Cited on its name, and almost nothing else.

A site that did not exist in June, instrumented on both sides for five weeks. Asked for by name it is quoted almost every time. Asked about the subjects it publishes on, it has never been quoted.

0

of **1,664** asks about the subjects it publishes on returned a citation. Not a low rate. None.

NINETY DAYS, FIVE WEEKS IN

A first reading of a site that did not exist in June, instrumented on both sides. A second, pre-registered reading follows on 15 October, and its result is not ours to choose.

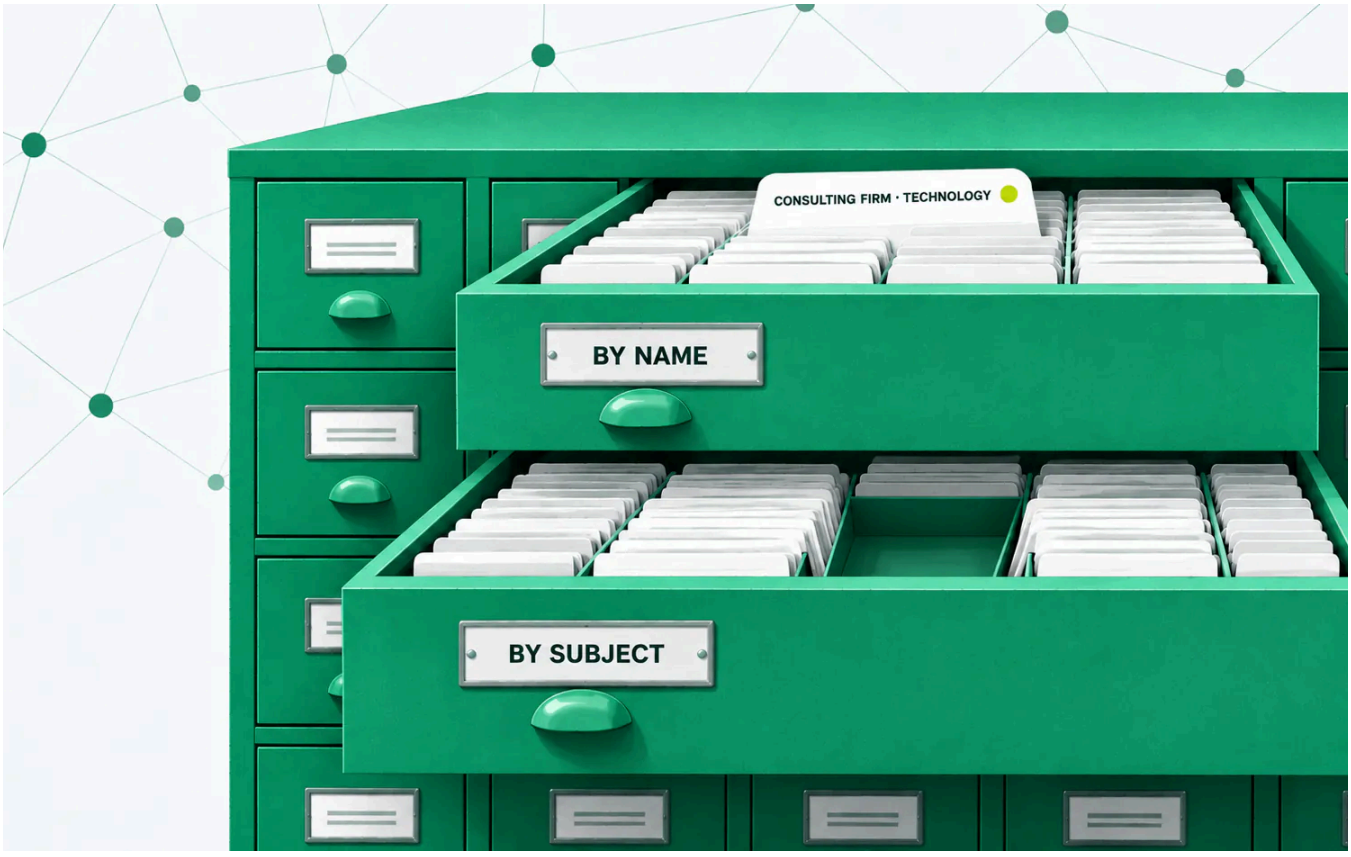
CORTEXA
CONSULTING

cortexa-consulting.ai/insights
Intelligence, applied.

Cortexa Consulting did not exist in June 2026. The domain was registered in July, the first article published on the seventh, and the site stayed behind a gate until 3 August. That makes it a poor business case and an unusually clean instrument. Every study of Generative Engine Optimization (GEO) we could find measures established brands, because that is where the traffic is. Their zeros are confounded by years of backlinks, existing entity records and prior press. Ours are not confounded by anything, because there was nothing there.

So this is a cold-start field study. We instrumented our own site on both sides: the answers an engine gives, and the server logs showing whether it came to read the page. We started measuring four days after the gate came down. You get one shot at this per domain and the window closes as the domain ages.

> **CORTEXA'S READ**
>
> The headline is not that a new site struggles. Everyone expects that. The headline is **which** questions it wins and which it loses, and the split is far sharper than we expected. Asked for by name, we are cited almost every time. Asked about the subjects we write about, we've never been cited.



The same record, and the same cabinet. Filed under the name, and missing from every subject we asked about.

