The Absence Manual

A technical manual on AI search visibility for B2B software.

Sairam Sivakumar, Broadcastwell
Version 1.0, August 2026
Built 2026-08-18

This PDF is a mirror. It is generated from the same source as the website, in the same build, so the two cannot drift. It is free and ungated: there is no form, no email field and no login anywhere in it or on the site.

Contents

The Absence Manual

What this is

The Absence Manual is a technical manual on AI search visibility for B2B software: how to measure how often an AI answer names your company, how to read what that measurement does and does not tell you, and what the evidence actually supports doing about it. Every finding in it is drawn from three already-published studies with DOIs, open data and open analysis code. It is free, ungated and permanently online. There is no form, no login, no paywall and no email field anywhere on this site, because an AI crawler issues a plain HTTP request and cannot fill in a form. A gated manual about citability would refute itself.

Author. Sairam Sivakumar, Broadcastwell.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

What it is built from

Volume	What it covers	DOI
Volume I	85 B2B software companies, 61 categories, 860 scored answers and 5,160 traced citations, one engine held constant. The dataset README additionally records 1,753 unique domains cited	10.5281/zenodo.21537014
Volume II	The Absence Ladder. All 616 absence records classified by question shape	10.5281/zenodo.21586091
Volume III	Cross-engine divergence. 280 questions, 40 categories, four engines, 853 answers collected 05:28 to 16:29 UTC on 2026-08-04, with a 25-question repeat subsample run three times	10.5281/zenodo.21789120

All three are CC BY 4.0, with their data and analysis code at github.com/Broadcastwell/state-of-geo-2026.

How to read the evidence labels

Every substantive claim in this manual carries one of three labels, so you always know what kind of thing you are being told. This is the manual's central discipline and it is applied without exception.

Measured. The claim comes from one of the three volumes, and the volume is named in the same sentence or the label. You can recompute it from the published data.

Reported. The claim comes from a source outside this research programme, cited with its publisher and date. You should check it at the source.

Reasoned. The claim is an inference drawn from the measured or reported material above it. It is argument, not measurement, and it is marked so you can disagree with the reasoning without disputing the data.

A claim with no label is a defect. If you find one, it is a mistake and it will be corrected.

How to read it

Every chapter has its own permanent URL and stands alone. Statistics carry their sample size in the sentence that states them. Limitations are stated in the body of the argument rather than collected at the end where they can be skipped. Where the evidence is strong the claim is unhedged, and where it is thin the chapter says so and names the base it rests on. Where a number is missing or two published sources disagree, the manual flags the gap rather than filling it with a plausible value.

Chapters

Appendices

Every chapter and appendix is published. A slug published once is permanent: version numbers change, URLs never do. See the changelog.

The whole manual as one PDF

Download the complete manual as a PDF. Every chapter, every appendix, the source table with its DOIs and both dated self-audit lines, in one file. It is generated from this same source in the same build, so the PDF and the pages cannot drift apart. There is no form, no email field and no login: the link downloads the file. That address always serves the current version, and the versioned filename is preserved alongside it.

For marketing leaders

If you are not going to read a technical manual, start with what AI search is doing to your category. Plain language, no jargon, one page.

Notes

Short standalone pieces cut from the chapters, each linking back to its parent.

- Same score, different problem, from Chapter 2.
- Half its own shortlist, from Chapter 6.

Licence and status

Prose and figures are published under CC BY 4.0. The site code is MIT. Nothing here has been through academic peer review: it is measurement published with the data and code that produced it, and the standing invitation is to recompute any figure you doubt from the public files and publish what you get. In the four-engine, five-run self-audit published alongside Volume II in July 2026, Broadcastwell was named in 0 of 200 answers and cited 0 times among 663 citations. Re-measured on the same ten questions and the same four engines on 18 August 2026, at one run per question rather than five, it was named in 1 of 40 answers and cited once among 674 citations. Both figures stand with their dates and neither replaces the other.

Version 1.0, August 2026.

For marketing leaders: what AI search is doing to your category

This page is the short version. It has no jargon and only the handful of numbers the argument needs. If you want the full technical treatment, the rest of this manual is free and there is no form anywhere on it.

Your buyers are asking a machine first

A meaningful share of the research that used to happen on a results page now happens inside a generated answer. Someone types a question about your category, and an AI product replies with a short list of vendors and a paragraph about each. That reply is the shortlist. For many buyers it is the only shortlist they will see before they start booking calls.

The important thing about that reply is how short it is. In our published research across 860 scored AI answers, the average answer named about two vendors. Not ten. Two.

Which means most companies are simply missing

When there are only two places, most companies get none. Across 85 B2B software companies we measured, 35% were never named once in ten questions about their own category. Not ranked low. Named zero times.

That is the first thing to understand about this channel: it is not a leaderboard where you are somewhere in the middle. It is a door you are either through or not through. Being absent is the normal condition, not the exceptional one.

There are two doors, not one

This is the part that changes what you should do, and almost nobody separates them.

The first door is whether the engine thinks you are one of the companies in your category at all. When a buyer asks "who are the best tools for X", the engine assembles a handful of candidates before it writes a word. If you are not among the candidates, nothing else matters. In our data, companies that were never named lost mostly these questions: broad ones, that do not mention a competitor by name, that simply ask who exists.

The second door is whether the engine prefers you when it is comparing you with someone specific. Companies that were named in most answers had cleared the first door. What they still lost was almost entirely head-to-head questions of the form "you versus them".

These need different work. Getting through the first door is about whether credible information about you exists in enough places for a machine to conclude that you belong in the category. Getting through the second is about whether there is something specific and checkable it can quote about you against a named rival. If you are stuck at the first door and someone sells you work aimed at the second, you will pay for content very few people will ever see.

Why the number you were shown may not mean what you think

If an agency or a tool has already given you an AI visibility score, three things are worth knowing before you act on it.

It probably covers one AI product, not all of them. These systems disagree with each other much more than people expect. Of 32 companies we tested on more than one engine, 14 were visible on some and invisible on others. Almost half would get a different verdict depending on which one happened to be measured. A single number labelled "AI search" is usually a number about one product.

It was probably measured once. Ask the same AI product the identical question twice, minutes apart, and it agrees with only about half of its own previous list. That is not a fault, it is how these systems work. But it means a single measurement is one draw from a range, and the difference between your score in March and your score in June can easily be the tool rather than your company.

On some questions there was no answer at all. Google does not generate an AI overview for every question. When it does not, there is nothing to appear in. If those questions were quietly dropped from the calculation, the percentage you were shown is bigger than it should be.

None of this means measurement is pointless. It means a number without its engine, its date, its questions and its number of runs cannot be interpreted, and most numbers in this market arrive without them.

What to do about it, in order

First, find out which door you are at. You do not need to buy anything to do this. Write down the ten questions a real buyer in your category would type. Put each one into an AI product with web search on. Write down only whether your company was named. Then look at the questions where you were not named and sort them into two piles: broad questions about who exists in the category, and questions naming you against a specific competitor.

Whichever pile is bigger tells you which door you are standing at. That is the whole diagnostic and it costs you an afternoon.

Second, be sceptical of anything sold before that diagnosis. A proposal that recommends the same programme regardless of which pile is bigger has not looked at your situation. Ask a supplier which door your data says you are at, and ask to see the questions that support the answer.

Third, ask what will exist when the work is finished. Not how many pages will be published, but what credible information about your company will exist in places you do not own. That is what the first door responds to, and it is the question most proposals answer least clearly.

Fourth, insist on being able to check the work later. Fix the question set on day one and never change it. Require that each question is asked more than once every time it is measured. Ask for the raw answers, not just a dashboard. All three are cheap at the start and impossible to add afterwards.

An honest note about who wrote this

Broadcastwell ran the research this page is based on and sells services in the category it measures. That is a real conflict and the only useful response to it is to publish everything, which we have: all the data, all the analysis code, and the questions. If you think a number here is wrong, you can recompute it and publish what you get.

We also measure ourselves on the same terms. In our own most recent check, across four AI products and ten questions about our own category, we were named once in forty answers. The one mention was an engine quoting our research while recommending other agencies. We publish that because a firm asking you to demand evidence should be willing to be measured by it.

Where to go next

If you are evaluating suppliers or tools, Chapter 12: How to buy GEO without getting sold a number is the practical companion to this page. It has twelve questions to put to any supplier, a description of what a defensible deliverable looks like, and a section on when the right decision is to hire nobody at all.

If you want the underlying research, the manual itself is free, ungated and has no email capture anywhere on it. It is one chapter per page and every number in it traces back to a published dataset you can download.

Chapter 0. Selection, not ranking

By the end of this chapter you will know why AI search is a selection problem rather than a ranking problem, what that changes about the questions worth asking, and why almost every metric imported from search engine optimisation answers a question the system is no longer asking.

The claim this chapter defends

Measured, Volume I. Across 860 scored answers, the average AI answer named 2.05 vendors. **Reasoned.** A results page with ten blue links is an ordering of a long list; an answer that names two vendors is a selection from a candidate set. Those are different mechanisms, and the difference is not cosmetic. Ordering has a position you can improve. Selection has a boundary you are inside or outside of. Every metric, every diagnosis and every remedy that a marketer imports from search engine optimisation assumes the first mechanism, and this chapter is about what happens when you apply it to the second.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

The number that breaks the old model

Measured, Volume I. The average answer named 2.05 vendors. The sweep covered 85 B2B software companies across 61 categories, put to one AI engine with live web search. Volume I reports 860 scored answers in total. **Reasoned.** Two point zero five is not a shorter results page. It is a different object. A page of ten links can carry the market leader, three challengers, two review sites and a comparison blog, and the buyer chooses among them. An answer naming two vendors has already chosen. The visible surface is not a ranked list that has been truncated. It is a decision that has been made and then narrated.

Ranking asks where, selection asks whether

Reasoned. Under a ranking model the operative question is positional: where do we sit, and what moves us up. Position implies a continuum, so effort maps onto small monotonic gains, and a bad position is a recoverable one. Under a selection model the operative question is binary at the first stage: are we in the candidate set the engine assembles before it writes anything. If the answer is no, position is undefined. There is no eighth place to climb out of. You are not ranked badly. You are not present, and presence and position are different states requiring different work.

What the distribution actually looks like

Measured, Volume I. The median company was named in 20% of the AI answers for its own category, and 35% were named in zero. **Measured, Volume I.** The median category leader appeared in 80% of its category's answers. **Reasoned.** Those two figures together describe a winner-take-most shape rather than a gradient. A gradient would put most

companies in the middle with thin tails. What the data shows instead is a large group at or near zero, a small group near saturation, and comparatively little in between. Selection systems produce that shape. Ranking systems, which have to place everybody somewhere, do not.

The sample this rests on, stated plainly

Measured, Volume I. The 85 companies were selected as plausible challenger brands rather than sampled at random, so the average named rate is lower than a random draw of all vendors would produce. Volume I says so in its own limitations. **Reasoned.** That skew matters for the level of the numbers and not for the shape of them. A challenger-skewed sample cannot tell you the industry base rate for zero visibility. It can tell you that within a population of plausible challengers, a large fraction are absent entirely while leaders sit near saturation, and that is the observation the selection model rests on.

Why the shape follows from the mechanism

Reasoned. If an answer has room for roughly two vendors and the engine assembles its candidates from sources that already agree about who the players are, then the same well-corroborated names will be selected repeatedly and everyone else will be selected almost never. Scarcity of slots plus consensus in the evidence layer produces concentration mechanically. This is a reading of the measured distribution rather than a separate finding, and it is offered as the reason the numbers look the way they do, not as evidence for itself.

Different engines make different-sized selections

Measured, Volume III. Mean vendors named per answer ran to 4.77 for Google AI Overviews, 8.83 for Claude, 10.69 for Perplexity and 14.72 for ChatGPT. **Measured, Volume III.** Volume III states directly that these are not comparable like for like with Volume I's 2.05, because its extraction layer is wider: Volume I counted only companies inside the study universe, while Volume III also captures vendors outside it. **Reasoned.** The comparison that does survive is relative. Whatever the counting rule, a Google AI Overview selects a materially smaller set than a conversational engine does, so a vendor competing for a slot in one is competing for a scarcer slot than in the other.

Scarcity is not the same everywhere

Reasoned. A vendor with a fixed amount of evidence behind it is not equally likely to be selected across engines, because the number of slots differs before any question of merit arises. That is a structural disadvantage on the shortest-list engine and a structural opportunity on the longest-list one. It also means a single-engine score conflates two things, how good your evidence is and how many slots the engine you happened to measure was handing out. Chapter 5 at /engine-divergence/ takes that apart with the cross-engine data.

Selection is unstable in a way ranking is not

Measured, Volume III. Asked the identical question again in the same collection window, Google AI Overviews agreed with roughly half of its own previous vendor set, at a mean agreement of 0.499 over 62 repeat pairs, and Claude at 0.442 over 74 repeat pairs.

Reasoned. A ranking is a stored artefact that changes when the index changes. A selection is generated at request time, so it can differ between two identical requests with nothing having changed in the world. That difference has no analogue in classical search results and it is why single-run measurement is a weaker instrument here than it was there. Chapter 6 at /measurement-noise/ measures it.

Sometimes there is no answer to be selected into

Measured, Volume III. Google returned an AI Overview for 92.1% of these buyer questions, 258 of 280. On the remaining 22 there was no AI answer at all. **Reasoned.** Under a ranking model there is always a results page, so the denominator is stable. Under a selection model the surface itself is conditional: on some questions the engine declines to generate, and a vendor cannot be present in an answer that does not exist. That makes the denominator a variable rather than a constant, which is a problem for anybody quoting a percentage. Chapter 7 at /overview-trigger-rate/ works through what it does to a score.

The engines do not select the same names

Measured, Volume III. Across 142 questions answered by all of Claude, Perplexity and Google AI Overviews, 2,238 distinct vendor mentions were recorded, of which 63.3% came from a single engine and 17.5% from all three. **Reasoned.** If selection were a close proxy for some underlying quality ranking, independent systems reading a shared public evidence layer would converge much harder than that. A near two-thirds single-engine share says the candidate sets are substantially engine-specific, which is what you would expect from four different retrieval stacks selecting under four different constraints, and not what you would expect from four measurements of one ranking.

Rank tracking has no analogue here

Reasoned. The daily rank check is the load-bearing habit of search engine optimisation, and it does not port. There is no position to track, the surface is regenerated per request, the set size varies by engine, and on some questions the surface does not appear. What can be tracked is presence and absence across a fixed question set, on named engines, over repeated runs, with the variance reported. That is a different instrument with a different cadence, and treating it like rank tracking produces a chart that moves for reasons which have nothing to do with the vendor.

Link volume is the wrong lever to reach for first

Reported. Aggarwal and colleagues showed that content-side interventions do move visibility inside a generative engine, in work introducing generative engine optimisation as a problem (arXiv:2311.09735, KDD 2024). **Reasoned.** That establishes the surface is movable, not that the classical levers are the ones that move it. Under selection, the binding question at the first stage is whether an engine can assemble you into a candidate set from evidence it can reach, which is a question about entity presence and third-party corroboration rather than about how many links point at a page. Chapter 9 at /category-door/ sets out what that evidence has to look like.

Two gates, which is the model this manual uses instead

Measured, Volume II. Across 616 absence records from 85 companies, the shape of what a company loses changes with its visibility: companies named in zero of ten answers lost category-level questions 55.3% of the time and head-to-head comparisons 20.0% of the time, while companies named seven or more times reversed that to 12.5% and 68.8%, with chi-square 61.8, df 15, $p = 1.3 \times 10^{-7}$. **Reasoned.** That is a selection model with two boundaries rather than one. The first is admission to the candidate set. The second is preference inside it. Chapter 2 at /absence-ladder/ is the full argument.

Why two gates beats one number

Reasoned. A single visibility percentage is a projection of a two-dimensional position onto one axis, and projections lose information. Two companies scoring identically can sit at different gates, and the work each needs is different in kind rather than in degree. The absence mix recovers the lost dimension at no cost, because it uses data any measurement already has: the questions you were not named in, kept with their text. The reason the industry reports one number instead is that one number is easier to sell, not that it is more informative.

What the incumbency literature adds, and what it does not

Reported. Chu and Hou document an incumbent advantage in large language model product recommendation, in a single consumer category across three model interfaces rather than in production search products (arXiv:2606.17443). **Measured, Volume II.** In this dataset the related question came out null: the Spearman correlation between category leader visibility and challenger visibility was -0.051 at $p = 0.64$ over 85 companies. **Reasoned.** Incumbents being advantaged and challengers being suppressed by incumbents are different claims. The first is about who gets selected. The second is about crowding out, and the crowding-out version does not appear in this sample.

The three questions the selection model makes worth asking

Reasoned. First, on which questions are we absent, and what do those questions have in common. Second, on which engines, since the candidate sets differ and a single-engine answer generalises poorly. Third, how stable is any of this, since a selection generated at request time carries run-to-run variance a single measurement cannot show. None of these three has a natural expression in the ranking model, which is the practical reason the ranking model has to be put down before anything useful can be measured.

What this model does not claim

Reasoned. It does not claim that classical search work is wasted, that links stopped mattering, or that ranking systems have disappeared from the buyer journey. It does not claim that being selected is the same as being bought. And it does not claim that the mechanism inside any engine is literally a two-stage filter; nobody outside the providers can observe that. The claim is narrower: the measured behaviour is better described by selection from a candidate set than by ordering of a full list, and the questions that description generates are better questions.

The honest limitation of this framing

Measured, Volume I and Volume II. The evidence behind this chapter is one engine with one run per question, on a challenger-skewed sample of 85 companies, plus a four-engine single-window collection in Volume III. **Reasoned.** That supports the direction of the argument and not a precise parameterisation of it. Nothing here has been tested as an intervention: no company in these datasets was measured, changed and measured again. Read the selection model as the frame that fits the observations, and hold it loosely enough to drop it if a larger or longitudinal dataset fits better.

What this means for your buying decision

Reasoned. Ask any supplier which of the two gates your absence data says you are at, and whether they can show you the question-level absence list that supports the answer. A supplier who answers only with a percentage is still working in the ranking model and will prescribe accordingly. A supplier who cannot produce the absence list does not have the data to diagnose you. If the diagnosis is not framed as presence in a candidate set on named engines, you are being sold a rank-tracking habit under a new name. Chapter 12 at /how-to-buy-geo/ turns this into the specific questions to put to a vendor.

Where to go next

Reasoned. If you want the diagnostic model in full, read Chapter 2 at /absence-ladder/. If you want the size of the problem before the model, read Chapter 1 at /named-zero-times/, which sets out how many companies are never named at all. If you want to know how much any single measurement can be trusted, read Chapter 6 at /measurement-noise/.

Sources

Volume I supplies the 2.05 vendors per answer, the 20% median, the 35% zero rate and the 80% leader median. Volume II supplies the absence ladder table, its chi-square and the leader null result. Volume III supplies the per-engine vendor counts, the repeat agreement figures, the 92.1% trigger rate and the consensus shares. External work is cited by author and identifier where it appears, and is reproduced from Volume III's reference list. Everything is at github.com/Broadcastwell/state-of-geo-2026.

Chapter 1. One in three vendors is never named

By the end of this chapter you will know how many companies in a challenger-skewed sample of B2B software were named zero times by an AI engine answering questions about their own category, how far the leaders sit from them, and which of those figures you are entitled to quote outside this sample.

