only five round-to-round reads, and on the pool-contraction reading, where a two-round narrowing at Rounds 4 and 5 was reversed by a single round and is reported here as fluctuation on the strength of one additional observation. With eight rounds, that call would have been better founded.

One MISSED cell. 503 usable of 504 designed. The lost cell is an instrument fault (sources panel refused to open on two reads, screenshot verified as a full signed-out SERP with a fired Overview), not a block and not an absent Overview. The same query and market also required a retry in Round 5, so this cell has a repeat instrument weakness.

16.2% of citations are unscored for domain authority. 200 of 667 distinct domains were scored, covering 3,602 of 4,298 citations. The remaining 467 domains, of which 398 are cited once or twice and 69 three or four times, are handled by the two-way bound rather than by

an assumption. The bound is wide (+16 to +25 points), and at its worst extreme TAIL sits at or below MID at every threshold, which is why no smooth three-step gradient is claimed anywhere in this paper.

Size classification is a decision, not a measurement. Even under a numeric authority rule, the choice of threshold and the choice of authority metric are decisions. We mitigate by reporting three thresholds and two extreme assumptions, not by claiming the decision away. Intent-level size figures (Table 7) were not covered by the sensitivity pass at all and are correspondingly weaker.

Twenty-one cells per market per round. All market-level percentages are directional. A market difference of a few cells is not a finding, which is exactly why the Quora result is presented as a demonstration of churn rather than as a geography claim.

One operator, one instrument, one vertical, one 17-day window. Nothing here is a population estimate. All captures were taken by a single operator through a single scripted browser tool, in one ecommerce query set, between 2026-07-17 and 2026-08-02, in four English-language markets.

Instrument risk, stated specifically. The capture tool's cited-array parser was observed contradicting its own screenshots in a separate study on 2026-07-29, and the parser has not been read line by line. The same instrument captured both lanes in Table 8, so a parser-level fault would correlate across the two samples and weaken the cross-lane convergence in exactly the dimension it is offered as evidence. Results that do **not** depend on the cited array are unaffected: the Overview fire rate reported in Results and the eye-checked positive screenshot rows. Results that **do** depend on it include the composition table, the pool-size deltas, the churn measure and the citation-versus-rank figures. Our mitigation is the per-round positive screenshot check, in

which a sampled row's cited set was matched by eye against the stored frame in every round; that is a sample, not an audit, and a full parser verification remains outstanding.

Organic capture is limited to the top 10. A cited site absent from our organic array may rank at 11 or beyond, or not at all, and this study cannot separate those cases. Any reading of Table 6 as evidence that cited sites "do not rank" overstates what the instrument can see.

Round 1 was captured before the pacing fix. Round 1 fired all 84 reads plus retries from one address in about an hour and tripped Google's rate limiting near query 19, losing two cells to a 403 that were back-filled clean the same day after a cooldown. Rounds 2 through 6 ran on the paced two-slot schedule with zero blocks across ten slots. Round 1 is therefore captured under a slightly different regime than the rest and is retained because its two affected cells were re-captured and verified.

A split-round smear. Each round is captured in two slots about 13 hours apart, so a round is a window rather than an instant. The split is identical in every round, making it a constant.

Measure 11 is NOT COLLECTED. The Shopping-module flag was never emitted by the harness in any round. It is not a null, and no claim about Shopping surfaces appears in this paper.

Two documentation discrepancies were found and reconciled. The study text described matched Pair A as changing sign five times, while the logged six-round series (+1.8, -0.2, -0.5, -2.8, +2.0, 0.0) turns sign exactly twice before closing at zero; and the study's summary block gave YouTube's range as 19 to 21 of 21 queries where the per-round log records 18 in Rounds 2 and 3, the per-round series being 21, 18, 18, 19, 20, 20. In both cases the per-round log is the primary record and is what this paper reports, and in both cases the summary

text has been corrected in the study record rather than left standing. No published number moves as a result, and the Pair A verdict of no stable direction holds under either description.

The author's own property was not tracked. By design, nothing in this study concerns whether our own site was cited, so no self-interested reading of the data is possible from it, and equally no first-party causal claim can be made from it.

Conclusion

On a designed 21-query ecommerce ladder captured in four markets over six rounds (503 usable cells of 504), the composition of Google's AI Overview citation set changes with query competitiveness while the size of the cited pool does not.

Small-site share of citations rises from head to tail by 16 to 25 percentage points under the numeric authority rule, positive at three thresholds and under both extremes of a two-way sensitivity analysis covering 83.8% of citations. The mechanism, decomposed under the separate curated classifier in the two rounds where the full split was recorded, is that big-brand share falls from 40% of head citations to 12 to 14% on the tail while mega-platform share stays roughly flat. Big brands stop competing on the tail; small sites do not win more of a contested space.

Cited-pool size does not track competitiveness. Three matched head-to-tail pairs produced three individually stable but mutually contradictory behaviors over six rounds each, so our pre-declared "no gradient" null is published as a result on that measure.