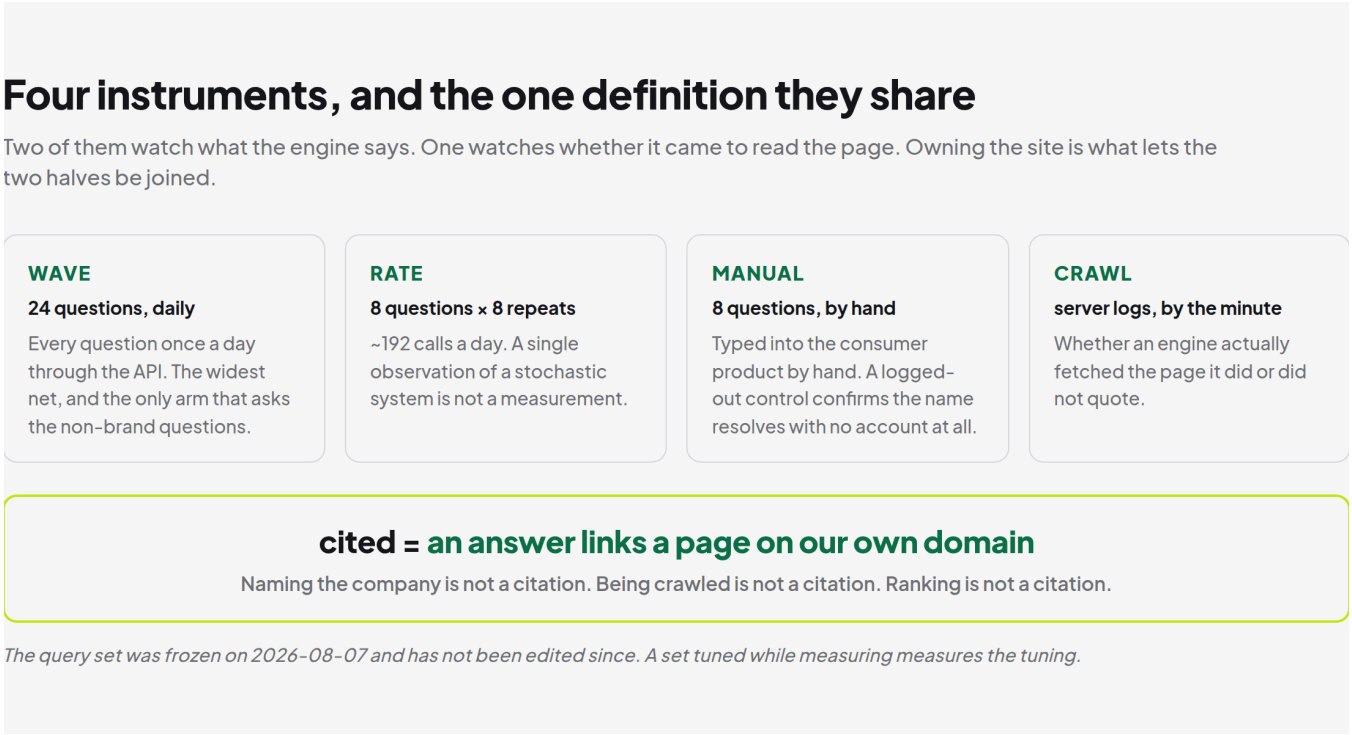
01

CORTEXA FIELD STUDIES Nº 01

What a cold start actually looks like

What was measured, and how?

A frozen set of 24 questions in three strata: eight that name the company, twelve about the topics we publish on, and four with buying intent. The set was fixed on 7 August and has not been edited since, because a query set you tune while measuring is a query set that measures your tuning. Each question exists in three paraphrase variants, pinned by checksum, on the same terms: add a variant, never edit one.



Two instruments watch what the engine says, one watches whether it came to read the page. Owning the site is what lets the two halves be joined.

- **Wave arm** — all 24 questions, once each, daily, through the Application Programming Interface (API). 35 days of data.
- **Rate arm** — the eight naming questions only, every variant, eight repeats per cell, daily. Roughly 192 calls a day, because a single observation of a stochastic system is not a measurement.
- **Manual arm** — the same eight questions typed into consumer ChatGPT by hand, in a Temporary Chat, three runs a day, rotating which variant goes first. The Temporary Chat was meant to stop the session remembering us from yesterday. Late in the study we found it does not, and the control below is how we checked what that cost.

- **Crawl arm** — minute-resolution server logs from the edge, so we can ask whether an engine fetched the page it did or did not quote.

The repeated sampling is not a flourish. Two 2026 preprints make the same argument independently: answers vary across runs, prompts and time, so visibility has to be treated as a distribution rather than a single result.[1] One reframes the whole problem as statistical estimation over a stochastic system.[2] We agree, and the design predates our reading of them, which is luck rather than foresight.

What was already built when we started asking?

This decides how to read the zero. A site nothing cites might simply be a badly made site, so here is what was in place before the first question was asked. None of it is exotic. It is the list an agency sells, done before measurement rather than after it.

- **Structured data on every page.** An Organization and WebSite entity graph with one canonical identifier that every other page references instead of restating. Articles declare themselves as articles, question-and-answer sections declare themselves as such, and every page carries a breadcrumb trail.
- **The unglamorous basics, done properly.** One H1 per page, descriptive alt text on every image, an absolute canonical URL on every route, meta descriptions written to length, and a sitemap that lists only pages we want indexed.
- **Submitted, not just published.** The sitemap went to Google Search Console and Bing Webmaster Tools rather than waiting to be discovered. Verifying the property looks like the larger of the two acts: Google was holding 21 of our URLs on 10 July, the day Search Console ownership was verified and 26 days before any sitemap reached it.
- **The front door held open.** Our robots file allows every AI crawler by name, bot challenges stay off in front of them, and a plain-text index of the articles sits at a fixed URL for anything that prefers to read a list.
- **Fourteen articles** live before the first question was asked, the last two published the day before, each written to answer one question and each carrying its sources.

Off-site, one corroborating record was already live before any of this began: a Crunchbase profile created in July. The rest came later, and their dates matter enough that we will return to them: a Clutch profile created on 25 August, with our town added to its description on the 26th, then Bizapedia, a Wikidata item, OpenCorporates, the state filing and a Dun & Bradstreet record through the end of August and the first half of September.

> **CORTEXA'S READ**
>
> Every item above is a real control, and together they bought exactly one thing: the name. Asked by name we are found, described accurately and located correctly. Asked about the subjects those same pages are written about, the same site is absent. The technical work is not what is missing.

What happens when nobody asks for you by name?