The claim this chapter defends

Measured, Volume I. Across 860 scored AI answers, 35% of the companies measured were named in zero of the answers for their own category, and the median company was named in 20% of them. **Measured, Volume I.** Over the same sweep the median category leader appeared in 80% of its category's answers. **Reasoned.** Those three figures together describe a market in which absence is the common condition and saturation is the exception, and they are the reason the rest of this manual is organised around absence rather than around position.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

What was measured

Measured, Volume I. Between 18 and 23 July 2026, ten standardised buyer questions were put to one AI engine with live web search enabled, for each of 85 B2B software companies across 61 categories. Each answer was scored on two binary outcomes: whether the company was named, and whether its own domain was cited as a source. Every cited URL was logged. The engine was held constant for the entire sweep, which buys internal comparability across categories at the cost of saying nothing about any other engine.

The scored corpus

Measured, Volume I. The sweep produced 860 scored AI answers and traced 5,160 source citations. **Reasoned.** Two binary outcomes per answer is a deliberately austere instrument. It records no sentiment, no ordering, no prominence within the answer and no wording. That austerity is what makes it reproducible: two people scoring the same answer will agree on whether a brand string appears far more often than they will agree on how favourably it appeared. The cost is that the dataset cannot say anything about how a company was described, only whether it was there.

One in three, named zero times

Measured, Volume I. 35% of the companies measured were named in zero AI answers for their own category. **Reasoned.** Zero is a qualitatively different result from a low score. A company named once in ten has demonstrated that the engine can reach it and will sometimes select it. A company named zero times in ten has produced no evidence that it is

in the candidate set at all, on any of the ten questions its own buyers would ask. The remedy for a low score and the remedy for a zero are not the same remedy, which is the argument Chapter 2 at /absence-ladder/ develops in full.

The median is 20%, which is also low

Measured, Volume I. The median company was named in 20% of the AI answers for its own category. **Reasoned.** The median company is therefore absent from four answers in five. Because the median sits at 20% while more than a third of the sample sits at zero, the distribution is compressed against the floor rather than spread evenly, and an arithmetic mean would be a poor summary of it. Quote the median and the zero rate together, because either one alone gives a misleading picture of where a typical company in this sample sits.

The leaders are somewhere else entirely

Measured, Volume I. The median category leader appeared in 80% of its category's answers. **Reasoned.** Set that against the 20% median for the measured companies and the shape becomes clear. This is not a market where everybody gets a slice proportional to merit. It is a market with a small saturated group and a large absent group, and the distance between the two is most of the range. Volume I describes the pattern as winner-take-most, and the phrase is doing accurate work rather than rhetorical work.

Winner-take-most is a consequence of scarcity

Measured, Volume I. The average answer named 2.05 vendors. **Reasoned.** If an answer has room for roughly two names, then a leader appearing in 80% of answers is consuming a large share of a very small total supply of slots. The 80% figure and the 2.05 figure are not two findings. They are the same finding seen from two directions: slots are scarce, and the same names keep taking them. Any strategy premised on gradually accumulating share in a market like that has to explain where the marginal slot is coming from.

The evidence layer underneath is fragmented

Measured, Volume I. The top 10 cited domains hold only 12% of all citations, and 56% of cited domains appear exactly once. **Reasoned.** That is the opposite of the concentration seen in the vendor names, and the contrast is the interesting part. A small set of vendors is named repeatedly, but the sources the engine reaches for to support those answers are spread thin across a long tail. Concentrated outputs drawn from a fragmented input layer is a specific structure, and Chapter 4 at /who-gets-cited/ works through what it implies about where evidence has to exist.

Being named and being cited are scored separately

Measured, Volume I. Each answer carried two independent binary scores, one for whether the brand was named and one for whether the company's domain was cited as a source. **Reasoned.** Keeping them separate is a design decision with consequences, because it makes it possible to observe a company whose pages are quoted as evidence in an answer

that recommends somebody else. That case is common enough to matter and is treated in Chapter 3 at /cited-not-recommended/. A measurement that collapses the two into one visibility number cannot see it.

The most-named brands cross category boundaries

Measured, Volume I. The published top-fifty brand file records, for every brand named across the sweep, its total mentions, the number of distinct category sweeps it appeared in at least once, and its share of all answers. The most-named brand, 6sense, carried 123 mentions, appeared in 14 distinct category sweeps, and was named in 14.3% of all answers. The second, Demandbase, carried 101 mentions across 13 sweeps at 11.7%. **Reasoned.** A brand named in fourteen different category sweeps is being selected well outside the category it was queried for, which means the competition for a slot is not confined to the vendors a buyer would consider adjacent.

What that does to a challenger's arithmetic

Measured, Volume I. The average answer named 2.05 vendors, and a single brand took 14.3% of all answers. **Reasoned.** Slots are consumed by names that travel across categories, so a challenger is not competing only against the other vendors in its own market. It is competing against whichever broadly-established names the engine reaches for when it is assembling a set, including names a category-restricted competitive analysis would never have listed. Any share-of-voice model built on a fixed competitor list will therefore understate how contested the available slots actually are.

The number you are not entitled to quote

Measured, Volume I. The 85 companies were selected as plausible challenger brands rather than drawn at random, and Volume I states in its own limitations that this skew lowers average named rates relative to a random sample of all vendors. **Reasoned.** So 35% is a property of this sample, not an industry base rate, and quoting it as "one in three B2B software vendors" without the sampling frame attached is a misuse of it. Inside the sample the figure is exact and reproducible. Outside it, the honest claim is that a large fraction of plausible challengers are absent entirely, with the size of the fraction unmeasured.

Why the skew was chosen, and what it buys

Measured, Volume I. Category leaders enter the dataset through mentions rather than through selection: they are named in answers, but they were not the companies being measured. **Reasoned.** That design targets the population where a visibility gap is actionable, since a company already at 8 of 10 has little to gain from a diagnosis. The cost is that the sample cannot describe the whole market, and the benefit is that both ends of the distribution are observable in the same sweep, the challengers as measured subjects and the leaders as named brands.

One engine, one run per question

Measured, Volume I. All answers came from a single engine with one run per question, and Volume I states that AI answers vary run to run and that repeated runs of the same question can return different leaders. **Reasoned.** Every figure in this chapter is therefore a point-in-time estimate with an unmeasured variance around it. Volume I says so plainly rather than presenting its numbers as fixed rankings. Chapter 6 at /measurement-noise/ measures that variance directly on a later sample and reports how wide it is, which is the correct context for reading any decimal place in this chapter.

What a single engine cannot tell you

Reasoned. Holding the engine constant is the right call for a study comparing categories with each other, because it removes engine as a confound. It is the wrong basis for a claim about AI search in general, because it measures one product. A company absent from this engine may be present on another, and the size of that effect is not something Volume I can speak to. Chapter 5 at /engine-divergence/ measures it on four engines and finds the difference is substantial.

The collection window is part of the result

Measured, Volume I. Collection ran between 18 and 23 July 2026. **Reasoned.** These are products under continuous change, so the sweep describes a state of the world during that window and not a durable property of the companies in it. A figure quoted without its window is a figure whose meaning has quietly drifted. This is why every headline number in this manual is written with its date attached, and why a visibility score offered to you without one should be treated as incomplete rather than merely informal.

Precision and rounding

Measured, Volume I. Volume I states that segment sizes vary and that percentages are rounded. **Reasoned.** So the difference between 20% and 21%, or between 35% and 36%, is not a difference this dataset can adjudicate. Read the figures at the resolution the source publishes them at and no finer. Where this manual needs a sharper figure it takes it from Volume II or Volume III, which report to more decimal places on their own samples, and it says which volume it came from rather than implying the sharper figure applies to Volume I's sweep.

What company-level anonymity costs and buys

Measured, Volume I. Company identities are withheld and replaced with stable identifiers, while aggregates and brand-level mentions of widely known market leaders are published. **Reasoned.** That means nobody can check whether the score assigned to any individual company is right, which is a real limitation on external verification. What can be checked is every aggregate, because the per-company rows are published with their scores attached to the identifiers. The trade buys the ability to publish per-company distributions at all, which a named dataset in this category would not have survived.

What this chapter does not claim

Reasoned. It does not claim that a zero-visibility company is a bad company, that a saturated leader deserves its position, or that any of these figures would replicate on a different question set. It does not claim causation in any direction. And it does not claim that being named is the same as being shortlisted by a human, because nothing in this dataset observes a buyer. The claim is narrow and it is about measurement: on these questions, on this engine, in this window, this is how often each company appeared.

What this means for your buying decision

Reasoned. Establish your own zero rate before you buy anything, because the remedy for a zero differs in kind from the remedy for a low score. Ask any supplier for the count of questions on which you were named zero times, not just an average, and ask what sampling frame their benchmark figures come from. If a supplier quotes a market-wide zero rate without naming the sample it was drawn from, they are quoting a number that has lost its denominator somewhere along the way. Chapter 12 at /how-to-buy-geo/ sets out the full list of questions.

Where to go next

Reasoned. Chapter 2 at /absence-ladder/ turns the zero rate into a diagnosis by classifying which questions the absences fall on. Chapter 3 at /cited-not-recommended/ takes the second scored outcome, domain citation, and shows why it moves independently of naming. Chapter 0 at /selection-not-ranking/ is the frame that makes this distribution expected rather than surprising.

Sources

Every figure in this chapter comes from The 2026 State of GEO, Volume I, published with its datasets under CC BY 4.0: the collection window, the 860 scored answers, the 5,160 traced citations, the 35% zero rate, the 20% median, the 80% leader median, the 2.05 vendors per answer, the top-ten and single-appearance domain shares, and the stated limitations on sampling, engine, run count and rounding. The per-company rows are in `company_visibility_anonymized.csv` at github.com/Broadcastwell/state-of-geo-2026.

Chapter 2. The Absence Ladder

The claim this chapter defends

Measured, Volume II. Across 616 recorded absences from 85 B2B software companies in 60 categories, the kind of question a company loses to an AI answer changes systematically as its visibility rises: companies named in zero of ten answers lose category-level questions 55.3% of the time and head-to-head comparisons 20.0% of the time, while companies named in seven or more of ten reverse that to 12.5% and 68.8%. The pattern is significant across all four tiers (chi-square 61.8, df 15, $p = 1.3 \times 10^{-7}$). **Reasoned.** Invisibility is not one condition. It has stages, and the work each stage implies is different.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

A single percentage is a symptom, not a diagnosis

Reasoned. A visibility score of 2 out of 10 tells you that eight answers went by without you. It does not tell you what those eight answers were about. Two companies can both score 2 and sit in entirely different strategic positions. One is losing every best-of shortlist in its category and never gets considered at all. The other appears on the shortlists and loses only head-to-head comparisons against one named rival. Those are different problems with different fixes and different costs, and a single percentage hides the difference completely. The number is real. It is just not a diagnosis.

So invert the unit of analysis

Measured, Volume II. The standard approach studies the answers where a company appeared and counts them. Volume II did the opposite. It studied the answers where the company did not appear, kept the full text of every one of those questions, and asked what they had in common. **Reasoned.** That inversion is the whole method. Presence data tells you how much visibility a company has. Absence data tells you what kind of visibility it is missing, and that is the thing you can act on. Everything in this chapter follows from measuring the losses rather than the wins.

What was measured

Measured, Volume II. 85 B2B software companies across 60 distinct product categories. For each company, ten buyer questions reflecting how a real purchaser researches that category. Each question was put to one AI engine with live web search enabled, one run per question. For every answer the study recorded whether the company's brand was named, whether its domain was cited as a source, and which competitor was named most often. The sample is challenger-skewed by construction: companies were selected as plausible non-leaders in categories with an identifiable incumbent, which is the population where a visibility gap is actionable at all.

Why the category count is 60 and not 61

Measured, Volume II. Volume I of this research programme reported 61 categories. That figure counts every category string in the collection sheet, including one category whose company was queued but never scored. Volume II analyses only companies with complete scores, which is 60 categories. The 85 companies are the same in both volumes. This is stated here rather than left as an apparent contradiction between two published papers, because a reader checking one number against the other will otherwise find a discrepancy that is not one.

How 616 absence records were built

Measured, Volume II. For every question where the company was not named, the study retained the full question text. That produced 616 absence records. Completeness was then validated directly rather than assumed: for all 85 companies, the absence count equals ten minus the visibility score, with zero mismatches. **Reasoned.** That check matters more than it looks. It means no absence was silently dropped and no answer was double counted, so the denominator of every percentage below is the complete set of losses rather than a convenience subset of them.

The six question shapes

Measured, Volume II. Each absence question was assigned exactly one shape by regular-expression rules applied in fixed precedence.

Shape	Rule	Example
Head-to-head comparison	contains "vs", "versus", or "comparison"	"Whitespace vs Artificial Labs: which is better?"
Alternatives-to-incumbent	contains "alternative"	"What are the best Sequel alternatives?"
Best-of shortlist	contains "best" or "top N"	"What is the best resource management software?"
Evaluation criteria	evaluate, choose, select, criteria, pricing model	"How should I compare pricing models when switching?"
Use case or problem	opens "how can/do/does" or contains "use case"	"How can this software reduce manual data entry?"
Definitional	opens "what is/are" without "best"	"What is reinsurance placement software?"

Measured, Volume II. 2.8% of records fell to *other*. Precedence matters and is published so the classification is reproducible: a question containing both "best" and "vs" is counted as a comparison, not a shortlist.

The result

Measured, Volume II. The distribution of absence shapes shifts systematically with visibility.

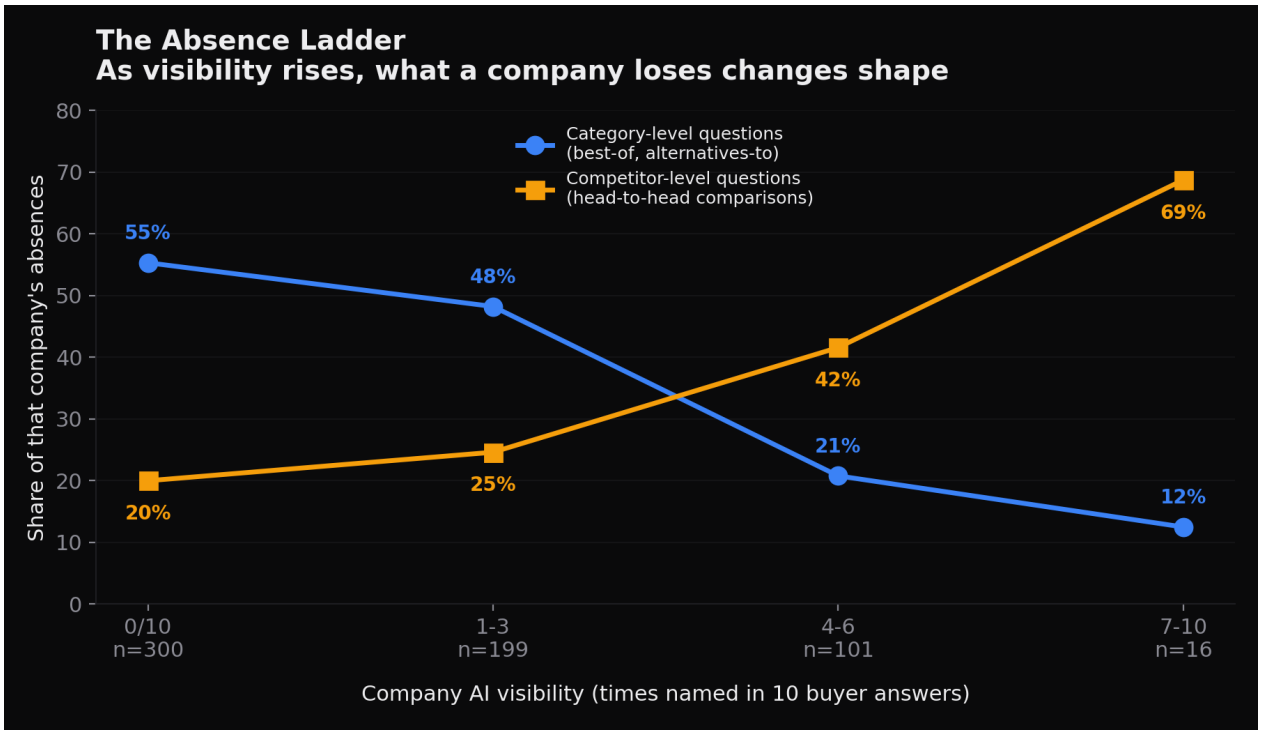


Figure 1. The absence ladder. Volume II, 616 absence records across 85 companies.

Visibility	Companies	Absences	Category-level	Head-to-head
Named 0 of 10	30	300	**55.3%**	20.0%
Named 1 to 3	25	199	48.2%	24.6%
Named 4 to 6	21	101	20.8%	41.6%
Named 7 to 10	8	16	12.5%	**68.8%**

Measured, Volume II. Category-level means best-of shortlists plus alternatives-to-incumbent queries.

What the top row rests on, stated before anyone asks

Measured, Volume II. Nine companies scored 7 or above. One of them scored 10 of 10 and therefore produced no absence records at all, so the top tier's absence analysis rests on the remaining 8 companies and their 16 records. Company counts across the four tiers sum to 84 in the table above and to 85 in the sample, for that reason. **Reasoned.** The 68.8% comparison figure is the most quotable number in this chapter and it rests on the smallest base in it. Anyone repeating it should repeat the base with it.

The statistics, in full

Measured, Volume II. All four results below are computed in Volume II from the same 616 records.

- Chi-square across tier by shape: $\chi^2 = 61.8$, $df = 15$, $p = 1.3 \times 10^{-7}$, Cramér's V = 0.18.
- Visibility against absence-is-category-level: $r = -0.284$, $p = 6.8 \times 10^{-13}$, n = 616.
- Visibility against absence-is-comparison: $r = 0.234$, $p = 4.1 \times 10^{-9}$.

- Excluding the small top tier entirely, the comparison effect survives: **r = 0.184, p = 5.9 × 10⁻⁶**, n = 600.

Reasoned. The last line is the one that decides whether the finding is an artefact of 16 records. It is not. Drop the top tier completely and the relationship between visibility and comparison-shaped absence still holds at n = 600.