No Overview-free result page was found among the 504 designed cells, including every roughly zero-volume tail query and every AI-shopping frontier query, falsifying our pre-declared null that the feature would not fire. For a publisher in this vertical there is no query specific enough to sit below it.

Data Availability Statement

The raw data, capture frames and analysis code listed below are deposited with this paper in one Zenodo record, https://doi.org/10.5281/zenodo.21923520, so that every figure above can be re-derived independently. The record's own file list is the authoritative inventory, and the manifest deposited alongside these files carries a SHA-256 for each one.

- **Raw capture rows:** 157 JSONL files (JSON Lines, one JSON record per line) across six round directories (`captures/round-1` through `captures/round-6`), one row per query per market per attempt, carrying the cited host list, the organic top 10, the fire flag, the panel state, the footer binding and the usability flag. Refused and lower-bound rows are retained in place with their error state so that nothing counted is hidden and nothing hidden was counted.
- **Capture frames:** 504 full-page PNG screenshots, one per designed cell. Attempts exceeded 504, because two Round 1 cells were back-filled after a block, one Round 5 cell was retried and Round 6 carried four single-market retries; retries overwrote their frame, so the count is one per designed cell rather than one per request. Approximately 697 MB in total. These include the frame for the single MISSED cell and the per-round positive-check frames referenced under Verification, Usability, and the Treatment of Failures.
- **Domain-authority scores:** `da-scores.tsv`, 200 rows of `domain <tab> authority score`, values ranging 1 to 100 with a median of 48.5.
- **Cited-domain volumes:** `cited-domains-by-volume.txt`, 667 rows of `domain <tab> citation count`, summing to 4,298 citations. Joining this file against the score file reproduces the 83.8% coverage figure exactly.

- **Analysis code:** `gradient-analysis.py`, carrying the frozen v4 size classifier, the band and intent assignments, the matched-pair logic and the churn computation, applied identically to all six rounds.
- **Figure code:** `paper-01-figures.py` regenerates all seven figures; the script is the artifact, and running it reproduces every figure. It also carries, in its header, the design constraints that keep the figures honest, including why Figure 3 must draw intervals rather than a line. For provenance: Tables 2, 4, 5, 6 and 7 and the churn figures derive from `gradient-analysis.py` over the raw rows, Table 3 from `da-sensitivity.py`, and Table 8's SEO-lane column from the separately deposited SEO-lane findings.
- **Study log:** the full round-by-round record, including every instrument fault, every retry, every pacing change and every reading we later downgraded.
- **Sensitivity code:** `da-sensitivity.py` reproduces Table 3 from the raw capture rows and the score file. It imports the frozen classifier's own loader, band map and root-folding rather than re-implementing them, so the two passes cannot drift apart. It prints the split on both a pooled-six-round and a Round 6 basis, so the basis of the published table is visible rather than asserted, and it prints the size of the disagreement between the curated host-level rule and the numeric root-level rule.

Ethical note: all captures were of publicly accessible search result pages, taken signed out, at a paced request rate with explicit cooldowns, with no attempt to circumvent any block. No personal data was collected.

References

Authoritas. (2025). *The state of AI Overviews: User intent research*. https://www.authoritas.com/seo-ai-research-whitepapers/the-state-of-aios-user-intent-research-dec-2024

Conductor. (2025). *AI Overview analysis and study* (September 2025 update). https://www.conductor.com/academy/ai-overviews-analysis/

Okabe, M., & Ito, K. (2008). *Color universal design (CUD): How to make figures and presentations that are friendly to colorblind people*. https://jfly.uni-koeln.de/color/

Seer Interactive. (2026). *AIO impact on Google CTR: 2026 update*. https://www.seerinteractive.com/insights/aio-impact-on-google-ctr-2026-update

Semrush. (2025). *The most-cited domains in AI: A three-month study*. https://www.semrush.com/blog/most-cited-domains-ai/

WpConsults Research Lab. (2026a). *Ecommerce AI Overview citation gradient: Study log, 2026-07-16 to 2026-08-02* [Study record]. Zenodo. https://doi.org/10.5281/zenodo.21923520

WpConsults Research Lab. (2026b). *Who the AI Overview actually cites, and how it shifts by country* [Standing finding]. Zenodo. https://doi.org/10.5281/zenodo.21923520

Appendix A

The Two Pre-Declared Nulls, as Written Before Capture

Reproduced verbatim from the method, fixed before the first capture on 2026-07-17.

Pre-declared nulls we WILL publish: "no gradient" (the pool is the same on head and tail), and "the AIO does not fire on long-tail / frontier ecommerce queries." Both ship as-is.

Outcome: the "no gradient" null **survives and is published** on the pool-size measure (Table 5) and **does not survive** on the composition measure (Table 3). The "no fire" null is **falsified**: no Overview-free result page was found among the 504 designed cells, as reported under The AI Overview Fires Everywhere in This Vertical. Both outcomes are reported at equal length, and neither was rewritten after the data arrived.