Nothing happens. This is the number the study exists to report:

- **Topic questions: 0 of 1,248.** Twelve questions about retrieval-augmented generation, agents, guardrails, evaluations and the rest of what we publish on, asked daily for 35 days.
- **Buying-intent questions: 0 of 416.** Four questions a prospect would plausibly type.
- **Combined: 0 of 1,664.** Not a low rate. Zero, with no exceptions, for the entire measurement period.

We publish carefully sourced explainers on precisely these subjects. Twenty-three of them, each with visible citations, each written to be the clearest page on its question. Across 1,664 opportunities to be quoted on that material, the engines quoted us zero times. Published surveys of the field report that roughly 80 to 95 percent of citations go to earned media and about a fifth to brand-owned pages.[3] Our figure is not a weaker version of that finding. It is the floor beneath it, which is what a site with no age and no external coverage has.

And when they do ask by name?

Then almost everything works, and the contrast is the finding:

- **Consumer ChatGPT, asked by name: 444 of 456 (97.4%).** Hand-collected, 19 collection days, 57 runs, from 20 August.
- **The same eight questions through the API: 3,327 of 4,165 (79.9%).** Same period, same wording, eight repeats per cell.

A site five weeks past its gate lift is not invisible. It is findable, quotable and correctly described, provided the person asking already knows the name. What it cannot do is enter a conversation it was not named in. Those are two different problems and the industry tends to price them as one.

What was the engine searching for?

We can answer that exactly, and we should have looked sooner. Two of the three engines record the sub-queries they issue before answering: the searches the model decided to run, in its own words. They have been sitting in our filed responses since the first wave, and reading them costs nothing we had not already paid for.

Across 1,656 question-instances, the pattern is the same on both:

- **Most non-brand questions never triggered a search at all.** They ran one on 124 of 1,104 instances. The rest were answered from the model's own training, with no retrieval step, so nothing about our site was consulted. Brand questions searched on 473 of 552.
- **When a non-brand question did search, the searches never named us.** Not one of 853 non-brand sub-queries contained the word “Cortexa”. Of 1,465 brand sub-queries, 1,448 did.

What the engines asked for on a topic question was “RAG vs fine-tuning pros and cons” and “Model Context Protocol specification”. Whoever ranks for those is the candidate list, and a domain ten weeks old is not on it.

CORTEXA'S READ

This changes what you would do about it. Our pages did not lose a comparison. They were never in the comparison. A better page or a cleaner schema is not consulted by a question that runs no search, and it does not help a search that went looking for something else. The first gate is whether the question retrieves anything. The second is whether the search names you. Everything the industry sells as content quality applies after both, and most of what we asked never got that far.

Two limits on this, and they matter. The third engine returns the pages it used and never the searches it ran, so about a third of the 1,664 is outside this analysis and its zero stays unexplained rather than explained differently. And we can see what was asked, not what came back or why one page was chosen over another. It says why we were absent. It says nothing about why anyone else was present.

Why did the same question get two different answers?

Those two figures come from the same eight questions in the same words over the same weeks. The gap between 97.4% and 79.9% is not our site changing. It is the surface changing: a consumer product and a developer API, nominally the same vendor, behaving differently on identical input.

That has a practical consequence for anyone buying a GEO report. A vendor measuring through an API and a client checking the chat app will disagree, both will be right about what they measured, and neither number describes the other. We keep the two series side by side and labeled, and we never pool them.

We are not the only ones who ran into this, and the larger version of it is instructive. When the same vendor's citation mix shifted in August 2026, two measurement firms disagreed about whether anything had happened. Promptwatch, sampling the consumer product, recorded the change on 8 August. dejan.ai, sampling 196,692 background queries through the same vendor's API, recorded none.[8] A third tracker watching both put one source's share of consumer citations falling from about 4.7 percent to about 0.4, while the API held near 0.67 across the same weeks.[9] One vendor, one question, two instruments, two answers, at a sample size we cannot reach. Our 97 and 79 are the small version of the same thing.

How do we know the 97% isn't the machine remembering us?

Because we checked, and the reason we checked is that we never took it on trust. The hand-collected arm was typed on the founder's own signed-in account, in a Temporary Chat, which the vendor documents as carrying no memory.[6] On 9 August the study wrote down that a reviewer would not accept that on trust, and named the specific risk: memory shaping the search query rather than the answer. That kind of contamination leaves no fingerprint in the citations at all.

Two things then happened. At the end of August the vendor added a personalized Temporary Chat, which can read saved memories where the default still does not. And on 15 September a temporary chat on that account answered a bare-name question by offering to help with the founder's own private working topics. Everything done outside a temporary chat for this business had been saved to the account, so the material