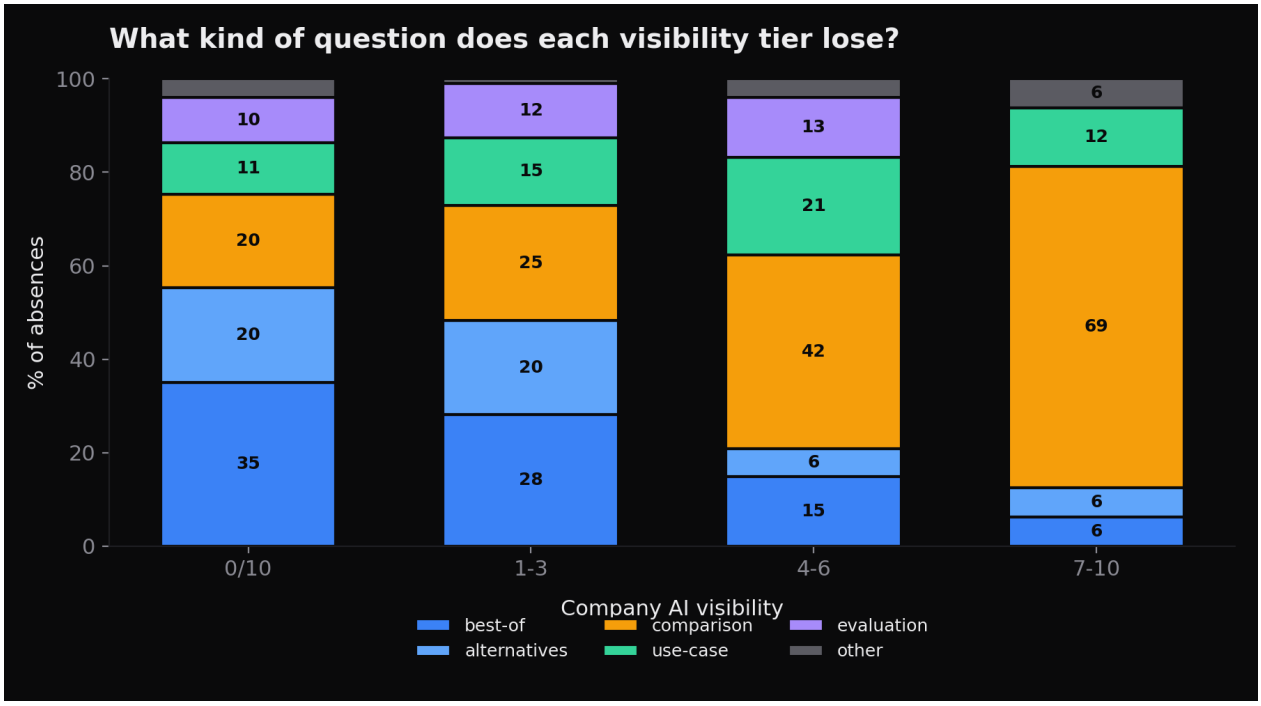


Figure 2. Shape mix by visibility tier. Volume II, 616 absence records.

Two gates, not one wall

Measured, Volume II, and reasoned from it. A company named zero times is failing at the category door. The engine does not consider it a member of the set when a buyer asks who the players are. More than half of its losses are questions that never mention a competitor by name. They simply ask who exists, and the answer does not include it. A company named seven or more times has cleared that door and is a recognised member of the category. What it loses now is the second gate: direct comparison against a specific named rival. Nearly seven in ten of its remaining absences are comparison questions.

Why the middle tiers matter to the argument

Measured, Volume II, and reasoned from it. If the two gates were simultaneous, the middle tiers would look like noisy blends with no order to them. They do not. The 1 to 3 tier sits close to the zero tier, at 48.2% category-level against 24.6% comparison. The 4 to 6 tier has already flipped, at 20.8% against 41.6%. The mix moves monotonically in both columns across all four tiers. That is what a sequential process looks like when you cut it into slices, and it is the reason the finding is described as a ladder rather than as a correlation.

The practical consequence, which is the point of the chapter

Reasoned. The work implied by each stage is different. A company at the category door needs entity presence and inclusion in the third-party sources engines consult when assembling a set. Its problem is membership. A company at the comparison gate needs comparison content that a model can quote against a specific named competitor. Its problem is preference. Prescribing the second to a company stuck at the first is a common and expensive mistake, and a single visibility percentage cannot tell you which one you are. The absence mix can.

What this does not license you to conclude

Reasoned. The ladder describes where a company sits, not why it got there and not what happens if it acts. Nothing in this dataset is an intervention study. No company was measured, changed and measured again. The finding is that absence shape and visibility level move together across a cross-section of 85 companies at one moment in time. Reading it as a causal claim that fixing category presence produces movement into the next tier goes beyond what was measured, however plausible that reading is.

Limitation: one engine, one run per question

Measured, Volume II. All answers behind this analysis came from a single AI engine with live web search, one run per question. AI answers vary between runs. Volume II did not measure that variance, and it cannot claim these proportions would replicate exactly. **Reasoned.** Treat the direction of each finding as the result, not the decimal. Chapter 6 of this manual measures that run-to-run variance directly on a different sample and reports how large it is, which is the correct place to look before quoting 55.3% as though it were a constant.

Limitation: the sample is challenger-skewed by construction

Measured, Volume II. Companies were selected as plausible non-leaders in categories with an identifiable incumbent. **Reasoned.** That is the right frame for the question being asked and it makes the sample unrepresentative of B2B software as a whole. The zero-visibility rate in this dataset is a property of this sample and not a property of the industry. Do not quote it as an industry base rate. The absence ladder itself is a within-sample relationship and is less exposed to that skew than any headline rate is.

Limitation: classification is rule-based, not human-annotated

Measured, Volume II. Question shapes were assigned by regular expression rather than by human annotation. That trades some accuracy for complete reproducibility. **Reasoned.** A human annotator would classify some questions differently, particularly the boundary between evaluation-criteria and use-case shapes. The rules and their precedence are published in full, and the 616 classified questions are published with their verbatim text, so anyone who disagrees with the rules can reclassify the whole set and recompute the table rather than argue about it.

Limitation: the question set is a design choice

Measured, Volume II. The ten questions per category were generated to reflect realistic buyer research rather than sampled from real search logs. **Reasoned.** A different question set would produce a different absence mix. The shapes are drawn from that set, which is why the tier comparison is internally consistent even where the absolute proportions are set-dependent. If your own buyers ask a materially different mix of questions than this bank does, the tier boundaries will still order correctly but the percentages will not transfer.

Limitation: 60 categories is enough to see a pattern

Reasoned. 60 categories is enough to see a pattern and not enough to call any individual category. Nothing here supports a statement of the form "in category X, challengers lose comparisons." The unit that carries the finding is the tier, pooled across categories, and the per-category cell counts are far too small to stand alone. Treat any category-level reading of this dataset as a hypothesis to be tested on a larger sample rather than as a result.

Do this yourself: build your own absence list

Reasoned. You do not need a tool. Write down the ten questions a real buyer in your category would type. Put each one to an AI engine with web search on. For each answer, record one thing only: were you named, yes or no. The questions where you were not named are your absence list. Keep the full text of each. If you were named in all ten you have no absence list and this chapter has nothing to say to you, which is itself useful information.

Do this yourself: classify by shape

Reasoned. Take your absence list and sort each question into one of the six shapes above, applying the precedence rules in order: comparison first, then alternatives, then best-of, then evaluation, then use case, then definitional. Anything that fits none of them is *other*. Then collapse best-of and alternatives-to-incumbent into a single category-level bucket, and keep head-to-head comparison separate. Those two buckets are the diagnostic. Count them.

Do this yourself: read which gate you are at

Reasoned. If category-level questions dominate your absence list, you are at the category door. The engine does not yet treat you as a member of the set, and comparison content against a named rival will be read by very few buyers because you are not appearing in the answers where rivals are shortlisted. If head-to-head comparisons dominate, you are at the comparison gate. You are in the set and losing preference inside it, and more category-presence work will move a number that is already moving.

Do this yourself: read the ambiguous case

Measured, Volume II, and reasoned from it. If the two buckets are close to even, you are in the transition the middle tiers describe, and the honest reading is that you are doing both jobs at once. The 4 to 6 tier in this data sits at 20.8% category-level against 41.6%

comparison, so an even split is if anything below where this sample's mid-tier companies sat. Sequence the work rather than splitting it: membership problems bound the return on preference work, so the category door is the constraint to clear first.

The maturity model, stated plainly

Reasoned. Stage one, the category door: the engine does not name you when a buyer asks who exists. Stage two, the comparison gate: the engine names you in the set and picks someone else when a buyer asks you against a specific rival. Stage three, which this dataset can only point at because one company reached it and produced no absences at all: named in every answer, no absence list to classify. The single visibility percentage tells you none of this. The absence mix tells you all of it.

What this changes

Reasoned. If the absence ladder holds, AI visibility is not one metric and should not be managed as one. The useful question is not what percentage of answers name us. It is what kind of question are we currently losing. That answer tells you whether you are fighting for admission to a category or for preference within one, and those require different work, different content and different evidence. The single number is a symptom. The shape of the absence is the diagnosis.

Where to go next

Reasoned. If you have not yet accepted the model this chapter is built on, read Chapter 0 at /selection-not-ranking/, which sets out why an answer is a selection rather than a ranking. If you want the size of the problem before the shape of it, read Chapter 1 at /named-zero-times/. If you are at the first gate, Chapter 9 at /category-door/ sets out what evidence admission appears to require; if you are at the second, Chapter 10 at /comparison-gate/ takes up preference inside the candidate set.

Sources and reproduction

Every number in this chapter comes from The 2026 State of GEO, Volume II, published with its data and analysis code. Three files carry the whole result: `challenger_visibility_v2.csv`, 85 rows of per-company results; `absence_questions_classified.csv`, all 616 absence questions with verbatim text, assigned shape and visibility tier; and `absence_shape_by_tier.csv`, the aggregate table behind Figure 1. They are at github.com/Broadcastwell/state-of-geo-2026. Anyone can reproduce every figure above from those three files.

Chapter 3. Cited is not recommended

By the end of this chapter you will know why an engine quoting your pages and an engine recommending your product are separate measurable events, how loosely the two move together, and why a supplier who reports them as one number is hiding the more useful of the two signals.

The claim this chapter defends

Measured, Volume II. Brand naming and domain citation correlate, but loosely: Pearson r = 0.512 and Spearman ρ = 0.577, at $p < 10^{-6}$. **Measured, Volume II.** 29.4% of companies were cited more often than they were named, and nine companies were named zero times while their own domain was cited as a source, one of them five times out of ten. **Reasoned.** Being quotable and being recommendable are different properties of the same company, and a content strategy aimed only at the first will not deliver the second.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

Two events, scored separately, on purpose

Measured, Volume II. For every answer the study recorded two independent binary outcomes: whether the company's brand was named in the answer, and whether its own domain appeared among the sources the engine cited. **Reasoned.** Those are different events in the machine. Citation is a statement about where the engine went for evidence. Naming is a statement about which vendors the engine put in front of the buyer. Scoring them together into one visibility percentage destroys the distinction, and the distinction turns out to carry most of the diagnostic value in this chapter.

The correlation is real and it is weak

Measured, Volume II. Pearson r = 0.512 between times named and times cited, with Spearman ρ = 0.577, both significant at $p < 10^{-6}$. **Reasoned.** A correlation of that size means the two signals share roughly a quarter of their variance and differ in the rest. It is strong enough that nobody should claim they are unrelated, and weak enough that nobody should treat one as a proxy for the other. If citation predicted naming reliably, the sensible strategy would be to maximise quotable content and wait. The measured relationship does not support that.

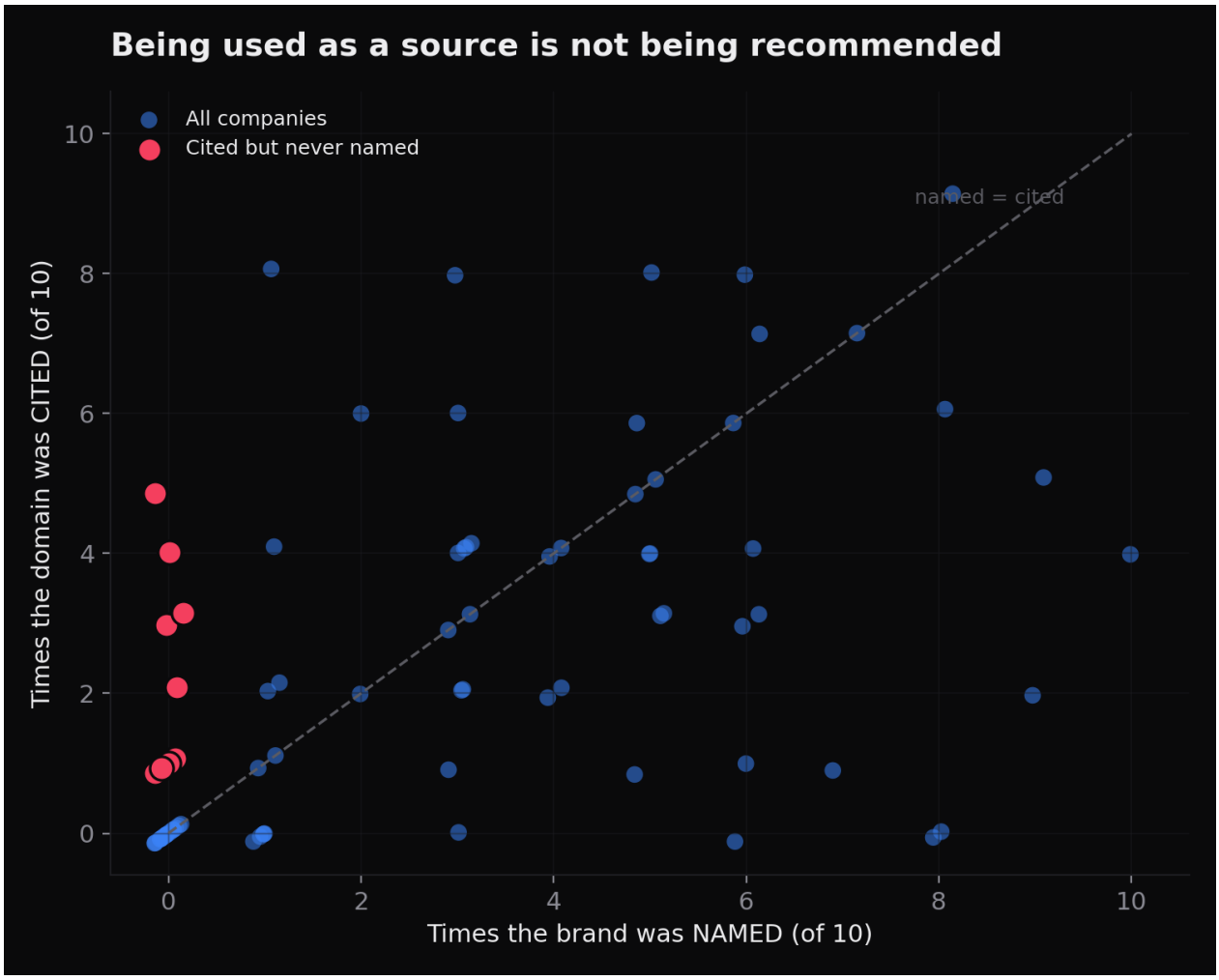


Figure 1. Naming against citation, per company. Volume II, 85 companies.

Nearly a third are cited more than they are named

Measured, Volume II. 29.4% of companies were cited more often than they were named. Among that group, the average company was named 2.2 times and cited 4.4 times. **Reasoned.** Twice as many citations as namings is not a rounding artefact. For those companies the engine is reading their material, treating it as usable evidence, and then writing an answer that recommends somebody else. Whatever is failing for them, it is not the production or the discoverability of content, because both are demonstrably working.

The nine that are cited and never named

Measured, Volume II. Nine companies were named zero times while their own domain was cited as a source, and one of those was cited in five of its ten answers. **Reasoned.** This is the cleanest available demonstration that the two signals are separate. A company cited half the time and named none of the time is supplying the evidence layer for answers that recommend its competitors. The pages are working. They are working for somebody else. No single visibility percentage can represent that state, because the percentage has to choose which of the two events it counts.

Why this happens, stated as inference

Reasoned. An engine assembling an answer has two distinct needs: candidates to recommend and evidence to support the recommendation. Material that explains a category well, defines terms, or lays out how to evaluate options serves the second need without positioning its author as a candidate for the first. That is a plausible mechanism for the measured pattern and it is not tested here. Volume II observes the outcome and does not observe the retrieval, so treat the mechanism as a hypothesis that fits rather than a finding.

The commercial consequence for the cited-not-named group

Measured, Volume II. For companies in this group the binding constraint is not content volume. **Reasoned.** Adding more of the material that is already being cited will raise citation further and has no measured tendency to raise naming. The work that plausibly moves naming is different in kind: being present in the third-party places where an engine assembles a candidate set, and being described as a vendor of a thing rather than as a commentator on it. Chapter 9 at /category-door/ sets out the standards for that.

What the reverse case means

Reasoned. The mirror position, named more often than cited, is the more comfortable one and it carries its own risk. A company recommended without its own domain being used as evidence is being selected on the strength of what other sources say about it. That is durable while those sources persist and fragile if they change, and it is invisible to any measurement that reports naming alone. Volume II publishes both columns per company precisely so that either asymmetry can be seen.

This manual's own self-audit is a case of it

Measured, self-audit of 18 August 2026. On a re-measure of the same ten published questions across four engines, at one run per question, Broadcastwell was named in 1 of 40 answers and cited once among 674 citations. The naming and the citation are the same answer: one engine quoted Broadcastwell's own published visibility page as a source. **Reasoned.** That is a citation of the research and not a recommendation of the firm. The engine used the page as evidence about somebody else's comparison. Reported as a visibility figure, it would look like a firm appearing in an answer. Read correctly, it is exactly the pattern this chapter describes, and the operator of the measurement is in it.

Why saying that plainly matters

Reasoned. A supplier who measures itself and reports the naming without reading what the naming consisted of has made the error this chapter is about, in public, on its own data. The honest report is that one engine cited a research page and recommended other agencies in the same answer. Anybody quoting the 1 of 40 as evidence of visibility, including this firm, would be quoting a number whose content contradicts its headline. Both self-audit figures are published on every page of this manual with their run structures attached for that reason.

What a measurement has to publish to be readable here

Reasoned. Three things. The naming count and the citation count as separate columns rather than a blended score. The identity of the source that was cited, so a citation of your own domain can be told apart from a citation of a page about you. And the answer text or enough of it to see whether a naming was a recommendation, a passing mention, or an attribution of evidence. Volume II publishes the first, Volume III publishes cited URLs per answer, and the third is the one most commercial tools do not offer.

Reading the four positions

Reasoned. Plotting naming against citation gives four positions and each implies different work. Named and cited: the engine both selects you and uses your material, which is the position to defend rather than to improve. Named but rarely cited: your selection rests on sources you do not own, which is durable while they last and outside your control. Cited but rarely named: the position this chapter is about, where supply of material is not the constraint. Neither: the category door, which is Chapter 1 at /named-zero-times/ and Chapter 9 at /category-door/.

Why the diagonal is the wrong expectation

Measured, Volume II. The published scatter of naming against citation places a line where the two counts are equal, and the companies scatter on both sides of it rather than along it. **Reasoned.** A tool that reports one blended visibility number is implicitly asserting that companies sit near that diagonal, because only then does one number stand in for both. The measured spread says otherwise. The further a company sits from the diagonal, the more information a blended score is discarding about it, and the companies furthest from the diagonal are exactly the ones with the most actionable diagnosis available.

The design decision that made this visible

Measured, Volume II. The two outcomes were scored separately and both columns are published per company. **Reasoned.** Nothing about the finding required a more sophisticated instrument than two checkboxes per answer. It required only that the two checkboxes were never added together. That is worth stating because it sets a low bar that most commercial reporting still does not clear: the separation costs nothing to collect and is lost only at the point where somebody decides a single headline number is easier to sell.

The limitation that bounds all of this

Measured, Volume II. All answers came from a single engine with one run per question, on a challenger-skewed sample of 85 companies in 60 categories. **Reasoned.** The 29.4% share and the count of nine are properties of that sample and that engine. The structural claim, that naming and citation are separable and can move in opposite directions, is the part that generalises, because it follows from the two events being different events. The proportions do not travel and should not be quoted as though they do.