Appendix B

Glossary

Terms are defined as this paper uses them. Where a term names a decision we made rather than a property of the world, the entry says so.

Table B1

Glossary of Terms as Used in This Paper

Term	Definition
AI Overview	Google's generated answer block, shown above the organic results, which attributes its answer to a set of cited web sources.
Band (HEAD, MID, TAIL)	The competitiveness tier assigned to each query before capture from its United States search volume and SEO difficulty: HEAD for high-volume competitive terms, TAIL for hyper-specific terms at roughly zero volume, MID between. A design decision, fixed before any data arrived.
Big brand	In the curated size classifier, a large commercial brand or established media property that is not one of the global platforms: ecommerce and SaaS vendors such as Shopify, Stripe and HubSpot, national and international publishers such as Forbes, the BBC and the Guardian, and the consumer security and review brands such as Norton, Avast and Trustpilot added when the rule was corrected at Round 4.
Cell	One query, captured in one market, in one round. The unit of observation: 21 queries by 4 markets by 6 rounds gives the 504 designed cells.
Chip (source-count chip)	The small count label on an Overview's sources panel, for example "6 sites", stating how many sources the panel holds. Logged on every capture and used as a cross-check on the cited list.
Citation instance	One appearance of one host in one cell's cited list. A host cited in three cells contributes three instances. The 4,298 total is a count of instances.
Cited host	A distinct hostname in a cited list, counted at full-host grain, so `community.shopify.com` is a different host from `shopify.com`. There are 695 across the study.
Cited set (cited pool)	The set of sources an Overview attributes for one query in one capture; at round level, the union of those sets across cells.
Corroboration	A candidate mechanism for citation: whether the same claim is available from enough independent sources for the answer layer

Term	Definition
	to treat it as consensus. Discussed, not manipulated in this study.
Domain authority	Ubersuggest's 1 to 100 estimate of a domain's overall strength, derived from its backlink profile. A vendor estimate, not a property of the domain, used here because it is reproducible and independent of our judgment.
Extractability	A candidate mechanism for citation: whether a page states a self-contained answer to the exact question in a liftable passage. Discussed, not manipulated in this study.
Frontier query	A tail query about behavior that barely exists yet, such as an AI agent buying products, whose volume is near zero because the underlying activity is new.
Generative engine optimization	The practice of trying to earn presence and citations inside AI-generated answer layers, as distinct from ranking in the organic results.
Jaccard similarity	A 0 to 1 measure of overlap between two sets: the size of their intersection divided by the size of their union. Used here to compare a cell's cited set across consecutive rounds. 1 means identical sets, 0 means no host in common.
Long tail	The large population of specific, individually low-volume queries, as against the small number of high-volume head terms.
Matched pair	Two ladder queries holding topic constant while varying specificity, a head form and a tail form of the same question, so a head-to-tail comparison is made on the same underlying subject.
Mega-platform	In the curated size classifier, the largest general-purpose platforms and marketplaces, whose pages are typically user or seller generated at global scale: YouTube, Reddit, Quora, Facebook, LinkedIn, Medium and Wikipedia, together with the global marketplaces such as Amazon, eBay and Etsy.
MISSED	The classification for a cell lost to a blocked or faulty instrument rather than observed. A MISSED cell is never counted as an absent Overview.
Root domain	The registrable domain into which hosts fold: `community.shopify.com` and `shopify.com` share the root domain `shopify.com`. The numeric authority rule operates at this grain, which is why it counts 667 domains where the host grain counts 695.
Search volume	The estimated monthly search count for a query, here from Ubersuggest United States data, used with SEO difficulty to assign bands.
SEO difficulty	A tool-estimated 0 to 100 score of how competitive a query's organic results are, here from Ubersuggest, used with search volume to assign bands.
SERP	Search engine result page: the page Google returns for a query, carrying the Overview, the organic results and any other modules.

Term	Definition
Signed out	Captured with no Google account logged in, so results are not personalized to an account's history.
Small independent	In the curated size classifier, everything that is neither a mega-platform nor a big brand: independent stores, blogs, niche publishers, and any host whose subdomain prefix marks it as a community, forum, support, help or blog page.
Sources panel	The expandable list attached to an Overview naming the pages it cites. A capture counts only if this panel was opened and scrolled to its end.
Usable row	A capture meeting every verification condition: Overview expanded, sources panel scrolled to exhaustion, market confirmed against the page footer. Only usable rows contribute cited lists.
User-generated content	Content written by a platform's users rather than by its operator, as on Reddit, Quora or a support forum.

Note. Size buckets are a classification decision rather than a measurement, as stated under Limitations. The mega-platform, big-brand and small-independent definitions describe the frozen curated classifier published with this paper; the numeric authority rule used for Table 3 is a separate rule operating on root domains, and the two do not always agree.