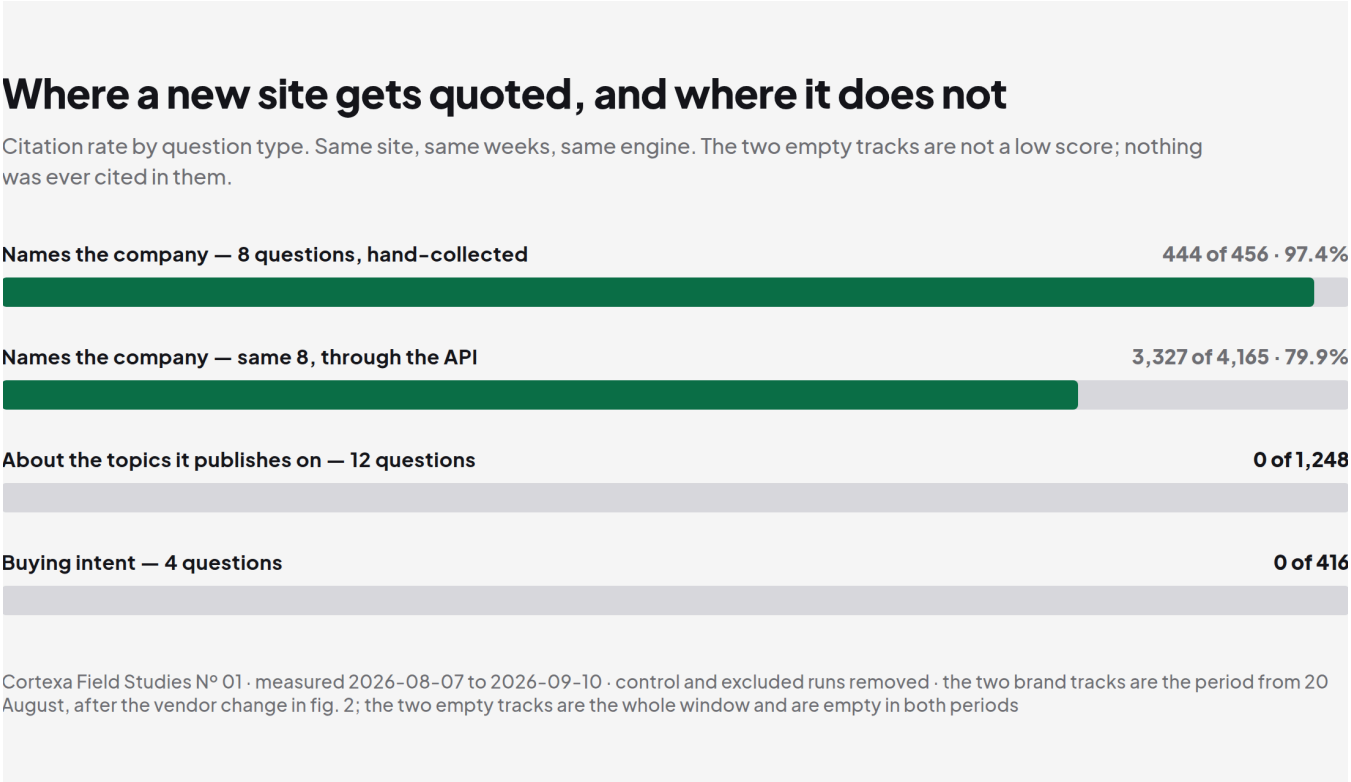
was there to be reached, and a chat that consults it before searching can shape the search without ever showing you that it did. Nothing in five weeks of data would have told us. It surfaced because one answer listed his projects back at him.

That puts a real question over the 97.4%. If the session knew us, the number measures recall rather than retrieval. So the next morning we ran the same questions with no account at all: four fresh logged-out sessions, one question each, no follow-ups.

- **All four were cited, on the first turn, with no prompting.** Two of them asked the study's own frozen questions word for word, so they compare directly against the signed-in series.
- **The site was named, described accurately, and located correctly** with no memory, no account, and no history to draw on.

So the name works cold. The 97% is retrieval, not recall. We would have published the limitation either way, and the result is the less interesting of the two outcomes, but it is the one the data gave us.

The same morning produced the sharpest single comparison in the study. One of those logged-out questions was “Cortexa company”, and it returned nothing, twice. Adding one word (“Cortexa consulting company”) returned us, first, with our own pages cited. Same session type, same device, one day apart. Everything else in this report that separates a hit from a miss also changes a question, a surface or an account. This one changes a single word.



The same site, the same weeks, the same engine. The two empty tracks are not a low score. Nothing was ever cited in them.

02

CORTEXA FIELD STUDIES Nº 01

What the instruments disagree about

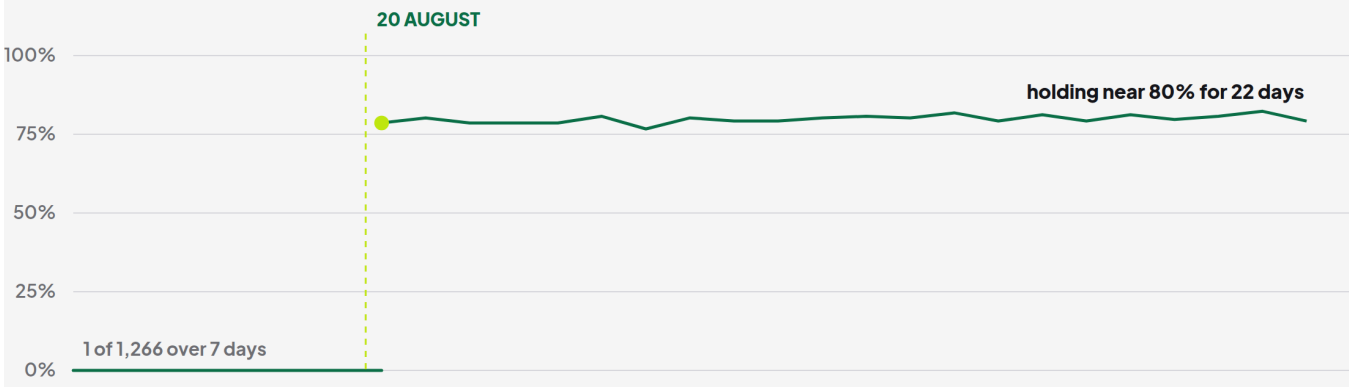
What happens when the vendor changes the rules?

On 20 August the API arm went from 1 citation in 1,266 attempts to a steady 80 percent, and stayed there. Nothing on our side changed that day. No content, no configuration, no new pages. The behavior of the engine changed.

- **Before 20 August: 1 of 1,266 (0.08%).** Seven days.
- **From 20 August: 3,327 of 4,165 (79.9%).** Twenty-two days, holding steady since.

The day the vendor changed the rules

Daily citation rate, one engine through the API. Nothing on our side changed on 20 August: no new pages, no configuration, no content. Pooling the two periods gives 61%, a number that describes neither.



Cortexa Field Studies Nº 01 · the pre-break period is drawn as one segment, annotated with its pooled figure, rather than seven invented points

Pooling the two periods gives 61%, a number that describes neither. The boundary is drawn rather than smoothed for that reason.

Wasn't that just your own work finally taking effect?