What run-to-run variance does to the nine

Measured, Volume II. Volume II states that AI answers vary between runs and that it did not measure that variance. **Reasoned.** So the nine companies named zero times are nine companies that were named zero times on one run each of ten questions. A second run might have named one of them once. The finding that survives that uncertainty is the group-level one: a substantial minority of companies sit above the diagonal, cited more than named. The exact membership of the extreme group is less stable than the existence of the group. Chapter 6 at /measurement-noise/ puts numbers on how unstable.

The one number that would help instead

Reasoned. If a single figure has to be reported, the useful one is not naming or citation but the difference between them, signed. A positive gap says the engine selects you more than it quotes you, and your position depends on third parties. A negative gap says the engine quotes you more than it selects you, and your material is working as evidence rather than as candidacy. A gap near zero says the two signals are moving together and neither is the obvious constraint. Volume II publishes both columns per company, so the gap is available to anyone who wants it.

What this predicts about content strategy, and how to falsify it

Reasoned. The prediction is specific and testable: for a company already cited more often than it is named, adding more of the same kind of material should raise citation and leave naming roughly where it was. Anyone can falsify that by publishing a before-and-after on a named question set with both columns reported. No such test exists in these volumes, which is why this is offered as an inference rather than a result. If it turns out to be wrong, the honest correction is a published counterexample rather than a quiet edit.

What this chapter does not claim

Reasoned. It does not claim that citation is worthless, that being quoted has no commercial value, or that the two signals are independent, which the measured correlation rules out. It does not claim that any specific content change would convert citation into naming, because no intervention was tested. And it does not claim that engines deliberately separate sources from recommendations. The claim is that the two outcomes are separately measurable, measurably different, and routinely reported as one thing.

How to check your own position in ten minutes

Reasoned. Take the ten questions a buyer in your category would ask. For each answer record two things rather than one: were you named, and does your own domain appear among the cited sources. Then plot the pair. If citation exceeds naming you are in the group this chapter describes and your constraint is not content supply. If naming exceeds citation your position rests on third-party sources you do not control. If both are zero, Chapter 1 at /named-zero-times/ is the relevant starting point.

What this means for your buying decision

Reasoned. Ask any supplier to show naming and citation as two separate columns before you accept a single visibility number, and ask which of the two their proposed work is meant to move. A supplier who cannot separate them cannot tell you whether your content is failing to be found or failing to position you as a vendor, and those need different work. If the answer to "what will improve" is more content, ask what the citation column already shows. Chapter 12 at /how-to-buy-geo/ has the full question list.

Where to go next

Reasoned. Chapter 4 at /who-gets-cited/ looks at the citation side across the whole corpus and asks which kinds of domain the engine reaches for. Chapter 2 at /absence-ladder/ classifies the questions you are absent from, which is the other half of a diagnosis. Chapter 10 at /comparison-gate/ covers what a model needs in order to lift a claim about you into a comparison answer.

Sources

Every figure in this chapter comes from The 2026 State of GEO, Volume II: the Pearson and Spearman correlations, the 29.4% share, the 2.2 against 4.4 averages within that group, and the nine companies cited but never named. The per-company naming and citation counts are published in `challenger_visibility_v2.csv` at github.com/Broadcastwell/state-of-geo-2026. The self-audit figures are the operator's own, published with their dates and run structures in the block at the foot of this page.

Chapter 4. Who the engines actually cite

By the end of this chapter you will know what kinds of source an AI engine reached for when it answered B2B software buying questions, how concentrated or fragmented that evidence layer is, and which of those figures carry which denominator, because two of the headline numbers in this chapter are computed on different bases and are routinely quoted as though they were not.

The claim this chapter defends

Measured, Volume I. Among the 100 most-cited domains, 77.4% of citations resolved to vendor-authored content, 10.2% to review platforms, 6.5% to independent media and 5.9% to analyst firms. **Measured, Volume I.** Community sources, meaning Reddit, Quora and Stack Overflow, accounted for 0.0% of all 5,160 citations. **Reasoned.** The evidence an engine reached for was overwhelmingly published by vendors, and the places practitioners assume carry weight in software buying did not appear at all.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

Two denominators, and why the difference matters

Measured, Volume I. The composition figures above are computed across the 100 most-cited domains, and those 100 domains account for 36% of all citations. The community figure is stated against the full corpus of 5,160 citations. **Reasoned.** So 77.4% vendor-authored is a statement about the head of the distribution, not about every citation the engine made. Quoting it as "77% of all AI citations are vendor content" overstates what was measured by attaching a corpus-wide denominator to a head-of-distribution figure. This manual states the base every time, and you should demand the same of anyone quoting it at you.

What was counted

Measured, Volume I. Every cited URL across the sweep was logged, producing 5,160 traced source citations. The 100 most-cited domains were then classified by hand into vendor-authored, review platform, analyst, independent media, community, encyclopaedia and social categories. **Reasoned.** Hand classification of a hundred domains is tractable and auditable, and it is also the limiting factor on the resolution of this chapter. Nothing is known about the composition of the citations outside those 100 domains beyond the domains themselves, which are published.

The head of the distribution, concretely

Measured, Volume I. The single most-cited domain was pipeline.zoominfo.com with 132 citations, classified vendor-authored. Second was gartner.com with 110, classified analyst. Fourth was g2.com with 57, classified review platform. **Reasoned.** The top of the list is not

a list of publishers. It is mostly a list of software companies publishing about their own category, interleaved with the two or three third-party institutions that survive at that altitude. A vendor reading this list is looking at the pages that were doing the work of explaining its market to an engine.

Below the top few, it is vendors all the way down

Measured, Volume I. Ranks three, five, six and seven of the most-cited domains were algolia.com with 69 citations, salesmotion.io with 50, guideflow.com with 47 and improvado.io with 46, all four classified vendor-authored. **Reasoned.** The pattern at the top is not a handful of vendor domains surrounded by publishers. It is a run of individual software companies, punctuated by one analyst domain and one review platform. If you are looking for the publications that explain your market to an engine, in this dataset they are largely your competitors' own pages.

The most-named brand and the most-cited domain are different companies

Measured, Volume I. The most-named brand across the sweep was 6sense, with 123 mentions. The most-cited domain was pipeline.zoominfo.com, with 132 citations. **Reasoned.** The company whose material the engine leaned on hardest and the company the engine recommended most often were not the same company. That is the finding of Chapter 3 at /cited-not-recommended/ visible at the very top of both distributions, and it is a useful corrective to the assumption that the way to be recommended is to become the most-quoted source in your category.

What the taxonomy cannot separate

Measured, Volume I. Classification was assigned per domain, into vendor-authored, review platform, analyst, independent media, community, encyclopaedia and social. **Reasoned.** A per-domain label cannot distinguish a comparison page from a pricing page on the same vendor domain, nor an editorially independent article from sponsored placement on a media domain. It also forces single labels onto domains that do more than one thing. The consequence is that the composition figures are reliable about which organisation published a source and unreliable about what kind of content the source was, which is a real limit on how far they can be pushed.

The evidence layer is fragmented, not concentrated

Measured, Volume I. The top 10 domains held only 12% of all citations, and 56% of cited domains appeared exactly once. **Reasoned.** That is a long tail, and it sits in tension with the concentration seen on the vendor-naming side, where a small set of names took a large share of the slots. Concentrated outputs assembled from a fragmented input layer is a specific structure: there is no small set of publications to be present in, because there is no small set of publications doing the work.

How many distinct places the engine went

Measured, Volume I dataset documentation. The published dataset README records 1,753 unique domains cited across the sweep. This figure appears in the dataset documentation rather than in a paper body, and is labelled accordingly here. **Reasoned.** 1,753 distinct domains behind 5,160 citations is roughly three citations per domain on average, which is another way of saying the same thing the single-appearance share says: the engine went to a very large number of places, most of them once.

What fragmentation rules out

Reasoned. It rules out a media-placement strategy as the primary route. If twelve percent of citations sit in ten domains, then winning a place in all ten would still leave the great majority of the evidence layer untouched. It also rules out the reverse assumption, that the tail is where the work is, since a single appearance in one of the many domains cited exactly once is worth very little on its own. What it points to instead is breadth of corroboration rather than placement in any particular outlet.

The community result is the surprising one

Measured, Volume I. Community sources accounted for 0.0% of all 5,160 citations. **Reasoned.** Practitioner folklore holds that software buyers read Reddit threads and that engines therefore surface them. On these questions, on this engine, in this window, that did not happen once in five thousand one hundred and sixty citations. The honest reading is narrow: it is a fact about one engine's retrieval on B2B software buying questions in July 2026, and it says nothing about consumer categories or about other engines, one of which is measured differently in Chapter 5 at /engine-divergence/.

Why a zero is worth more attention than a small number

Reasoned. A category at two or three percent would be a minor contributor. A category at zero across the whole corpus is a structural absence, and structural absences are more informative than small positives because they cannot be explained by sampling noise in the same way. It also means any strategy premised on community presence feeding AI answers has no support in this dataset and needs its own evidence before anyone spends against it.

Analyst and review presence, in proportion

Measured, Volume I. Within the top 100 domains, review platforms took 10.2% of citations and analyst firms 5.9%. **Reasoned.** Both are present and neither is dominant. That is a more useful reading than either the claim that review sites are the gatekeepers of AI answers or the claim that they no longer matter. They are one input among several, they punch above their number of domains because a single review platform can carry many citations, and a vendor absent from them is absent from roughly a tenth of the head of the evidence layer.

What this implies about where evidence has to exist

Reasoned. If the head of the layer is mostly vendor-authored and the tail is enormous and thin, then the practical requirement is that material describing your category, and describing you as a member of it, exists in enough independent places that an engine assembling a candidate set encounters it more than once. That is a statement about corroboration rather than about volume, and the standards for it are the subject of Chapter 9 at /category-door/.

Your own domain is a separate question

Measured, Volume I. Each answer was scored for whether the company's own domain appeared among cited sources, independently of whether the brand was named. **Reasoned.** So a vendor-authored citation share of 77.4% at the head of the distribution does not imply that your vendor-authored pages will be cited, still less that being cited will get you named. Chapter 3 at /cited-not-recommended/ shows those two outcomes moving apart, including nine companies cited as a source while never being named at all.

Limitation: the classification is by hand and by domain

Measured, Volume I. The type assignment was made by hand, at domain level, for the 100 most-cited domains only. **Reasoned.** Domain-level classification cannot distinguish an independent review published on a vendor's domain from a product page on the same domain, and hand assignment is a judgement that another analyst would make differently at the margins. The classification and the underlying counts are published, so anyone who disagrees with a call can reclassify and recompute rather than argue about it.

Limitation: one engine, one run, one window

Measured, Volume I. All citations came from a single engine with live web search, one run per question, collected between 18 and 23 July 2026. **Reasoned.** Citation behaviour is a property of a retrieval stack, and retrieval stacks differ between products and change over time. Chapter 5 at /engine-divergence/ reports the same composition question measured across four engines in August 2026 and finds the vendor-owned share varies widely between them, which is the correct caution to carry into any use of the figures in this chapter.

Limitation: citation is not endorsement and not traffic

Reasoned. A cited URL is a URL the engine used while composing an answer. It is not a claim that the source was read favourably, that the buyer clicked it, or that it influenced the recommendation. Nothing in this dataset observes a buyer, and nothing in it observes the weight the engine placed on any individual source. Read the composition as a description of what the engine reached for, and not as a ranking of influence.

The measurement standard this chapter implies

Reasoned. Any citation statistic offered to you should arrive with four things attached: the corpus size it was computed over, the subset it was computed on if that subset is not the whole corpus, the classification scheme and who applied it, and the collection window.

Volume I publishes all four, which is why its figures can be quoted precisely and why the precision is worth insisting on. A citation composition chart with none of the four is a chart whose numbers cannot be checked or compared with anybody else's.

Why comparing citation studies is harder than it looks

Reasoned. Two studies can report very different vendor-owned shares without either being wrong, because the share depends on which subset was classified, how the taxonomy draws its lines, which engine was measured and when. Chapter 5 at /engine-divergence/ shows the same quantity varying substantially between four engines measured in a single window, which puts a floor under how much of the disagreement between published studies is real difference rather than methodological difference.

What this chapter does not claim

Reasoned. It does not claim that vendor-authored content is the best evidence, that engines prefer it on quality grounds, or that publishing more of it will get you cited. It does not claim community sources are worthless, only that they did not appear here. And it does not claim these proportions hold outside US English B2B software buying questions on one engine in one week, because the sample cannot support that and Volume I says so.

What this means for your buying decision

Measured, Volume I, and reasoned from it. Ask any supplier which denominator their citation statistics use, because head-of-distribution and whole-corpus figures differ substantially and are quoted interchangeably in this category. Ask whether their plan depends on placement in a small number of outlets, and if so, how that squares with the top ten domains holding 12% of citations. If a proposal leans on community presence as the route into AI answers, ask for the measurement that supports it. Chapter 12 at /how-to-buy-geo/ has the full list.

Where to go next

Reasoned. Chapter 5 at /engine-divergence/ splits citation composition by engine and shows how much it moves. Chapter 3 at /cited-not-recommended/ explains why being in the evidence layer is not the same as being in the answer. Chapter 9 at /category-door/ turns the fragmentation finding into standards for what evidence about you has to look like.

Sources

Every figure in this chapter comes from The 2026 State of GEO, Volume I: the 5,160 traced citations, the composition of the 100 most-cited domains and their 36% share of all citations, the 0.0% community share of the full corpus, the 12% held by the top ten domains, the 56% single-appearance share, and the individual domain rows quoted above. The 1,753 unique-domain count is taken from the published dataset README and labelled as such. The per-domain file is `cited_source_domains_top100.csv` at github.com/Broadcastwell/state-of-geo-2026.

Chapter 5. One engine is not four

By the end of this chapter you will know how much two AI search engines agree about which vendors belong in an answer, how differently they behave in list length and citation composition, and how often a company's verdict flips depending on which single engine somebody happened to measure.

The claim this chapter defends

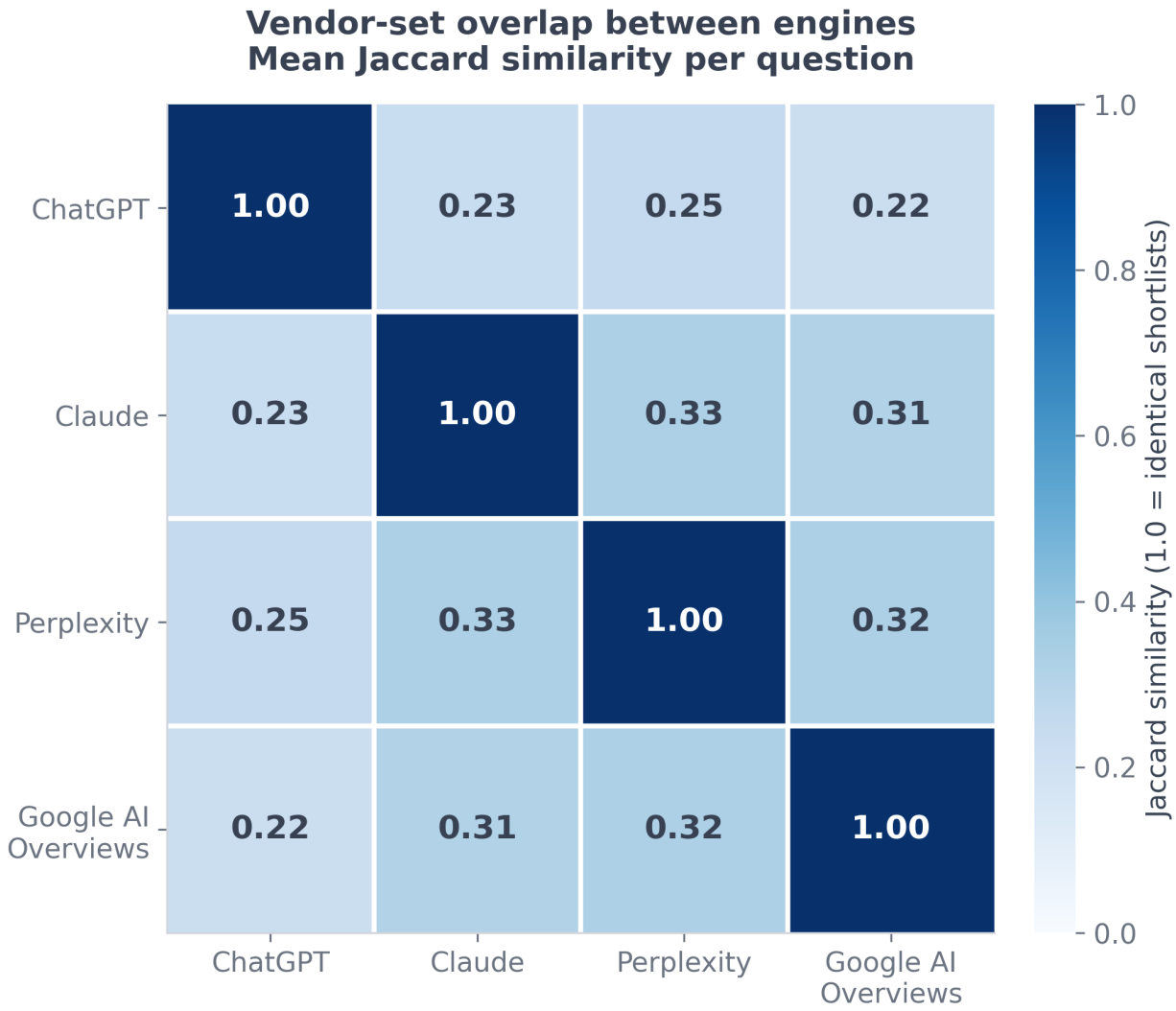
Measured, Volume III. Overall mean pairwise vendor-set overlap between engines was 0.313, with a 95% confidence interval of 0.295 to 0.330, over 590 engine-pair observations, and a median of 0.294. **Measured, Volume III.** Of 32 measured companies tested on at least two engines, 14 were visible on some engines and invisible on others. **Reasoned.** A visibility score produced on one engine is a measurement of that engine. It is not a measurement of AI search, and for nearly half the companies here it would have produced a materially different verdict if a different engine had been chosen.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

What was measured

Measured, Volume III. 280 B2B software buyer questions across 40 categories were put to four production AI search engines, ChatGPT, Claude, Perplexity and Google AI Overviews, in a single collection window running 05:28 to 16:29 UTC on 2026-08-04. The set of vendors named in each answer was extracted, and agreement was measured per question. The categories and companies are inherited from Volume I, so results tie back to that single-engine baseline rather than starting from an incomparable frame.

How much two engines share



Diagonal fixed at 1.0. Off-diagonal cells are means over questions both engines answered.

Figure 1. Mean vendor-set overlap per question, all engine pairs. Volume III.

Measured, Volume III. The most alike pair was Claude and Perplexity at 0.327, over 153 questions. The least alike was ChatGPT and Google AI Overviews at 0.223, over 18 questions. **Reasoned.** Every pair sits closer to a quarter than to a half. Two engines answering the same buyer question mostly name different vendors.