It is the first thing we asked too, and it is the better explanation on its face: the site had been submitted, crawled and indexed, and one morning it all switched on. Our own server logs are what rule it out.

The engine's crawlers had been reading the site for two weeks before that morning, and on some days reading it hard: 264 requests on 8 August, 343 on the 13th, 94 on the 18th. Through every one of those days the API cited us once in 1,266 attempts. Then on 20 August the crawlers barely came at all, nine requests in

the whole day, and the rate went to 79 percent. The pages had been fetched and not used, and then they were used on the quietest day of those two weeks.

The off-site records rule it out from both directions. Most of them did not exist yet: the Clutch profile was created on 25 August, five days after the step, and the rest landed through early September. The one that did exist, a Crunchbase profile live since July, had been in place for roughly six weeks by the 20th, through every one of those zero days. Whatever changed that morning, it was neither the records that arrived later nor the one that was already there.

There is one thing we cannot see, and it is worth saying plainly: a crawl is observable and an index is not. So the account we would defend is both at once. Our pages sat in that pipeline for two weeks, and on 20 August something on the engine's side started using them. The timing was theirs; the substance was ours.

A second engine is the cleanest version of that argument, and we had it running the whole time without noticing what it was for. Gemini answered the same 24 questions on the same days through the same runner, and it did not step: 87.5 percent of brand questions cited before 20 August, 85.8 percent after. On the morning one vendor's API went from 1 citation in 1,266 attempts to a steady 80 percent, another engine asking the same things of the same web moved 1.7 points the other way. Whatever happened was that vendor's, not the web's and not ours.

The sub-queries say it from the other side. What the engines asked for did not change on 20 August: brand searches named us 97.5 percent of the time before the break and 98.3 percent after, and the non-brand searches named us on neither side. What changed was what came back.

There is one published benchmark for how long that wait normally runs, and we land inside it. Profound measured about 900 new pages that ChatGPT or Claude cited over a 60-day window in March to May 2026: half were cited within 6.81 days of publication, three quarters within 18.68, nine in ten within 37.10.[7] Our step falls 17 days after the gate lifted, inside their third quartile.

We report that rather than leave a reader to find it, and we are not claiming a match. Read the denominator. Their sample is pages that were cited, so the curve describes how long a successful citation took, not how often one arrives; a page never cited is not in it. That is exactly the difference between their question and ours, because most of what we asked was never cited at all. The shape differs too. A distribution over pages predicts a ramp, and ours moved all eight brand questions on one day and held for twenty-two. Both can be true: the wait was ordinary in length, and the way it ended was not.

> **CORTEXA'S READ**
>
> Which is the part a client should take away. You cannot schedule this. You can only be ready before it happens, because when the engine's side moves it uses whatever it already has. Everything in the list above was in place two weeks early, and that is the only reason there was anything to incorporate.

Pooling those into one rate gives 61 percent, a number that describes neither period and would be the most quotable figure in this paper. We refuse to quote it. Any series that spans 20 August has to be reported as two series, and the tooling that computes these totals now refuses to emit a single pooled rate across the break.

> **CORTEXA'S READ**
>
> If you are paying for a GEO dashboard, this is the number that should worry you. A tenfold change in measured visibility, on the engine's side, in one day, with no action by the site being measured. We cannot tell you whether the same thing happened to every site that morning or only to ours, and that is the point: any report that does not carry a change log cannot tell you whether your agency earned the improvement or it arrived on its own.

So what did move the needle?

One thing did, and it was not content. On the evening of 25 August we created a profile on Clutch, a business-services directory, and filled every descriptive field its free tier allows. The next day the engines began citing it.

- 24 and 25 August: Clutch appears in zero answers.
- **26 August: it appears in thirteen.** One day after the profile existed, across two of that day's three runs.
- **Since then: 121 citations** of that profile in the hand-collected arm alone, and not a single day at zero.

A third-party page describing us, created from scratch, was quotable within 24 hours. Our own articles, published over two months and written to a stricter standard, are at zero on every topic question. Both statements come from the same dataset over the same weeks.

One answer stated the mechanism out loud rather than leaving us to infer it, giving our location and attributing it to the Clutch profile as its only source. That is a single observation and we treat it as one.

Why does the bare name still fail?

“Cortexa” on its own remains unanswerable. In the repeated-measures arm that question is 0 of 680: it has never once produced a citation of our site, in either period, under any variant.

The reason is a crowded namespace, and naming it neutrally is part of the method. Our probes have surfaced a software firm in Washington state, a therapy-practice analytics platform, a consultancy in Hungary, a learning-technology company in the United Kingdom, a dissolved British management consultancy with nearly our exact name, and several others. Eight distinct third-party domains carrying the name have appeared in answers. None of them is doing anything wrong. They were there first, or they are simply also there.

Choosing a contested name costs measurable visibility, and the cost lands specifically on the question a stranger asks first.

We paid that cost twice, and the second time was our own doing. On 10 September, while this study was running, we named the imprint this paper appears under “Cortexa Research”. Sixteen days earlier one of our own control runs had asked the engine to list the Cortexas, and it had returned “Cortexa Research” twice, in two separate lists, describing a real open-research platform that is not us. The entry was already sitting in our namespace file. Nobody connected it to the naming decision until the paper was laid out.

The imprint became Cortexa Field Studies before publication. We report it because it is the clearest illustration in this paper of how the failure happens. Not carelessness about the namespace: we were measuring it daily, in writing, with an instrument built for the purpose. We measured it in one place and named something in another.

Were we crawled, or just ignored?

This is the part most GEO measurement cannot do, and the log-analysis literature concedes it: logs alone will not tell you whether a page influenced an answer. We own the site, so we have both halves. Across 42 days the edge logged 13,354 requests from seventeen named crawlers. Thirteen of those are Artificial Intelligence (AI) crawlers and account for 7,416; the other 5,938 are Googlebot, Bingbot, Applebot and Google's Search Console fetch tool, which are not AI crawlers and are reported separately for that reason.