The pair sample sizes are not equal, and that is reported

Measured, Volume III. The Claude and Google AI Overviews pair rests on 247 questions, the Perplexity and Google AI Overviews pair on 136, Claude and Perplexity on 153, and all three pairs involving ChatGPT on 18. **Reasoned.** Pairs involving the two engines whose collection was cut short carry wide uncertainty and should not be read as precisely as the pairs with two hundred questions behind them. Volume III publishes the sample size next to every pair figure rather than presenting one pooled number, which is the practice this manual asks buyers to demand.

The spread between pairs is itself a finding

Measured, Volume III. The spread between the most and least similar pair was 0.104. **Reasoned.** So the question "how much do AI engines agree" does not have one answer even inside a single study on a single question set in a single window. Any published figure for cross-engine agreement is a figure for a specific pair, and pairs differ by roughly a third of their own magnitude. A number quoted without its pair is a number that has lost most of its meaning.

Most vendor mentions come from a single engine

Measured, Volume III. Across 142 questions answered by all of Claude, Perplexity and Google AI Overviews, 2,238 distinct vendor mentions were recorded, of which 63.3% came from exactly one engine and 17.5% from all three. **Measured, Volume III.** The same pattern holds at other set sizes: the single-engine share stayed between 63.3% and 69.8% whichever engines were in the room. **Reasoned.** A near two-thirds single-engine share is the clearest statement of the chapter's thesis. The candidate sets are substantially engine-specific rather than four views of one underlying ranking.

Engines return very different amounts

Measured, Volume III. Mean vendors named per answer was 4.77 for Google AI Overviews, 8.83 for Claude, 10.69 for Perplexity and 14.72 for ChatGPT. Mean answer length ran from 1,344 characters for Google AI Overviews to 7,627 for ChatGPT. **Reasoned.** Competing for a slot in a Google AI Overview is competing for a materially scarcer slot than competing inside a conversational answer, before any question of relative merit arises. That alone will make single-engine visibility scores disagree, and it is the reason the divergence analysis in Chapter 6 at /measurement-noise/ has to control for list length.

Agreement depends on which vendors you count

Measured, Volume III. Treating the engines as raters making a binary mention decision, Fleiss kappa across three engines was 0.015 over all vendors and 0.609 when restricted to the study universe. Across two engines it was −0.129 over all vendors and 0.548 restricted. **Reasoned.** Engines agree reasonably about the well-known vendors in a category and almost not at all about the long tail of names one engine mentions and the others never do. That is precisely the region a challenger brand occupies, which makes the low all-vendor figure the more relevant one for anybody reading this manual.

The sensitivity analysis says the same thing

Measured, Volume III. Recomputed on universe-only vendor sets, mean pairwise overlap rose to 0.506 with a 95% interval of 0.456 to 0.551, against 0.313 on the full sets, and the single-engine share of vendor mentions fell to 41.7% from 67.3%. **Reasoned.** Both readings are correct and they answer different questions. If you want to know whether engines agree about the established players, the restricted figure is the one. If you want to know whether they agree about who exists in a category at all, the full-set figure is the one, and it is much lower.

How much of each engine is unique to it

Measured, Volume III. Across the three-engine set, 40.6% of Claude's vendor slots were unique to Claude, 49.6% of Perplexity's were unique to Perplexity, and 23.2% of Google AI Overviews' were unique to it. Across the two-engine set, 64.1% of Claude's slots were unique against Google AI Overviews' 34.7%. **Reasoned.** The engine returning the longest lists contributes the most names nobody else names, and the engine returning the shortest lists contributes the fewest. Uniqueness is therefore partly a property of list length rather than of independent judgement, which is a caution to carry into any claim that one engine "finds" vendors others miss.

What a blended multi-engine score hides

Reasoned. Averaging across engines produces a number that no engine would have produced, and it hides the case that matters most. A company saturated on one engine and absent on another can blend to the same mid-range score as a company that is mediocre everywhere, and those two companies need completely different work. Given that 14 of 32 companies here were visible on some engines and invisible on others, a blend is not a summary of the situation for nearly half the sample. It is a description of a company that does not exist.

The company-level result is the commercial one

Measured, Volume III. Of 32 measured companies with questions in their category, all 32 were tested on at least two engines. 14 were visible on some engines and invisible on others, 18 were named at least once by every engine that answered for them, and none were named by no engine at all. **Reasoned.** 14 of 32 is not an edge case. For those companies the answer to "are we visible in AI search" is genuinely engine-dependent, and a single-engine report would deliver a confident verdict that a different single-engine report would contradict.

The tie back to the single-engine baseline

Measured, Volume III. Against the Volume I single-engine baseline, the correlation across 32 companies was Pearson r = 0.70, joined at category level because Volume I published company-level results anonymised. Volume III states that a category-level join is weaker than a per-company join. **Reasoned.** A correlation of 0.70 means the single-engine baseline carried real signal about multi-engine visibility and was far from a complete description of it. Substantial but incomplete is consistent with the 14 of 32 result rather than in tension with it.

Citation behaviour differs at least as much as naming

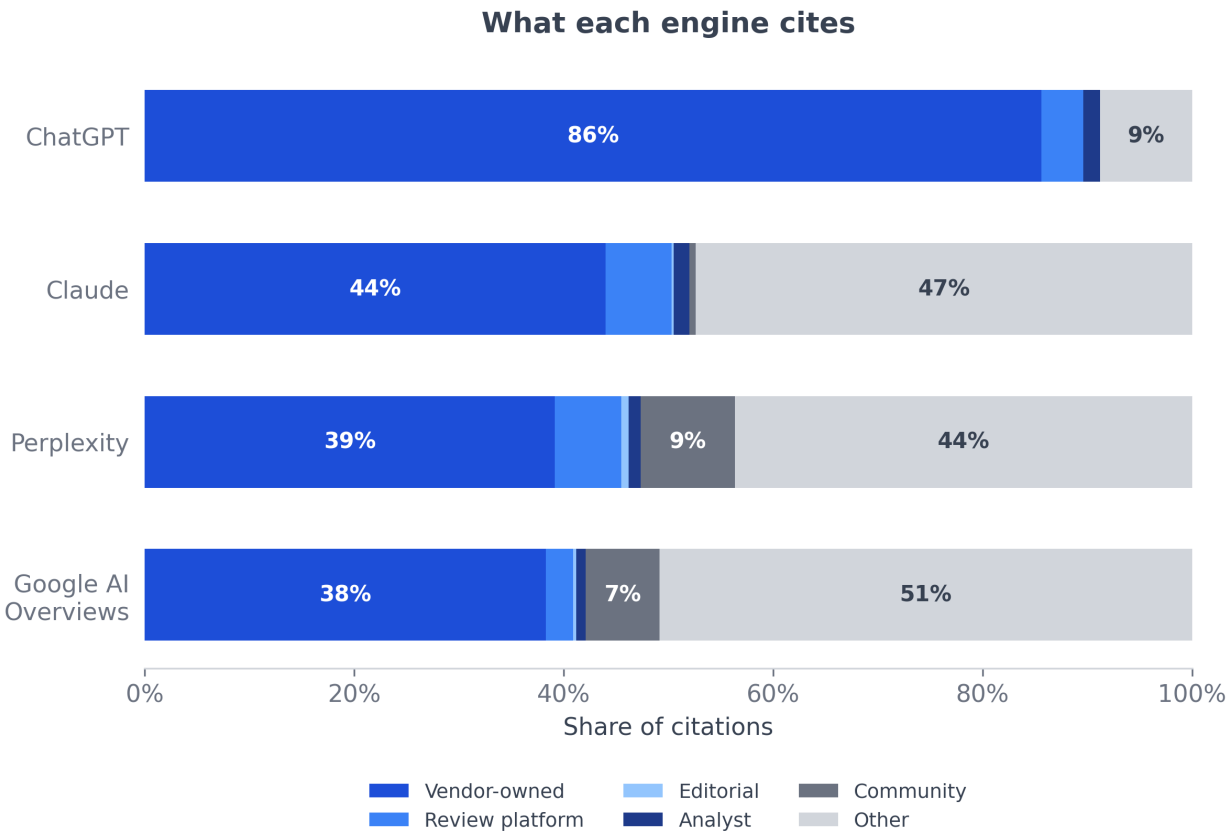


Figure 2. Citation source mix per engine. Volume III.

Measured, Volume III. Vendor-owned share of citations was 85.6% on ChatGPT over 125 citations, 44.0% on Claude over 3,563, 39.2% on Perplexity over 3,016 and 38.3% on Google AI Overviews over 2,849. Community sources ran from 0.0% on ChatGPT to 9.0% on Perplexity. **Reasoned.** The composition of the evidence layer is engine-specific too, so a citation-mix statistic from one engine does not transfer either. Chapter 4 at /who-gets-cited/ reports the single-engine composition that this table splits apart.

Citations per answer differ by a factor of three

Measured, Volume III. Mean citations per answer were 6.9 for ChatGPT, 12.7 for Claude, 19.2 for Perplexity and 11.0 for Google AI Overviews. **Reasoned.** An engine that cites nineteen sources per answer is reaching into a materially different part of the evidence layer than one citing seven. That is another axis on which a single-engine measurement fails to generalise, and it means being present in the sources one engine favours is no guarantee of being present in the sources another one reaches for.

Where independent work lands

Reported. Jack, Lehman, Maloney and Xu compute per-prompt cross-provider overlap of recommended brand sets over commercially framed prompts and report roughly 0.35, in work covering two providers as model pools rather than production search products, with no Google surface and no released dataset (arXiv:2606.26116). **Measured, Volume III.** This study's overall pairwise figure is 0.313. **Reasoned.** Two independent measurements on

different systems and different corpora landing in the same band is weak but real corroboration that the low agreement is a property of the systems rather than an artefact of one instrument.

Limitation: the collection was interrupted

Measured, Volume III. Three of the four API accounts ran out of credit during collection. One was topped up and finished, two were not. The study as executed holds 18 ChatGPT, 280 Claude, 175 Perplexity and 280 Google AI Overviews answers, against 280 planned for each. **Reasoned.** Every ChatGPT pair therefore rests on 18 questions and carries wide uncertainty, and the four-engine intersection is reported at its true size rather than dressed up. Volume III's answer counts do not reconcile across the different tables it publishes, which is recorded as an open flag in this manual and not reconciled here.

Limitation: one window, and endpoints rather than apps

Measured, Volume III. Everything was collected between 05:28 and 16:29 UTC on 2026-08-04, and the engines are production endpoints rather than the consumer applications people type into. Google AI Overviews has no official API and is collected by scraping. **Reasoned.** Results describe what these endpoints returned in that window. The retrieval layer, system prompt and ranking logic of the consumer products are not observable from outside, and this manual does not claim otherwise.

What this chapter does not claim

Reasoned. It does not claim any engine is better, more accurate or more commercially valuable than another. It does not claim the four engines have equal buyer share, because nothing here observes buyers. It does not claim the divergence figures would replicate on a different question set or in a different month. And it does not claim that measuring four engines is always necessary, only that a claim about AI search made from one engine is a claim the evidence does not support.

What this means for your buying decision

Reasoned. Ask which engines a score covers and refuse to accept "AI search" as an answer. Ask for the per-engine breakdown rather than a blend, because a blend can hide a company that is saturated on one engine and absent on another. If a supplier measures one engine, ask them to say so on the front of the report rather than in a footnote, and price the work accordingly. Chapter 12 at /how-to-buy-geo/ sets out the full question list.

Where to go next

Reasoned. Chapter 6 at /measurement-noise/ asks whether the divergence measured here is larger than each engine's own run-to-run variance, which is the test that decides how much of this chapter is signal. Chapter 7 at /overview-trigger-rate/ covers the questions on which one of these engines returns no answer at all. Chapter 8 at /valid-measurement/ turns all of it into criteria a measurement has to meet.

Sources

Every figure in this chapter comes from The 2026 State of GEO, Volume III: the pairwise overlap matrix and its sample sizes, the 0.313 overall mean and its interval, the consensus shares, the per-engine vendor counts and answer lengths, the Fleiss kappa values, the universe-only sensitivity analysis, the company-level 14 of 32 result, the correlation against the Volume I baseline, and the per-engine citation table. External work is cited by author and identifier and is reproduced from Volume III's reference list. Data and code are at github.com/Broadcastwell/state-of-geo-2026 under `volume-iii/`.

Chapter 6. Your Score Is Noisier Than Your Vendor Admits

The claim this chapter defends

Measured, Volume III. Asked the identical buyer question three times inside one collection window, an AI search engine agrees with roughly half of its own previous shortlist: mean vendor-set agreement of 0.499 for Google AI Overviews across 62 repeat pairs and 0.442 for Claude across 74 repeat pairs, on a stratified subsample of 25 questions. **Reasoned.** Every single-run visibility number sold in this industry carries that variance and cannot show it. Any measured difference between two engines contains an unknown amount of the same variation, so a divergence headline reported without a noise floor measured on the same questions in the same window is not interpretable.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

These systems are not deterministic

Reasoned. Ask the same engine the same question twice and the shortlist moves. This is not a defect, it is how the products work: retrieval is live, ranking is probabilistic, and the answer is generated rather than looked up. The consequence for measurement is severe and mostly unstated. Any measured difference between two engines therefore contains an unknown amount of the same variation you would get by asking one engine twice. Unless the noise floor is measured on the same questions in the same window, a divergence headline is not a finding about engines. It is a number with an unstated denominator.

What was measured

Measured, Volume III. 280 B2B software buyer questions across 40 categories, put to four production AI search engines: ChatGPT, Claude, Perplexity and Google AI Overviews. Collection ran in a single window, 05:28 to 16:29 UTC on 2026-08-04. From each answer the study extracted the set of vendors named, then measured how much two vendor sets overlapped. The categories and companies are inherited from Volume I, so the results tie back to that single-engine baseline rather than starting from a fresh and incomparable frame.

The repeat subsample is the whole design

Measured, Volume III. A stratified subsample of 25 questions, spread across the 40 categories and balanced across question shapes, was run three times on all four engines. **Reasoned.** That subsample is what makes the rest of the study readable. Between-engine agreement is compared against within-engine agreement on exactly those questions, on the

same scale, from the same window, measured by the same instrument. Without it there is no way to tell a real engine difference from ordinary nondeterminism, and every cross-engine comparison published without one has that problem whether or not it says so.

What went wrong during collection, stated plainly

Measured, Volume III. Three of the four API accounts ran out of credit during collection. One was topped up and finished. Two were not. The study as executed therefore holds 18 ChatGPT, 280 Claude, 175 Perplexity and 280 Google AI Overviews answers, against 280 planned for each. Rows carrying a billing error are excluded from every denominator in the paper and are retained in the raw data with an error status, so the exclusion is checkable by anyone who wants to check it rather than taken on trust.

Why that disclosure is a feature and not an embarrassment

Measured, Volume III. The two engines that did not finish were deliberately not refilled. A refill collected in a second window would make every cross-engine pair span two windows, so engine drift between windows would read as engine divergence and inflate the exact quantity the study measures. Staying short was the more honest option. The four-engine intersection is reported at its true size of 18 questions rather than dressed up, and every figure carries the sample it was computed on. **Reasoned.** This disclosure is the reason the chapter can be trusted, not a caveat on it.

The decisive test, reported per engine and never pooled

Measured, Volume III. For each engine the comparison is like for like on the same 25 stability questions. *Within* is every pair of repeats of the same question on that engine. *Between* is that same engine against the others on the first run of those same questions. An engine needs at least 12 repeat pairs before a verdict is asserted for it.

Engine	Repeat pairs	Within-engine	Between-engine	Gap (95% CI)
Claude	74	0.442 (0.389 to 0.494)	0.277 (0.222 to 0.333), n = 37	0.165 (0.082 to 0.242)
Google AI Overviews	62	0.499 (0.441 to 0.556)	0.240 (0.186 to 0.295), n = 35	0.258 (0.178 to 0.338)

Measured, Volume III. ChatGPT and Perplexity contributed no repeat pairs and therefore no verdict. That is stated rather than hidden.

Engines name very different numbers of vendors

Engine	Mean vendors named per answer
Google AI Overviews	4.77
Claude	8.83
Perplexity	10.69
ChatGPT	14.72

Reasoned. That spread is not a curiosity. It is a threat to the headline. Set-overlap between two similar-length lists runs mechanically higher than between two lists of very different length, so part of any within-versus-between gap could be list length rather than divergence.

The asymmetry that has to be controlled

Reasoned. Within-engine pairs compare an engine with itself, so they are length-matched by construction: the same engine returns roughly the same number of vendors on every run. Between-engine pairs are not length-matched at all, because they compare an engine that names five vendors against one that names nine or fifteen. That asymmetry systematically depresses the between-engine figure for reasons that have nothing to do with which vendors the engines picked. Left uncontrolled, it would manufacture a gap out of arithmetic. This is the chapter's methodological spine.

Three controls, each recomputed from the same pairs

Measured, Volume III. First, restrict both vendor sets to the study universe, which removes every out-of-universe name and narrows the length spread. Second, truncate each answer to its first five named vendors, which forces both sides of every comparison to the same maximum length. Third, replace the overlap measure with a coefficient normalised by the smaller of the two sets, which is the least forgiving of a length artefact because it cannot be depressed by one side simply being longer. All three are recomputed from the same repeat pairs, not from a fresh sample.

The split result, reported honestly because it is the point

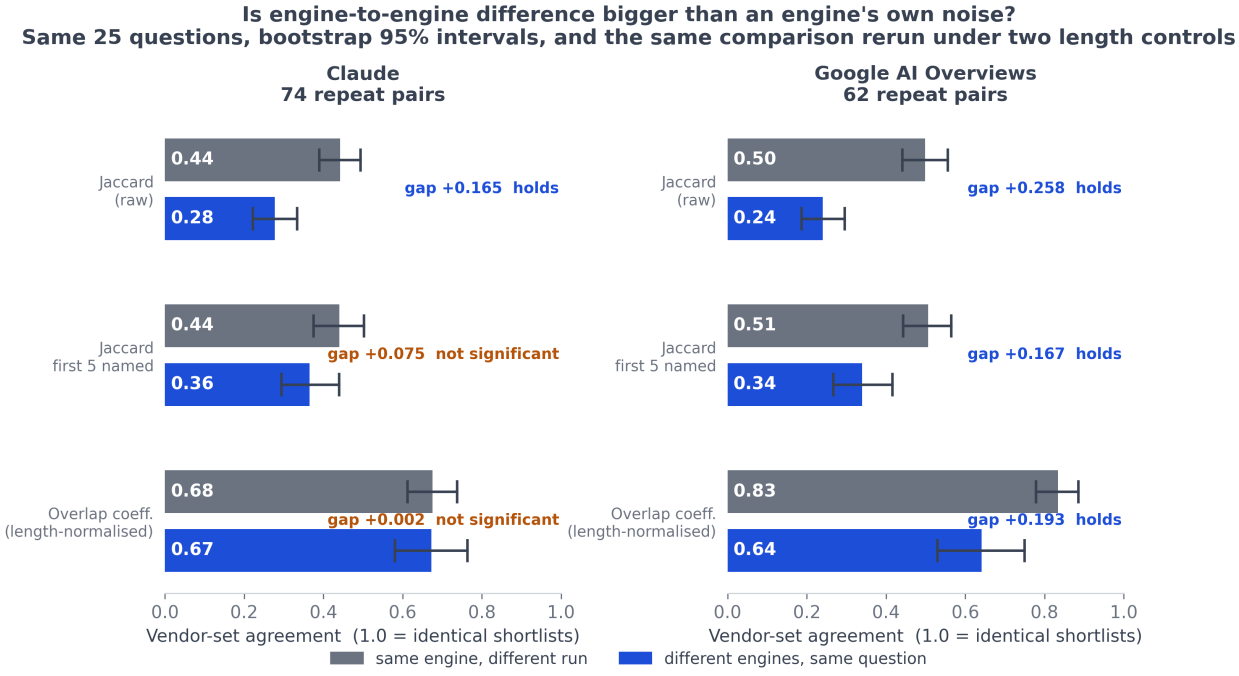


Figure 1. Within-engine versus between-engine agreement, per engine, under three measures. Volume III, same 25 questions, bootstrap 95% intervals.