That produces the finding we would least like to report. There are dated cases where an engine fetched our pages during the same minutes it was answering a question, and then answered without citing them. Crawled, read, and passed over. “Not indexed yet” is a comfortable explanation that our own logs rule out for those runs.

Who else shows up wearing a crawler's name?

A log records what a request says it is, not what it is. The user agent is a line of text the sender chooses. We took that for granted until we looked at what some of this traffic was asking for.

On 13 September, three days after the measurement window closed, the site took 6,187 requests against a normal day of about 1,400, and seven crawlers that had never visited arrived at once. None of it was crawling. The requests were for environment files, SSH private keys, Terraform state and Docker credentials, many with directory traversal in the address. Inside the 42 days the window does cover, 1,912 requests like these arrived on 21 separate days, and that count does not include the 13 September burst. This was not one bad afternoon. It has been going on for most of the study, it is getting broader, and it did not stop when we stopped counting.

All of it was refused. Nearly every request returned an error, and not one request for a secret, a key or a traversal path returned anything at all. There was nothing to find: this site is not a server handing out files from a folder, so the file that scan is hunting for has no web address here. The reason it is worth reporting is the names it used.

- **It called itself GPTBot, ClaudeBot, PerplexityBot and OAI-SearchBot.** Those are the user agents that every site chasing visibility in AI answers is being advised to welcome.
- **We welcome them on purpose,** and we keep bot challenges switched off in front of them, because the entire strategy is to be readable by answer engines.

That is the trade nobody mentions. The names on the allowlist become the cheapest disguise available, because the sites most likely to wave them through are precisely the sites that have stopped checking. We would still make the same call, and we are not recommending anyone close the door on AI crawlers over this. We are saying the door has a cost, and that we only found ours by keeping our own logs.

It costs us a number too, and the honest thing is to say so next to the number. Crawl counts are built from self-declared names. Across these 42 days roughly a third of that traffic carries a user agent our edge could not confirm, and on the two heaviest days the named AI crawlers were almost entirely unconfirmed. Unconfirmed is not the same as fake, since our plan does not label every legitimate request either, and that is exactly why a raw crawl total should not be read as attention. We treat it as a ceiling, and any claim about whether an engine came to read a page is made from the confirmed subset.

The linkage evidence above survives that, for a reason worth stating rather than assuming. Scanners collect refusals; real crawlers collect pages. The cases where an engine fetched our articles and then answered without citing them are all successful requests for real URLs, which is a population a scanner never reaches.

What did the dashboard say?

Zero. Clutch's own analytics panel reported no AI-driven visibility for our profile during a window in which this study counted 23 citations of that exact profile in two days. Both instruments were watching the same page over the same hours and disagreed completely.

We are not accusing anyone of bad arithmetic. The likely explanation is mundane: different denominators, different engines sampled, different definitions of a citation. That is the point. A preprint in this field describes the failure as denominator disagreements disguised as performance disagreements.[4] If two competent instruments can differ by 23 to nothing on one page in two days, a single dashboard number is not a measurement of anything you can act on.

What did we get wrong?

Enough that it belongs in the paper. This study has a findings ledger with a retraction category, and it is not empty.

- **A parser bug turned a perfect run into an empty one.** The consumer product changed its output format mid-study and our reader silently recorded eight citations as zero. Every guard passed. We found it by eye.
- **We pooled the wrong denominators five times.** Each time by computing from a remembered design rather than the filed data. There is now a tool whose only job is to refuse that.
- **One question cannot be scored by matching text.** Two of four failures on the location question name the correct state inside a sentence denying it. A keyword count reads those as successes.
- **A location hypothesis died because we assumed rather than recorded.** Two days of runs were made through a virtual private network nobody wrote down.

We report these because a GEO study that presents only its clean numbers is asking to be believed rather than checked. Every figure above survived a recount against the filed data. Several earlier versions of those same figures did not. And the sub-query finding, the one that explains the headline number, came out of responses we had been filing daily for six weeks without reading the field it was in.

03

CORTEXA FIELD STUDIES Nº 01

What we registered before we knew

Does fixing your entity records help?

We registered that question as a prediction before intervening, on 25 August, and wrote down what would prove us wrong. The hypothesis: engines decide whether you are a candidate answer from structured records about your company, not from the quality of your pages. The intervention corrected our location across five profiles, created the Clutch record, added a Wikidata item and set a business-identifier record that names our domain.

The pre-registered test is whether four cold asks of the bare name move off a baseline of zero. Three confounds were recorded in advance: the vendor regime break of 20 August, ten weeks of company age as a large proportional change, and the possibility that a record is deleted rather than answered. The widest-reaching record did not become effective until 8 September, seven days before the read, so a null result will be partly explained by non-delivery rather than only by the hypothesis being wrong. We wrote that down before knowing the answer, and a second read is already registered for 15 October.

> CORTEXA'S READ
>
> The read happened on 15 September and it is a null. Four cold asks of the bare name, zero citations, which is exactly where the baseline was. The secondary indicator did not move either: that question is 0 of 680 across the whole series and has returned zero on every single day of it. Publishing this is the commitment we made on 27 August, before the data existed, and the confound we registered in advance still stands. The widest-reaching record became effective seven days before the read, so part of this null is a record that had not propagated rather than a hypothesis that is wrong. The 15 October reading is what separates those two, and we do not get to choose which one it finds.

What would we tell a client?

- **Being cited by name is not visibility.** It is the easy half, and a new site can win it in weeks.
- **Topic citations are the hard half** and good content did not buy them in five weeks. Anyone promising otherwise on a new domain is selling something we could not produce with a strict evidence standard and daily measurement.
- **Third-party records moved faster than anything we wrote.** A Clutch profile was quotable in a day.
- **Check your name before you buy the domain.** A contested name has a measurable cost on the first question anyone asks.

- **Demand a change log from any GEO vendor.** A tenfold swing happened on the vendor's side in one day, and no dashboard we saw mentioned it.