Measured, Volume III. On Google AI Overviews the gap survives all three controls. On Claude it does not.

The full robustness table

Engine	Measure	Within	Between	Gap (95% CI)	Separates
Claude	Jaccard, full vendor sets	0.442 (n=74)	0.277 (n=37)	0.165 (0.082 to 0.242)	yes
Claude	Jaccard, restricted to the study universe	0.648 (n=44)	0.458 (n=24)	0.189 (−0.030 to 0.403)	no
Claude	Jaccard, truncated to the first 5 named vendors	0.440 (n=74)	0.365 (n=37)	0.075 (−0.014 to 0.177)	no
Claude	Overlap coefficient, normalised by the smaller set	0.675 (n=70)	0.673 (n=33)	0.002 (−0.112 to 0.113)	no
Google AI Overviews	Jaccard, full vendor sets	0.499 (n=62)	0.240 (n=35)	0.258 (0.178 to 0.338)	yes
Google AI Overviews	Jaccard, restricted to the study universe	0.625 (n=20)	0.300 (n=20)	0.325 (0.050 to 0.575)	yes
Google AI Overviews	Jaccard, truncated to the first 5 named vendors	0.507 (n=62)	0.340 (n=35)	0.167 (0.066 to 0.262)	yes
Google AI Overviews	Overlap coefficient, normalised by the smaller set	0.834 (n=58)	0.641 (n=32)	0.193 (0.075 to 0.317)	yes

What the Claude column means

Measured, Volume III. The raw gap of 0.165 falls to 0.075 with a confidence interval of −0.014 to 0.177 under truncation, and to 0.002 with an interval of −0.112 to 0.113 under the overlap coefficient. Both intervals include zero. **Reasoned.** On Claude, what looks like cross-engine divergence cannot be distinguished from the fact that the engines being compared name lists of different lengths. That is the finding, not the finding the study set out to get, and it is reported because the controls were specified in advance rather than chosen after seeing which ones were flattering.

What the Google AI Overviews column means

Measured, Volume III. For Google AI Overviews the gap stays positive at 95% under every control, including the overlap coefficient, which normalises by the smaller of the two sets and is the hardest of the three to fool. **Reasoned.** The separation there is not length. It is a real difference between what Google AI Overviews returns and what the other engines return, larger than Google AI Overviews' own run-to-run noise on the same questions in the same window. Divergence survives a length control on one engine and vanishes on another, and reporting only the survivor would have been the easy mistake.

Why the pooled figure is not the headline

Measured, Volume III. Pooled across all engines, within-engine agreement is 0.468 (95% CI 0.427 to 0.505) over 136 repeat pairs, against between-engine agreement of 0.247 (95% CI 0.201 to 0.298) over 51 pairs. That looks like a clean separation and it is not one, because 54.4% of those within-engine pairs are Claude. **Reasoned.** A pooled mean is dominated by whichever engine contributed the most pairs and can report a separation that only one engine actually has. That is an unstated denominator, which is precisely the failure this chapter exists to name.

The sentence buyers actually need

Reasoned. There is a second reading of the within-engine column and it matters more commercially than the divergence result does. **Measured, Volume III.** Even the best case here is an engine agreeing with itself on roughly half its own shortlist across runs of the identical question, at 0.499 for Google AI Overviews and 0.442 for Claude. **Reasoned.** Every single-run visibility number in this industry, including the one in Volume I of this research programme, carries that variance and cannot show it. The same company measured twice can produce two different scores, and a single-run number cannot tell you how far apart the two could have been.

What this means if you are buying a visibility score

Reasoned. Anyone selling a number described as "your AI search visibility" without naming the engine, the date and the question set is selling a number whose denominator is unstated. Those three are not optional metadata. The engine determines which product was measured. The date determines which version of a continuously changing product was measured. The question set determines what "visibility" even means, because a company visible on use-case questions and invisible on best-of shortlists will score anywhere between 2 and 8 depending on the mix.

What a defensible score looks like

Reasoned. A defensible visibility measurement names its engine, states its collection window, publishes its question set, reports how many runs each question got, and either measures its own run-to-run variance or says openly that it did not. None of that is expensive. The reason it is rare is not cost. Chapter 12 of this manual turns this into a list of questions to put to a vendor before signing, and every one of them is answerable in a sentence by anyone doing the work properly.

What this chapter does not claim

Reasoned. It does not claim that engines are unreliable in a way that makes measurement pointless. Variance is not noise-in-the-pejorative-sense, it is a property of the system that can be quantified and reported, and this chapter quantifies it. It also does not claim that the two engines with no repeat data behave like the two that have it. **Measured, Volume III.** ChatGPT and Perplexity contributed zero repeat pairs. Nothing here licenses a statement about their run-to-run stability in either direction.

Limitation: a single collection window

Measured, Volume III. Everything here was collected between 05:28 and 16:29 UTC on 2026-08-04. **Reasoned.** These are products under continuous change. The numbers are a measurement of that window and should not be read as stable constants. A repeat of this study in three months could return materially different agreement figures without anything being wrong with either measurement, which is itself part of the argument the chapter is making.

Limitation: Google AI Overviews is scraped, not API-served

Measured, Volume III. There is no official API for Google AI Overviews, so it is collected by scraping rather than by a served endpoint. **Reasoned.** That introduces a dependency whose failure modes are not fully observable from the measurement side. **Measured, Volume III.** Where no overview was returned, the result is recorded with its own status rather than as an error, and treated as Google not producing one. Google returned an AI Overview for 92.1% of these buyer questions. On the remaining 22 there was no AI answer to be visible in at all.

Limitation: engines here are products and endpoints, not consumer apps

Measured, Volume III. The engines measured are production endpoints, not the consumer applications people actually type into. The retrieval layer, system prompt and ranking logic of the consumer products are not observable from outside. Results describe what these endpoints return and not what a person sees in an app. **Reasoned.** That distinction is routinely collapsed in industry reporting and it should not be, because the two can differ in ways nobody outside the providers can measure.

Limitation: the sampling frame generalises to nothing outside its own scope

Measured, Volume III. The categories and companies are inherited from Volume I, which was itself a convenience sample of B2B software categories. **Reasoned.** Nothing here generalises to consumer categories, to non-English queries, or to markets outside the United States English locale used for collection. The finding about run-to-run variance is likely to be general because it follows from how the systems work, but this study measured it in one place and says so.

Do this yourself: measure your own noise floor

Reasoned. Take five of your buyer questions. Put each one to the same engine three times, on the same day, within a couple of hours. Write down the set of vendors named in each answer. For each question, compare run one against run two, run one against run three, and run two against run three, and count how many vendors appear in both sets divided by how many appear in either. Average those. That average is your engine's noise floor on your questions, and it is the number every other visibility figure you are shown should be read against.

Do this yourself: audit the score you were sold

Reasoned. Ask four questions of whoever gave you a visibility number. Which engine produced it. On what date, in what window. What exactly were the questions, in full text. How many runs did each question get, and if the answer is one, what is the run-to-run variance on those questions. A supplier doing the work properly answers all four immediately. A supplier who cannot answer the fourth has sold you a point estimate from a distribution and has not told you the width of it.

Where to go next

Reasoned. If you want the requirements a measurement has to meet before it is worth reading at all, Chapter 8 at /valid-measurement/ turns the variance in this chapter into a specification. If you want the cross-engine picture that the single-engine noise floor sits inside, read Chapter 5 at /engine-divergence/. If you are about to buy this work from somebody, Chapter 12 at /how-to-buy-geo/ turns the four questions above into the full set to put to a supplier.

Sources and reproduction

Every number in this chapter comes from The 2026 State of GEO, Volume III, published with all collected answers, the extraction prompt, the scoring rules and the analysis scripts under CC BY 4.0. The bootstrap uses 2000 replicates with a fixed seed, so the confidence intervals reproduce exactly, and every figure is generated from the computed results file rather than typed, so a caption cannot drift from the dataset. All of it is at github.com/Broadcastwell/state-of-geo-2026 under `volume-iii/`.

Chapter 7. The AI Overview does not always appear

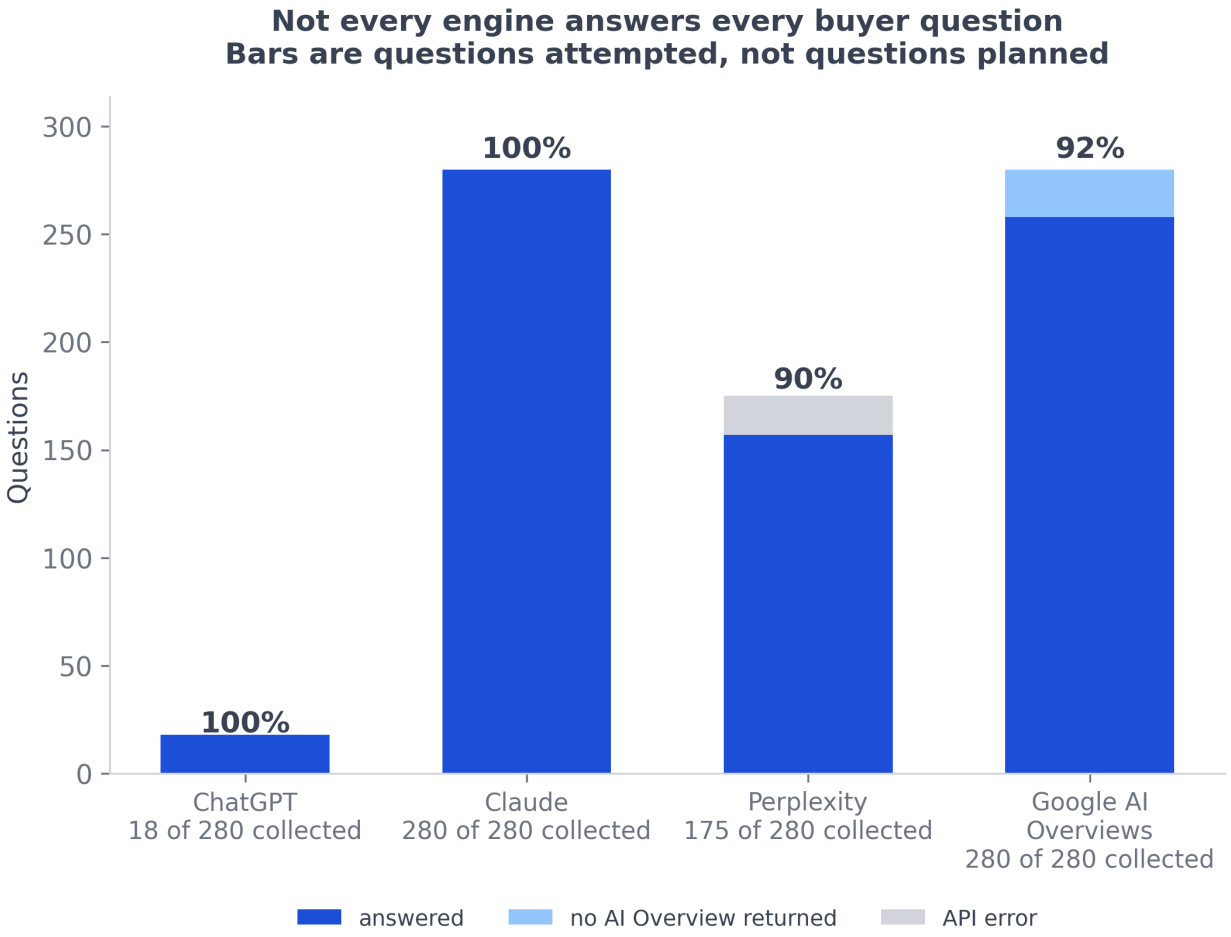
By the end of this chapter you will know how often one major engine declined to generate an answer at all, how much that rate moved with the shape of the question, and why a visibility percentage is meaningless until you know which of two denominators it used.

The claim this chapter defends

Measured, Volume III. Google returned an AI Overview for 92.1% of these buyer questions, 258 of 280. On the remaining 22 there was no AI answer to be visible in at all. **Measured, Volume III.** The rate moved with question shape, from 88.8% on best-of questions to 100% on comparison questions. **Reasoned.** A score computed over the answers that appeared is not the same quantity as a score computed over the questions asked, and the gap between the two is set by a trigger rate that the vendor being measured does not control.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

Two different reasons an answer can be missing



Three of the four accounts ran out of API credit during collection, which is why the bars differ in height.

Figure 1. Answered against attempted, per engine. Volume III.

Measured, Volume III. Volume III records a Google query that returns no AI Overview under its own status, `no_aio`, separately from an API error. **Reasoned.** The distinction is load bearing. An error is a fault in the measurement. A declined answer is a result about the engine, and treating the two the same way corrupts both the numerator and the denominator of anything computed downstream.

The trigger rate is a property of the question, not the vendor

Measured, Volume III. By question shape, an AI Overview was returned for 72 of 80 alternatives questions at 90.0%, 71 of 80 best-of questions at 88.8%, 47 of 47 comparison questions at 100%, 34 of 38 evaluation questions at 89.5%, and 32 of 33 use-case questions at 97.0%. **Reasoned.** Comparison questions triggered an overview every single time. Best-of questions, the shape most people picture when they imagine an AI shortlist, triggered least often. No vendor influenced any of that, and yet it determines how many chances that vendor had to appear.

The worked illustration, using only published counts

Measured, Volume III. Of 80 best-of questions, 71 returned an AI Overview. **Reasoned.** So a vendor measured on best-of questions has the questions that returned no answer silently removed if the measurement counts only the answers that appeared. Its score is then a fraction of 71 rather than of 80. On comparison questions no such removal happens, because all 47 returned an answer. The identical vendor, measured on two shapes with the same underlying performance, will show a different score purely because the two shapes lose different numbers of questions from the denominator.

Which denominator is right depends on the question you are asking

Reasoned. Both denominators are defensible and they answer different things. Over answers that appeared, the figure tells you how well the vendor competes when there is an answer to compete in. Over questions asked, it tells you how often a buyer asking that question sees the vendor at all. The second is closer to the commercial reality and it is the smaller number, which is why suppliers reporting a single figure have an incentive toward the first. Volume III states the failure mode plainly: a score that silently drops those questions from its denominator will overstate coverage.

The shape mix of a question bank becomes a lever

Reasoned. Because trigger rates differ by shape and this manual's own bank contains several shapes, the composition of the bank changes the headline trigger rate and therefore the size of the denominator effect. A bank weighted toward comparison questions will show a higher trigger rate and a smaller gap between the two denominators than one weighted toward best-of. That makes the shape mix a disclosure requirement rather than an implementation detail, and Appendix A at /question-bank/ sets out what a bank has to declare.

The other engines answered everything they attempted

Measured, Volume III. ChatGPT answered 18 of 18 attempted and Claude 280 of 280, both at 100%. Perplexity answered 157 of the questions it attempted, at an answer rate of 89.7%. **Reasoned.** Only the Google surface declined to generate. The conversational engines answered whenever they were reachable, so the trigger-rate problem in this dataset is specific to the AI Overview surface rather than general to AI search. A measurement covering only conversational engines will not encounter it, and a measurement including Google will.

Where the collection shortfall sits, and what it is not

Measured, Volume III. Three of the four API accounts ran out of credit during collection, which is why the attempted counts differ between engines. Rows carrying a billing error are excluded from every denominator and retained in the raw data with an error status. **Reasoned.** That is a fault in the measurement and Volume III separates it from the `no_aio`

result, which is a fact about the engine. This manual records Volume III's differing answer counts as an open flag and quotes whichever figure belongs to the claim being made without asserting any relation between them.

What the wider literature reports, and how it compares

Reported. Seer Interactive reports AI answer trigger rates of 95.4% for comparison queries over 280 queries and 85.9% for question-format queries over 1,413, published April 2026 on data running to February 2026. **Measured, Volume III.** This study's 92.1% sits below Seer's general-population rates. **Reasoned.** That is the direction you would expect: narrow B2B software buying questions are not consumer queries, and a narrower intent seems to trigger an overview less reliably. Two independent measurements finding comparison questions at or near the top of the trigger ordering is a small piece of mutual corroboration.

The rate is moving, which makes the date part of the figure

Reported. Semrush reports that the share of commercial search results carrying an AI Overview grew 71 percent between November 2025 and April 2026, published 2 July 2026. **Reasoned.** A trigger rate measured in one month is therefore not a constant to be reused later. Any visibility score whose denominator depends on the trigger rate inherits that drift, so two scores computed months apart are not comparable even if the question set, the engine and the vendor are all identical. This is a second reason, on top of run-to-run variance, that a score without a date is not a score.

Trigger rate compounds with slot scarcity

Measured, Volume III. Google AI Overviews returned an answer on 258 of 280 questions and named a mean of 4.77 vendors when it did, the smallest set of the four engines. **Reasoned.** Two structural constraints therefore stack on the same surface: on 22 questions there was no answer to be in, and on the rest there were fewer places in it than any other engine offered. Neither constraint is about the vendor. Both reduce the number of opportunities before merit is considered, and a score that reports only the outcome attributes the whole of that reduction to the vendor's performance.

What this does to a trend line

Reasoned. If the denominator can move between measurements, a visibility trend can rise or fall without the numerator changing at all. A vendor named in the same absolute number of answers will show a higher percentage in a month when fewer questions triggered an answer, because the denominator shrank underneath it. Anybody presenting a month-on-month visibility chart therefore owes you the denominator for each point on it, and a chart with a single series and no denominators cannot be read as improvement or decline.

The surface is not the whole page

Reasoned. On a question where no AI Overview appears, the buyer still sees a results page and still finds vendors on it. The absence of an overview is not the absence of an outcome, it is the absence of the specific surface being measured. This matters for how a score should

be described: it is a measurement of presence in one generated surface, not a measurement of whether a buyer researching that question encounters the vendor. Conflating the two overstates what AI visibility work can be expected to do.

Coverage and performance are different quantities

Reasoned. Splitting the two makes both readable. Coverage is the share of your questions on which an answer exists at all. Performance is the share of the existing answers that name you. A vendor with strong performance and weak coverage has a different problem from one with the reverse, and the combined percentage cannot distinguish them. Reporting the pair costs nothing beyond keeping the `no_aio` count, which any measurement already has if it bothered to record it.

What a buyer should be able to see

Reasoned. Three counts, published together: questions asked, answers returned, and questions on which the vendor was named. Everything else in a visibility report can be derived from those three, and no report that omits the middle one can be checked. If a supplier cannot supply the second count, they either did not record whether an answer appeared or are computing over a denominator they have not disclosed, and the two are indistinguishable from outside.

Limitation: one surface, one window, scraped collection

Measured, Volume III. The trigger figures cover one engine's surface, collected between 05:28 and 16:29 UTC on 2026-08-04, and Google AI Overviews has no official API so it is collected by scraping. A `no_aio` result means the collection returned no overview, which the study treats as Google not producing one. **Reasoned.** That treatment is reasonable and not directly verifiable from outside, so the trigger rate carries an unmeasured dependency on the collection path. Volume III says so, and a reader should discount accordingly rather than treating 92.1% as an exact property of the engine.

Limitation: the shape counts are small in places

Measured, Volume III. The shape breakdown rests on 80 alternatives questions, 80 best-of, 47 comparison, 38 evaluation, 33 use-case and 2 other. **Reasoned.** The comparison result of 47 out of 47 is a clean 100% on a sample of 47, which is enough to say comparison questions triggered reliably and not enough to say they always will. The evaluation and use-case cells are smaller again. Read the ordering of the shapes as the finding and the individual decimals as indicative.

Why this is the easiest of the four disclosures to demand

Reasoned. Of everything this manual asks a buyer to require, the answered count is the cheapest to supply. Recording whether an answer came back costs nothing at collection time, needs no extra calls, and is already in any raw log worth keeping. A supplier who cannot produce it has either discarded it deliberately or never captured it, and both are informative about the rest of the measurement. Treat the answered count as a competence test rather than as a favour you are asking for.

What this chapter does not claim

Reasoned. It does not claim to know why Google declines to generate on some questions, because nothing here observes the decision. It does not claim the 22 questions share any property beyond the shape breakdown reported. It does not claim other engines will develop the same behaviour. And it does not claim that a missing overview means no AI presence for the buyer, because the same buyer may ask the same question elsewhere and Chapter 5 at /engine-divergence/ shows how different the answer can be.

What this means for your buying decision

Reasoned. Ask for three counts rather than one percentage: questions asked, answers returned, and answers naming you. Ask which denominator the headline figure uses and require it in writing, because the two differ by however many questions returned no answer. If a supplier cannot tell you how many of your questions produced no AI answer at all, they cannot tell you what their percentage is a percentage of. Chapter 12 at /how-to-buy-geo/ has the full list of questions to ask.

Where to go next

Reasoned. Chapter 8 at /valid-measurement/ turns the denominator requirement into one criterion among several that a measurement has to meet. Chapter 6 at /measurement-noise/ covers the other reason a single figure moves without the vendor changing. Appendix A at /question-bank/ sets out the question shapes whose mix determines the trigger rate in the first place.

Sources

Every figure in this chapter comes from The 2026 State of GEO, Volume III: the 92.1% trigger rate and its 258 of 280 base, the per-shape breakdown, the per-engine attempted and answered counts, the separation of `no_aio` from API error, and the collection window. External trigger-rate figures are attributed to Seer Interactive and Semrush with their publication dates and are reproduced from Volume III's reference list. Data and code are at github.com/Broadcastwell/state-of-geo-2026 under `volume-iii/`.

Chapter 8. What a valid measurement requires

By the end of this chapter you will have a list of requirements you can hold any supplier's visibility number against, and a shorter list of failure modes that make a number meaningless regardless of how it was produced. This chapter is written as criteria for a buyer, not as a procedure for an operator.

The claim this chapter defends

Reasoned. A visibility number is only interpretable if nine things are true of it: the question bank was fixed before collection, published in full and covering the shapes that matter, each question was run more than once, the engine set is named, the scoring rule is deterministic and stated, the denominator is declared, every exclusion is disclosed, and the operator's own commercial position is declared. **Measured, Volume III.** Each of those requirements exists because a published measurement demonstrated what goes wrong without it, and the failure modes at the end of this chapter are all observed rather than hypothetical.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

Requirement one: the question bank is fixed before collection

Measured, Volume III. Of 280 questions, 276 came from the study's existing question bank and 4 from the published Volume I dataset. No question was regenerated. **Reasoned.** A bank that can be regenerated between measurements is a bank that can be tuned, deliberately or not, toward questions the subject happens to do well on. Fixing it in advance removes that degree of freedom. The test for a buyer is simple: ask whether the questions used this quarter are byte-identical to the questions used last quarter, and ask to see both.

Requirement two: the bank is published in full

Reasoned. Publishing the bank is what makes a score checkable by somebody who did not run it. Without the question text, a reader cannot tell whether a high score reflects broad presence or a bank weighted toward definitional questions where almost everyone appears. It also prevents the quiet substitution of brand-name questions, which measure whether an engine knows a company exists rather than whether it competes. Appendix A at /question-bank/ sets out the design rules at the level the source volumes publish them.

Requirement three: the bank covers the shapes that matter

Measured, Volume III. Each category was required to carry at least one best-of, one alternatives, one comparison, one use-case and one pricing or evaluation question. 33 of 40 categories met that in full, and the 7 that did not are listed explicitly in the published

sample manifest rather than quietly padded. **Reasoned.** Shape coverage matters because Chapter 2 at /absence-ladder/ shows absence concentrating in different shapes at different visibility levels. A bank missing a shape cannot detect the gate that shape diagnoses.

Why listing the seven gaps is the point

Reasoned. A study that declares 33 of 40 rather than reporting full coverage is telling you where its own instrument is incomplete. That disclosure costs nothing to make and is almost never made, which is why its presence is a useful signal about everything else in the report. When a supplier says their coverage is complete, ask which categories were short and what they did about it. Silence there usually means padding rather than perfection.

Requirement four: every question is run more than once

Measured, Volume III. A stratified subsample of 25 questions, balanced across shapes, was run three times on all four engines, and an engine needed at least 12 repeat pairs before a verdict was asserted for it. **Measured, Volume III.** Within-engine agreement came out at 0.499 for Google AI Overviews over 62 repeat pairs and 0.442 for Claude over 74. **Reasoned.** Engines agree with roughly half of their own previous shortlist on a repeated identical question. One run therefore samples a distribution and reports a point, and the point is presented as though it were the distribution.

How many repeats is enough

Reasoned. Three runs on a stratified subsample is the minimum this manual can point at a published precedent for, and it is a floor rather than a target. What matters more than the count is that the repeats exist, that they are run in the same window as the main collection, and that the variance they reveal is reported alongside the headline rather than in an appendix. A supplier who runs everything once and offers to run more for a fee has priced a correctness requirement as an upgrade.

Reported. Schulte, Bleeker and Kaufmann argue that a single visibility measurement is close to meaningless because the same query produces a distribution rather than a value, and that visibility should be measured repeatedly (arXiv:2604.07585).

Requirement five: the engine set is named on the front page

Measured, Volume III. Of 32 measured companies tested on at least two engines, 14 were visible on some and invisible on others. **Reasoned.** For nearly half the sample the verdict depends on which engine was asked, so a report headed "AI search visibility" without an engine list is not describing a measurable quantity. The disclosure has to be prominent rather than a footnote, because a reader who takes the headline at face value will act on a claim the data does not support. Chapter 5 at /engine-divergence/ has the underlying numbers.

Requirement six: the scoring rule is deterministic and published

Measured, Volume III. Brand matching is word-boundary and case-sensitive when the brand is a single plain alphabetic word, so a company called Pitch does not match the ordinary verb, while a brand carrying a dot, digit, hyphen or space matches case-

insensitively. Domain matching is at hostname level with multi-domain values split, so subdomains count and near-misses do not. **Reasoned.** Rules like these are boring and they are the difference between a score two people can reproduce and a score that depends on who ran it. Appendix C at /scoring-spec/ reproduces the specification at the level the source publishes it.

Where a language model may and may not be used in scoring

Measured, Volume III. The deterministic layer is authoritative for every company in the study universe. A second, model-based extraction layer exists only to catch vendors outside that universe, and any claim it makes about a universe company is discarded unless the deterministic layer also finds the string in the answer. **Reasoned.** That ordering is the safeguard: it means the model-based layer can widen coverage but cannot inflate or deflate a measured company's numbers in either direction. A supplier whose headline score depends on a model's judgement about whether you were mentioned has no such safeguard.

Requirement seven: the denominator is declared

Measured, Volume III. Google returned an AI Overview on 258 of 280 questions. **Reasoned.** A score computed over answers that appeared is a different quantity from a score computed over questions asked, and the two differ by however many questions produced no answer. Any report should carry three counts, questions asked, answers returned and answers naming the subject, from which everything else can be derived. Chapter 7 at /overview-trigger-rate/ works through the size of the effect.

Requirement eight: exclusions are disclosed and the raw rows retained

Measured, Volume III. Rows carrying a billing error were excluded from every denominator and retained in the raw data with an error status, so the exclusion is checkable. **Reasoned.** Every measurement drops something. The distinction between an honest measurement and a flattering one is whether the dropped rows are visible. Ask what was excluded, on what rule, and whether the excluded rows are still in the delivered data. A supplier who cannot answer has either not recorded exclusions or does not want them counted.

Requirement nine: the operator's own position is declared

Measured, Volume III. The study that supplies most of this chapter carries an operator disclosure stating that the firm running the measurement sells services in the category it measures, that it is excluded from the sample and from every ranking, and that the mitigation is publication rather than the absence of the conflict. **Reasoned.** Every commercial visibility measurement has this conflict. The question is not whether it exists but whether it is stated and whether the data is open enough that a sceptic can recompute the result. A supplier who both measures you and sells you the remedy should say so on the first page.

Failure mode zero: the number that has never been checked against itself

Measured, Volume III. The study reports its own repeat-run agreement rather than assuming stability, and reports it per engine rather than pooled. **Reasoned.** The cheapest possible integrity check on any measurement instrument is to point it at the same thing twice and see whether it agrees. A supplier who has never done that with their own tool does not know its precision, which means they cannot tell you whether a change between two reports is a change in your visibility or a change in nothing at all. Ask when they last measured their own repeatability and what the answer was.

Failure mode one: the pooled mean with an unstated denominator

Measured, Volume III. Pooled across engines, within-engine agreement was 0.468 over 136 repeat pairs against between-engine 0.247 over 51 pairs, and 54.4% of those within-engine pairs came from one engine. **Reasoned.** A pooled figure is dominated by whichever engine contributed most, so it can report a separation that only one engine actually has. Volume III declines to headline it for that reason. When a supplier gives you one blended number across engines, ask how many observations each engine contributed.

Failure mode two: the number with no date

Reported. Semrush reports that the share of commercial search results carrying an AI Overview grew 71 percent between November 2025 and April 2026, published 2 July 2026. **Reasoned.** The surfaces being measured are changing fast enough that a figure without a collection window cannot be compared with anything, including a later figure from the same supplier. A date is not metadata on a visibility score. It is part of the value.

Failure mode three: the score that cannot be recomputed

Reasoned. If the question set, the raw answers and the scoring rule are not delivered with the number, nobody can check it, including the buyer paying for it. This is the requirement most commercial tools fail, and it is the one worth being least flexible about, because every other requirement in this chapter is unverifiable without it. A deliverable that consists of a dashboard and no underlying rows is a claim, not a measurement.

What this chapter deliberately does not provide

Reasoned. This is a list of criteria for judging a measurement, not a procedure for producing one. The published protocol behind the three volumes is available in full in the source repository for anyone who wants to replicate it, and the appendices to this manual reproduce it at exactly the level the volumes already publish and no deeper. Nothing here describes how any commercial measurement is operated, and a reader looking for that will not find it in this manual.

The honest limitation of these criteria

Reasoned. These requirements are drawn from what the three volumes did and from what went wrong when parts of it were missing. They are not validated against outcomes, because no dataset here links measurement quality to commercial results. A measurement can satisfy every requirement in this chapter and still be measuring something a buyer does not care about. What the criteria guarantee is that the number means what it says, which is a lower bar than being useful and a necessary one.

What this means for your buying decision

Reasoned. Put the nine requirements to a supplier in writing before you sign, and treat any that cannot be answered in a sentence as a finding rather than a formality. The three that most often fail are repeat runs, the declared denominator, and delivery of the raw rows. If a supplier meets all eight, the number they give you may still be uninteresting, but it will at least be a number. Chapter 12 at /how-to-buy-geo/ turns this list into the specific questions to ask.

Where to go next

Reasoned. Chapter 12 at /how-to-buy-geo/ is the buyer-side companion to this chapter. Appendix A at /question-bank/, Appendix B at /absence-rules/ and Appendix C at /scoring-spec/ reproduce the published protocol the requirements are drawn from. Chapter 6 at /measurement-noise/ is the evidence behind the repeat-run requirement.

Sources

The requirements in this chapter are drawn from the published methods and results of The 2026 State of GEO, Volume III: the fixed question bank and its provenance, the shape coverage requirement and the 33 of 40 disclosure, the repeat subsample and the 12-pair threshold, the within-engine agreement figures, the two-layer scoring rule and its precedence, the trigger rate and its denominators, and the exclusion policy for billing errors. The company-level 14 of 32 result is also Volume III. External work is cited by author and identifier and reproduced from Volume III's reference list. Everything is at github.com/Broadcastwell/state-of-geo-2026.

Chapter 9. Getting named at the category door

By the end of this chapter you will know what the first gate actually tests, which properties of the evidence about a company plausibly determine whether it clears that gate, and one measured result that overturns the most common excuse for not clearing it.

The claim this chapter defends

Measured, Volume II. Companies named in zero of ten answers lost category-level questions, meaning best-of shortlists plus alternatives-to-incumbent queries, 55.3% of the time, across 30 companies and 300 absence records. **Reasoned.** The first gate is membership, not preference. More than half of what a zero-visibility company loses is questions that never mention a competitor by name and simply ask who exists. The work implied is therefore about being recognisable as a member of a set, and this chapter is about what that requires of the evidence rather than about tactics for producing it.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

What the category door actually tests

Measured, Volume II. At the zero tier, 55.3% of absences were category-level and 20.0% were head-to-head comparisons. At the seven-and-above tier those proportions reverse to 12.5% and 68.8%. **Reasoned.** A company failing at the door is not losing an argument against a rival. It is not in the room where the argument happens. The engine, assembling a set of candidates before it writes anything, does not include it. Nothing about product quality, pricing or positioning is being adjudicated at that point, because adjudication happens later.

The strong-incumbent excuse does not survive the data

Figure 1. Challenger visibility against leader visibility. Volume II, 85 companies, reported as a null result.

Measured, Volume II. The Spearman correlation between category leader visibility and challenger visibility was −0.051 at p = 0.64 over 85 companies. **Reasoned.** A dominant incumbent does not crowd the challenger out. The most common explanation offered for zero visibility, that the category is owned by somebody else, has no support here.

The counterintuitive part is which categories are worst

Measured, Volume II. Challenger mean visibility was 2.35 where the leader scored 9 or 10, 3.48 where the leader scored 7 or 8, and 1.57 where the leader scored 6 or below, with a Kruskal-Wallis result across those bands at p = 0.052. **Measured, Volume II.** In seven of the 85 company sweeps no brand scored above 5 of 10, and challengers in those categories did worst. **Reasoned.** Challengers did worst in weakly defined categories, not in strongly dominated ones. Volume II offers the plausible reading that a category with an obvious leader is one the engine models well enough to name a second and third option, and reports it as a null result to be retested rather than as a finding.

What that implies about the first job

Reasoned. If a weakly modelled category suppresses everyone in it, then part of the work at the category door is not about the company at all. It is about whether the category itself is legible: whether it has a stable name, whether sources describe it consistently, and whether a set of vendors is routinely enumerated within it. A company trying to be selected into a set that the engine does not cleanly recognise is solving the wrong problem first. This is inference from a null result and should be held loosely.

Standard one: the entity is unambiguous

Reasoned. An engine assembling candidates has to resolve a name to a company. Names that collide with ordinary words, with other companies, or with product lines create resolution failures that no amount of publishing fixes. Volume III's scoring rules take the same problem seriously from the measurement side: brand matching is case-sensitive for a single plain alphabetic word precisely so that a company called Pitch does not match the verb. If a measurement has to work that hard to identify you, so does a retrieval system.

Standard two: the category attachment exists outside your own pages

Measured, Volume II. Volume II's stated practical consequence for companies at the door is entity presence and inclusion in the third-party sources engines consult when assembling a set. **Reasoned.** The operative words are third-party. A company's own site asserting membership of a category is an assertion. The same claim appearing in sources the company does not control is corroboration, and corroboration is what a set-assembly step can act on without taking one vendor's word for it. This is a standard about where the claim lives, not about how much of it there is.

Standard three: corroboration is broad rather than placed

Measured, Volume I. The top 10 cited domains held only 12% of citations and 56% of cited domains appeared exactly once. **Reasoned.** A fragmented evidence layer means there is no small set of outlets to win. Presence in any one of them is a small contribution, and the plausible route is breadth: the same description of you, as a vendor of a specific thing, existing in many independent places. Chapter 4 at /who-gets-cited/ has the underlying composition, and this standard follows from it rather than from any tested intervention.

Standard four: the description is consistent across sources

Reasoned. If three sources describe a company three different ways, each one contributes to a different possible category attachment and none accumulates. Consistency here is not a branding preference, it is what allows independent mentions to reinforce rather than fragment. This is the least testable standard in the chapter, because nothing in the three volumes measures description consistency, and it is included as inference with that stated plainly.

The size of the group standing at the door

Measured, Volume II. 35.3% of the companies measured were named zero times in their own category, the median challenger scored 2 of 10 with a mean of 2.75, and the median category leader scored 8 of 10 with a mean of 7.73. **Measured, Volume II.** The median leader-minus-challenger gap was 5 points, and in 23.5% of sweeps the gap was 8 points or more. **Reasoned.** The door is where most of this sample is standing. A five-point median gap between the leader and the challenger in the same category is not a matter of relative preference among candidates, it is one company being routinely assembled into sets and another routinely not.

The extreme case, and why it is not about competition

Measured, Volume II. In 16 sweeps the challenger scored zero while the leader scored 8 or above. **Reasoned.** Read alongside the null result on leader strength, those 16 cases are not evidence that the leader displaced the challenger, because across the full sample leader strength does not predict challenger visibility at all. The more consistent reading is that the two companies differ in whether the engine recognises them as members of the category, and that the leader's score is a symptom of the same recognition rather than a cause of the challenger's absence.

Standard five: the evidence is reachable without a form

Reasoned. An engine assembling a candidate set issues plain requests. Material behind a form, a login or a download gate is not readable by the process that decides whether you exist in a category, whatever its value to a human reader who has already found you. This is the same reasoning that keeps this manual ungated, and it is the one standard in the chapter that is a direct consequence of how retrieval works rather than an inference from measured outcomes.

Standard six: being quotable is not enough

Measured, Volume II. Nine companies were named zero times while their own domain was cited as a source, one of them in five of its ten answers. **Reasoned.** These companies cleared every bar related to being findable and useful, and still failed at the door. Whatever the door tests, it is not whether your material can be retrieved and quoted. Chapter 3 at /cited-not-recommended/ covers this group in full, and it is the clearest evidence that content supply and category membership are separate problems.

What the ladder says about sequencing

Measured, Volume II. The absence mix moves monotonically across all four visibility tiers, with category-level absence falling from 55.3% to 48.2% to 20.8% to 12.5% while comparison absence rises from 20.0% to 24.6% to 41.6% to 68.8%. **Reasoned.** That ordering is consistent with two gates cleared in sequence rather than worked on in parallel. It implies that comparison work performed by a company at the door will be read by very few buyers, because the answers where comparisons happen are answers the company is not appearing in. Chapter 10 at /comparison-gate/ is the second gate.