What this study cannot tell you

One site, one sector, one engine family, five weeks. The 90-day design is not finished and the second pre-registered read is a month out. Everything here describes a business-to-business consultancy in a crowded namespace, and a consumer brand with an uncontested name would very likely produce different numbers. We have no control site, because we have only one domain that was new in July.

The measurement window runs from 7 August to 10 September, which is the window the design specified before any data existed. The automated arms stopped there. They ran daily throughout rather than dropping to twice weekly as the schedule allowed, so the window holds 35 days of collection against the 20 the plan budgeted for. Everything reported here is drawn from inside it. Collection continues for the 90-day design and the second pre-registered read, and neither is in this paper.

We can see the searches an engine ran and not the pages it weighed against each other. Selection is the half of the mechanism this study does not reach, and it is the next thing we would measure. The third engine does not expose its searches at all, so its share of the zero is outside even the half we can see.

What we can offer is the thing the aggregate studies cannot: dated failures, both halves of the linkage, a pre-registration published before its result, and a set of numbers that can be recounted. Every figure names the rule it was counted under. That includes the one judgment call in the hand-collected arm: on 1 September a run failed outright and was collected again, and the figure keeps the failure and discards the second attempt. Counted the other way it reads 99.1%. A rerun that is allowed to replace a failure means the arm only ever reports its good days.

Sources

1. Schulte et al., “Don’t Measure Once: Measuring Visibility in AI Search (GEO)” (preprint, 2026)
   https://arxiv.org/abs/2604.07585
2. Sielinski, “Quantifying Uncertainty in AI Visibility: A Statistical Framework for Generative Search Measurement” (preprint, 2026)
   https://arxiv.org/abs/2603.08924
3. “Optimizing Visibility in Generative Engines: A Critical Survey of Generative Engine Optimization (2023–2026)” (preprint)
   https://arxiv.org/abs/2607.14035
4. “Generative Engine Optimization at Scale: Measuring Brand Visibility Across AI Search Engines” (preprint)
   https://arxiv.org/abs/2606.20065
5. “From Stochastic to Stable: Rank Stability and Structural Sufficiency in AI Visibility Measurement” (preprint)
   https://arxiv.org/abs/2607.10341
   Background on why single-shot visibility rankings are unstable; not cited inline.
6. OpenAI — Temporary Chat FAQ (Help Center)
   https://help.openai.com/en/articles/8914046-temporary-chat-faq
   A temporary chat starts unpersonalized and uses no memory; a personalized one, added at the end of August 2026, can read saved memories. Neither writes new ones while the chat stays temporary.
7. Blyskal (Profound) — “How long does it take Claude and ChatGPT to cite a new page?” · n ≈ 900 new pages cited by ChatGPT or Claude agents over a 60-day window, agent log behavior, March–May 2026
   https://www.linkedin.com/posts/joshua-blyskal_how-long-does-it-take-to-get-cited-in-chatgpt-activity-7459597424102223873-Ywuj
   Industry measurement, not peer-reviewed; figures verified against the source. The sample is pages that WERE cited, so the percentiles describe how long a successful citation took, not how likely one is.
8. DEJAN — “Our OpenAI fanout data shows no rise in site: operator use” (196,692 API fan-out queries)
   https://dejan.ai/blog/site-fanouts/
   Figures verified against the source. Read alongside Promptwatch’s consumer-interface dataset, which recorded the opposite: promptwatch.com/data/chatgpt-site-operator-fanouts. The disagreement is the finding.
9. elmoHQ — ChatGPT interface citation share against the same vendor’s API, August 2026
   https://www.elmohq.com/blog/chatgpt-ui-reddit-citations-api
   Industry measurement, not peer-reviewed; figures verified against the source.

Appendix · how every figure was captured

Most of the numbers in this paper are our own readings rather than someone else's published figures, and on the page the two look identical. This is the difference. Below are the questions we asked and when we froze them, every headline number beside the command that reproduces it, what each collection day looked like, and the counting decisions the numbers rest on.

1 · The questions, fixed before anything was measured

Written and frozen before the first reading, and not edited since. A query set you tune while measuring is a query set that measures your tuning. The checksums are there so you can tell whether the set below is the one that produced the numbers.

Question set	24 questions, frozen 2026-08-07 · sha256 `f2cd9687db2d44b58930d3a690c30c87…`
Wording variants	3 per question, frozen 2026-08-09 · sha256 `f7611854b31af3e7eaa97fcfd03a6213…`
Measured	2026-08-07 to 2026-09-10 · the engine changed on 2026-08-20, and nothing is pooled across it

ID	ASKS ABOUT	THE QUESTION, WORD FOR WORD
D01	Names the company	Who is Cortexa Consulting?
D02	Names the company	What does Cortexa Consulting do?
D03	Names the company	Cortexa
D04	Names the company	Cortexa Consulting services
D05	Names the company	Is Cortexa Consulting a marketing agency or a technology consultancy?
D06	Names the company	Who founded Cortexa Consulting?
D07	Names the company	Cortexa Consulting North Carolina
D08	Names the company	What is Cortexa Ground Truth?
T01	Topic	What does agentic AI actually mean?
T02	Topic	Why does AI make things up?
T03	Topic	What is RAG in simple terms?
T04	Topic	Should we fine-tune a model or use RAG?
T05	Topic	What is a token in AI and why does it cost money?
T06	Topic	Is prompt engineering a real skill?
T07	Topic	Why do AI demos fail in production?
T08	Topic	What are evals in AI?
T09	Topic	What does multimodal AI mean for marketing?
T10	Topic	Which AI agents actually work in production?
T11	Topic	What is MCP, the Model Context Protocol?
T12	Topic	What is a vector database and why would I need one?
C01	Buying intent	AI consulting for marketing agencies
C02	Buying intent	market research for technology clients
C03	Buying intent	white label AI consulting for agencies
C04	Buying intent	who does AI market research for tech brands

2 · Every figure, and the command that reproduces it

Run the command against the filed readings and you should get the figure beside it. The second line under each command is the decision that number rests on, which is the part no tool can tell you.

THE FIGURE IN THE PAPER	HOW TO REPRODUCE IT
444 of 456 · 97.4% consumer ChatGPT, asked by name	`node tools/geo-probe/derive.mjs` From 20 August only. Cold reads excluded by marker, not by glob. The 2026-09-01 failed run is KEPT and its rerun discarded.
3,327 of 4,165 · 79.9% the same eight questions through the API	`node tools/geo-probe/rate-report.mjs` From 20 August only. Summed from the cells actually collected, never 192 × days — three days are partial.
1 of 1,266 · 0.08% the same arm before the break	`node tools/geo-probe/rate-report.mjs` Reported as its own series. The tool refuses to emit a pooled rate across 20 August; pooling reads 61%, which describes neither period.
0 of 1,248 · 0 of 416 · 0 of 1,664 topic, buying-intent, and the two combined	`node tools/geo-probe/derive.mjs` Whole window, both regimes, three engines. Zero in each regime separately, which is why these are not split at 20 August.
13,354 from 17 named crawlers · 7,416 from 13 AI crawlers the edge log	`node tools/geo-probe/crawl-report.mjs` 2026-07-31..09-10, the crawl arm's own 42 days. Googlebot, Bingbot, Applebot and the Search Console fetch tool are counted and reported separately; they are not AI crawlers.
0 of 853 non-brand sub-queries · 1,448 of 1,465 brand what the engines searched for	`node tools/geo-probe/fanout-report.mjs` Wave arm, two engines. The third exposes no sub-queries at all. Read off responses already filed; no new collection.
every figure above the paper against the data	`node tools/research-paper/verify-figures.mjs --strict` Diffs each printed figure against a fresh derivation and fails on drift. A figure it cannot check is reported UNVERIFIED, which also fails.

3 · Every collection day, counted

Answers that cited us, over questions asked, for each of the 35 days. The last two columns are the paper's headline: a zero that holds for 35 consecutive days is a different claim from a zero that is an average.

DATE	NAMES THE COMPANY	TOPIC	BUYING INTENT
2026-08-07	14/24	0/36	0/12
2026-08-08	14/24	0/36	0/12
2026-08-09	14/24	0/36	0/12
2026-08-10	14/24	0/36	0/12
2026-08-11	14/24	0/36	0/12
2026-08-12	14/24	0/36	0/12
2026-08-13	14/24	0/36	0/12
2026-08-14	14/24	0/36	0/12
2026-08-15	14/24	0/36	0/12
2026-08-16	14/24	0/36	0/12
2026-08-17	14/24	0/36	0/12
2026-08-18	14/16	0/24	0/8
2026-08-19	14/24	0/36	0/12
2026-08-20	18/24	0/36	0/12
2026-08-21	19/24	0/36	0/12

DATE	NAMES THE COMPANY	TOPIC	BUYING INTENT
2026-08-22	19/24	0/36	0/12
2026-08-23	19/24	0/36	0/12
2026-08-24	20/24	0/36	0/12
2026-08-25	20/24	0/36	0/12
2026-08-26	19/24	0/36	0/12
2026-08-27	20/24	0/36	0/12
2026-08-28	19/24	0/36	0/12
2026-08-29	19/24	0/36	0/12
2026-08-30	19/24	0/36	0/12
2026-08-31	19/24	0/36	0/12
2026-09-01	19/24	0/36	0/12
2026-09-02	20/24	0/36	0/12
2026-09-03	20/24	0/36	0/12
2026-09-04	20/24	0/36	0/12
2026-09-05	17/24	0/36	0/12
2026-09-06	19/24	0/36	0/12
2026-09-07	19/24	0/36	0/12
2026-09-08	18/24	0/36	0/12
2026-09-09	18/24	0/36	0/12
2026-09-10	18/24	0/36	0/12

4 · The counting decisions behind those numbers

Judgment calls, not arithmetic. Re-derive these figures without knowing them and you will get different answers that are also defensible, which is exactly why they are written down rather than remembered.

1. The 2026-09-01 variant-B run failed whole-run, 0 of 8, and was collected again the same day. The figure KEEPS THE FAILURE and DISCARDS THE RERUN. Keeping both reads 452 of 464; dropping the original reads 452 of 456. All three round near 97%, which is why the choice has to be written down rather than rediscovered.
2. Rows carrying `control` or `excluded_from_series` are filtered before anything is counted. The P-ENTITY-01 cold reads carry NEITHER marker on some days and are separated only by `d03cold` in the filename, so a glob over the manual directory eats them.
3. A citation means an answer links a page on our own domain. Naming the company is not a citation. Being crawled is not a citation. Ranking is not a citation.
4. Do not score a citation by string-matching a state name: four D07 misses NAME the correct state inside a sentence denying it.
5. Nothing is pooled across 20 August. The two periods are different instruments by measurement, and the pooled figure would be the most quotable number in the paper.
6. The wave arm and the rate arm are never pooled with each other either. One is a daily cross-section, the other a repeated-measures rate estimate.

WHAT WE SAID BEFORE WE KNEW

A second reading is already booked, and its result is not ours to choose.

The prediction, the date, and what would prove it wrong were written down and published before the intervention. Nothing below was added after seeing a number.

2026-08-25 — PREDICTION REGISTERED
the hypothesis, the falsifier, and the 0-of-4 baseline

2026-09-15 — FIRST READING
reported as found, including a null

2026-10-15 — SECOND READING
registered in advance, to survive a propagation delay

HOW TO CHECK US

The question set was frozen on 7 August and pinned by checksum before any of this was measured, and every figure names the rule it was counted under. The falsifier is stated: if the non-brand strata are still empty on 15 October, the intervention did not work.

ABOUT CORTEXA FIELD STUDIES

Cortexa Field Studies is where Cortexa Consulting publishes its own measurement. We run the instrument on our own property first, publish the method before the result, and say plainly when a reading comes back empty.

Intelligence, applied.