How you know you are at the door

Reasoned. Take the questions you were not named in, sort them into shapes using the published rules in Appendix B at /absence-rules/, then collapse best-of and alternatives-to-incumbent into one category-level bucket. If that bucket dominates, you are at the door. This is the same diagnostic as Chapter 2 at /absence-ladder/ and it requires no tool, only that you kept the text of the questions you lost.

What clearing the door does not get you

Measured, Volume II. Companies at the top tier still lost 16 answers between them, of which 68.8% were head-to-head comparisons. **Reasoned.** The door is a necessary condition and not a sufficient one. A company that becomes a recognised member of its category has changed which questions it loses rather than stopped losing. Any supplier promising that category-presence work produces recommendation is promising something the two-gate structure does not support.

Why the door is cheaper to clear than it looks

Measured, Volume II. The step from the zero tier to the 1 to 3 tier moves category-level absence from 55.3% to 48.2%, while the step from 1 to 3 up to 4 to 6 moves it from 48.2% to 20.8%. **Reasoned.** The largest change in the mix happens in the middle of the range rather than at the very bottom, which is consistent with recognition being something that accumulates and then tips rather than something bought one mention at a time. If that reading is right, the discouraging part is that early effort shows little movement, and the encouraging part is that the movement, when it comes, is not incremental. Both halves of that are inference from a cross-section and neither is a promise.

The honest limit on everything in this chapter

Reasoned. Nothing in the three volumes is an intervention study. No company was measured, changed and measured again, so none of the six standards above has been shown to move a visibility score. They are inferences from a cross-sectional pattern plus one null result, and they are labelled as inference for that reason. A supplier presenting the same standards as proven mechanisms is overstating what the evidence supports, and so would this chapter if it did the same.

What would falsify these standards

Reasoned. A published before-and-after on a fixed question set, with repeat runs and both naming and citation columns reported, in which a company at the door increased third-party category corroboration and did not move. That test is cheap, nobody has published it, and until somebody does the standards in this chapter remain plausible rather than demonstrated. This manual will link to the first credible one that appears, including one that contradicts it.

What this means for your buying decision

Reasoned. Establish which gate you are at before you buy work aimed at either. If category-level questions dominate your absence list, ask a supplier what evidence about you will exist outside your own domain when the engagement ends, and in how many independent places. Treat any proposal that answers mainly in terms of pages published on your own site as aimed at the wrong gate. And discount heavily any claim that your problem is a strong incumbent. Chapter 12 at /how-to-buy-geo/ has the question list.

Where to go next

Reasoned. Chapter 10 at /comparison-gate/ covers the second gate and what content a model can lift a claim from. Chapter 11 at /how-long-it-takes/ is honest about cadence and about how little the volumes can say on elapsed time. Chapter 2 at /absence-ladder/ is the measured basis for the two-gate model.

Sources

The measured figures in this chapter come from The 2026 State of GEO, Volume II: the absence ladder table across all four tiers, the leader null result with its Spearman correlation and Kruskal-Wallis band comparison, the seven sweeps in which no brand scored above 5 of 10, the nine companies cited but never named, and Volume II's own stated practical consequence for companies at the door. The fragmentation figures on the citation layer are Volume I. Everything is at github.com/Broadcastwell/state-of-geo-2026.

Chapter 10. Getting quoted at the comparison gate

By the end of this chapter you will know what the second gate tests, why it is the gate that most visible companies are stuck at, and what properties a comparative claim has to have before a model can use it. This chapter sets standards for the evidence, not tactics for producing it.

The claim this chapter defends

Measured, Volume II. Companies named in seven or more of ten answers lost head-to-head comparisons 68.8% of the time, across 8 companies and 16 absence records. **Measured, Volume II.** Across all 616 absence records, visibility correlates with absence-being-a-comparison at $r = 0.234$, $p = 4.1 \times 10^{-9}$, and the effect survives at $r = 0.184$, $p = 5.9 \times 10^{-6}$ over 600 records when the small top tier is excluded entirely. **Reasoned.** Once a company is a recognised member of its category, nearly everything it still loses is a question that names a specific rival, and that is a different problem from membership.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

What counts as a comparison question

Measured, Volume II. A question was classified as a head-to-head comparison if it contained "vs", "versus" or "comparison", and that rule took precedence over every other shape, so a question containing both "best" and "vs" was counted as a comparison. **Reasoned.** The precedence matters for reading the tier table. Comparison is the greediest category in the scheme, which makes the 68.8% at the top tier a conservative reading of the alternatives-and-shortlist share rather than an inflated reading of the comparison share.

The gate is where visible companies live

Measured, Volume II. Comparison absence rises across the tiers from 20.0% at zero visibility, to 24.6%, to 41.6%, to 68.8%. **Reasoned.** The ordering is monotonic across all four tiers, which is what a sequential process looks like when cut into slices. It also means that success at the first gate delivers you to the second rather than to the end. A company that does category work well should expect its absence mix to shift toward comparisons, and should read that shift as progress rather than as a new failure.

Why the top-tier figure needs its base attached

Measured, Volume II. Nine companies scored 7 or above, but one scored 10 of 10 and produced no absence records at all, so the top tier rests on 8 companies and 16 records. **Reasoned.** A percentage quoted to one decimal place on a base of 16 records is a small base carrying the most quotable number in the ladder, which is a bad combination. The

reason to trust the direction anyway is the correlation that survives dropping the tier entirely, at n = 600. Quote the correlation when you need the finding to hold, and quote 68.8% only with its base beside it.

The comparison surface is the one that always exists

Measured, Volume III. On a separate four-engine collection, Google returned an AI Overview for 47 of 47 comparison questions, a rate of 100%, against 88.8% for best-of questions. **Reasoned.** Comparison questions reliably produce a generated answer. So unlike the category door, where some questions produce no surface at all, the comparison gate is a contest that always takes place. Losing one is losing a contest that happened rather than missing one that did not.

Standard one: a specific named rival appears

Reasoned. A model answering "A versus B" has to find material that addresses A against B. Content that compares a company against unnamed alternatives, or against a generic category, does not supply that. Volume II's own stated practical consequence for companies at this gate is comparison content that a model can quote against a specific named competitor, and the specificity is the operative part. A comparison page that avoids naming anybody is a page that cannot be retrieved for the question being asked.

Standard two: the claim is liftable as a unit

Reasoned. A generated answer is assembled from fragments. A claim that only makes sense after three paragraphs of setup cannot survive extraction, whereas a sentence carrying its own subject, its own object and its own qualifier can. This is the same property that makes a claim checkable by a human reader skimming, and it is why the chapters in this manual are built from short self-contained sections. It is inference about retrieval behaviour rather than a measured result, and it is marked as such.

Standard three: the claim is falsifiable

Reasoned. "Faster" is not liftable in any useful sense, because it has no referent, no measure and no way to be wrong. A claim with a number, a condition and a scope is a claim a model can attribute and a reader can check. The discipline is the same one this manual applies to itself: every figure carries the sample it was computed on. Content that cannot be wrong is content that carries no information, and there is no reason to expect a retrieval system to prefer it.

Standard four: the comparison exists somewhere you do not own

Measured, Volume I. Among the 100 most-cited domains, 77.4% of citations resolved to vendor-authored content, with review platforms at 10.2%. **Reasoned.** Vendor-authored material clearly does get cited, so a vendor's own comparison page is not disqualified. But a comparison published by the party that wins it carries an obvious discount, and the measured presence of review platforms in the evidence layer indicates the engine reaches for third-party comparison material too. Breadth across both is the standard, not a preference for either.

Standard five: being quoted is not the same as winning

Measured, Volume II. 29.4% of companies were cited more often than they were named, and nine were cited as a source while never being named at all. **Reasoned.** A comparison page can be retrieved, quoted and used to support an answer that recommends the other company. Supplying the evidence layer for your own defeat is a real and measured position. The standard is therefore not "be quotable in comparisons" but "be quotable in a way that positions you as the answer", and the two come apart often enough to be worth separating. Chapter 3 at /cited-not-recommended/ has the numbers.

Standard six: the material is reachable without a gate

Reasoned. A comparison behind a form, a login or a download is not available to the process that assembles the answer. This is the least arguable standard in the chapter because it follows from how retrieval works rather than from any inference about ranking. It is also the one most often violated, because comparison material is exactly the kind of asset marketing teams instinctively put behind a lead capture.

The engines disagree most about the tail

Measured, Volume III. Treating engines as raters, agreement across three engines was 0.015 over all vendors and 0.609 when restricted to the study universe of established names. **Reasoned.** Engines converge on the well-known vendors in a category and diverge on everybody else. A challenger fighting a comparison against a named incumbent is fighting on the incumbent's side of that split, where the engines agree, which means the contest is more consistent across engines than category membership is. The practical reading is that comparison outcomes should be more stable engine to engine than category presence, and that a loss on one engine is more likely to be a loss on all of them.

The rival you are compared against is not your choice

Measured, Volume II. For every answer the study recorded which competitor was named most often. **Reasoned.** The pairing in a comparison question comes from the buyer and from whatever the engine considers the natural counterpart, not from the vendor's competitive positioning deck. A company may be routinely compared with a rival it does not consider a competitor, and comparison material aimed at the rival it wishes it were compared with will not be retrieved for the question actually asked. The absence list tells you which pairings are real, which is the only reliable source for that.

Superlatives are the common failure

Reasoned. Most vendor comparison material is built from claims that cannot be false: better, faster, more flexible, more scalable. Those survive legal review precisely because they assert nothing. A model assembling an answer that has to distinguish two vendors gets no discriminating information from them, and neither does a buyer. The standard that follows is uncomfortable for marketing teams and simple to state: if the claim could be made word for word by the competitor, it is not doing the work.

What the middle tier tells you about timing

Measured, Volume II. The 4 to 6 tier sits at 20.8% category-level absence against 41.6% comparison absence, over 21 companies and 101 records. **Reasoned.** By the middle of the range the mix has already flipped. That suggests comparison work becomes the binding constraint well before a company reaches high visibility, rather than only at the top. A company at 4 or 5 of 10 that is still spending exclusively on category presence is probably working on the gate it has largely cleared.

What the ladder does not say about causation

Measured, Volume II. The ladder describes a cross-section of 85 companies at one moment, and Volume II states that nothing in the dataset is an intervention study. **Reasoned.** So the observation that high-visibility companies lose comparisons does not establish that producing comparison content raises visibility. It establishes what the remaining losses look like at each level. Any supplier converting that into a promised mechanism is adding a claim the data does not carry, and this chapter is not making it either.

The variance caveat applies here more than anywhere

Measured, Volume III. Asked the identical question again in the same window, engines agreed with roughly half of their own previous shortlist, at 0.499 and 0.442 mean agreement over 62 and 74 repeat pairs. **Reasoned.** A comparison outcome measured once is a single draw. Losing one comparison answer is very weak evidence about anything, and a supplier reporting movement on a handful of comparison questions between two single-run measurements is reporting noise. Chapter 6 at /measurement-noise/ is the reason to require repeats before believing any of this moved.

How to tell you are at this gate

Reasoned. Classify your absence list using the published rules in Appendix B at /absence-rules/, applying comparison first because it takes precedence. If comparison-shaped questions dominate what you lose, you are at the second gate. If the two buckets are close to even you are in transition, and the sequencing argument in Chapter 9 at /category-door/ says to clear the membership constraint first because it bounds the return on everything else.

One company in this sample cleared both gates

Measured, Volume II. One company scored 10 of 10 and therefore produced no absence records at all. **Reasoned.** That is a single case and it proves nothing about method, but it does establish that the ceiling in this dataset is reachable rather than theoretical. It also means the ladder has an end: there is a state in which there is no absence list left to classify, and a company approaching it should expect its remaining losses to be almost entirely comparisons before they disappear.

What would falsify the standards in this chapter

Reasoned. A published before-and-after on a fixed question set, with repeat runs, in which a company at the comparison gate added checkable, third-party, named-rival comparison material and did not move. Nobody has published that test either way. Until somebody does, these six standards are inferences from a cross-sectional pattern and one stated practical consequence, and they are labelled as inference throughout rather than presented as a mechanism.

What this means for your buying decision

Reasoned. If comparison questions dominate your absence list, ask a supplier which named rivals the work will address and how many of the resulting claims will be checkable rather than adjectival. Ask what will exist off your own domain. Then ask how movement will be measured, and refuse an answer that involves a single run of a handful of questions. Chapter 12 at /how-to-buy-geo/ has the full list of questions to put to a vendor before signing.

Where to go next

Reasoned. Chapter 9 at /category-door/ is the first gate and the one to clear before this one. Chapter 11 at /how-long-it-takes/ is honest about cadence and about how little the evidence says on elapsed time. Chapter 2 at /absence-ladder/ is the measured basis for both gates.

Sources

The measured figures in this chapter come from The 2026 State of GEO, Volume II: the absence ladder across all four tiers, the comparison shape rule and its precedence, the correlation between visibility and comparison-shaped absence with and without the top tier, the composition of the top tier, and Volume II's stated practical consequence for companies at this gate. The trigger rate on comparison questions and the repeat agreement figures are Volume III. The citation composition figure is Volume I. Everything is at github.com/Broadcastwell/state-of-geo-2026.

Chapter 11. How long this actually takes

By the end of this chapter you will know why nobody, including this manual, can currently tell you how long AI search visibility work takes to show an effect, what the measured data does constrain, and what a defensible measurement cadence looks like given that constraint.

The claim this chapter defends

Reasoned. None of the three volumes behind this manual measures elapsed time to effect. No company in any of them was measured, changed and measured again. Every timeline statement in this chapter is therefore inference or external report and is labelled accordingly. The measured figures that do appear below are about measurement precision and about absence shape, never about elapsed time. **Reasoned.** What the data does constrain is different and more useful: it sets a floor on how quickly any change could be detected, and that floor is the binding limit long before biology, budget or effort become relevant.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

Saying plainly what is not known

Reasoned. The honest position is that the elapsed time between changing the evidence about a company and seeing a change in how often engines name it is unmeasured in this research programme and, as far as this manual can establish, unmeasured in any published open dataset. That is not a hedge. It is the state of the field. A supplier quoting a confident number of weeks is quoting a number that has no published basis, and the correct response to it is to ask where the number came from.

Why this chapter exists anyway

Reasoned. A buyer still has to decide how long to run an engagement, how often to measure, and when to conclude that something is or is not working. Refusing to discuss time because it is unmeasured leaves that decision to whoever is willing to invent a figure. What can be done responsibly is to derive the constraints that do follow from measured facts, and to be explicit that the remainder is judgement.

The constraint that dominates everything: detection

Measured, Volume III. Asked the identical question again inside the same collection window, Google AI Overviews agreed with 0.499 of its own previous vendor set across 62 repeat pairs and Claude with 0.442 across 74. **Reasoned.** An instrument that disagrees with itself on roughly half a shortlist cannot resolve a small change. Any real movement has

to exceed that variance before it is visible at all, which means the question "how long until I see something" is answered first by measurement precision and only second by how fast the world changes.

What that does to a two-point comparison

Reasoned. Comparing a single-run measurement in March with a single-run measurement in June compares two draws from two distributions. The difference between them contains the effect of any work done, plus the run-to-run variance of the engine, plus whatever the engine itself changed in between. With agreement between identical repeated runs sitting near half, the variance term is large enough to produce a visible move in either direction with no underlying change whatsoever. Two-point single-run comparisons are therefore not evidence, and no length of engagement fixes that.

The first thing that should move is not the score

Measured, Volume II. Absence shape shifts monotonically across visibility tiers, from 55.3% category-level at zero visibility to 12.5% at seven or above, with comparison absence rising from 20.0% to 68.8%. **Reasoned.** If the two gates are sequential, then a company clearing the first gate should see the composition of its losses change before the headline percentage moves much, because it starts appearing in the answers where competition happens and starts losing those instead. The absence mix is therefore a plausible leading indicator and the visibility percentage a lagging one. This is inference from the cross-sectional pattern and has not been observed longitudinally.

Why that matters commercially

Reasoned. If the mix moves first, then an engagement judged solely on the headline percentage will look like a failure during precisely the period in which it may be working. It also gives a buyer something to ask for that is harder to fake than a score: a shape breakdown of the absence list at each measurement, with the question text retained. A supplier who cannot produce that cannot show a leading indicator even if one exists.

The surface is moving underneath the measurement

Reported. Semrush reports that the share of commercial search results carrying an AI Overview grew 71 percent between November 2025 and April 2026, published 2 July 2026. **Reasoned.** So over a six-month engagement the denominator of any Google-surface score can move substantially for reasons entirely external to the company being measured. A visibility figure that rose over that period may reflect a surface appearing on more questions rather than a company appearing in more answers, and the two are separable only if the answered count is recorded. Chapter 7 at /overview-trigger-rate/ covers the mechanics.

What the literature establishes, and what it does not

Reported. Aggarwal and colleagues showed that content-side interventions move visibility inside a generative engine, in the work that introduced generative engine optimisation as a problem (arXiv:2311.09735, KDD 2024). **Reasoned.** That establishes the surface is

movable. It does not establish how long movement takes in production search products, on commercial buyer questions, at the scale of an ordinary vendor's evidence footprint. Movability and speed are different claims and only the first has support.

The engines change on their own schedule too

Measured, Volume III. The four engines were collected in a single window on one day precisely so that engine drift between windows could not be mistaken for engine divergence, and Volume III states that a refill collected in a second window was deliberately declined for that reason. **Reasoned.** If a study designed around this problem treats a gap of hours as material, then a gap of months between two commercial measurements contains an unknown amount of product change. Some of any movement a buyer is shown over a quarter belongs to the engine rather than to the work, and no measurement design available to a vendor can separate the two.

Three things move at once, and only one is yours

Reasoned. Between any two measurements, the evidence about the company changes, the engine changes, and the measurement itself varies. A single number at each end cannot attribute movement to any of the three. Repeat runs isolate the third. Nothing available to a buyer isolates the second. That is a permanent limitation of measuring a system you do not control, and the honest consequence is that attribution claims in this category should be weaker than they usually are.

What a baseline has to include to be worth having

Reasoned. A baseline taken at the start of an engagement should carry the full question text, the engine set, the run count, the answered count, the naming and citation columns separately, and the absence list with each question's shape. Everything in that list is cheap at the time and impossible to reconstruct later. A baseline that is only a percentage makes every subsequent comparison uninterpretable, which is a common way for an engagement to become unfalsifiable without anyone intending it.

A defensible cadence, offered as judgement

Reasoned. Measure on a fixed question set, with repeat runs at every measurement, on a named engine set, at intervals long enough that the expected effect could plausibly exceed the noise floor. Treat the first measurement as a baseline rather than a result, the second as a check on the instrument rather than on the work, and expect the earliest interpretable read at the third. This is a structure, not a schedule, and the interval that suits a given company depends on how much is changing in its evidence layer.

Why measuring more often is worse, not better

Reasoned. Frequent single-run measurement produces a chart that moves constantly, and constant movement invites interpretation. Every wiggle will be explained by somebody, usually in the direction that suits whoever is explaining. Given a noise floor near half the

shortlist, a weekly chart is close to a random walk with a narrative attached. The discipline is to measure less often, with more runs each time, and to accept a slower answer in exchange for one that means something.

What a promise of ninety days is actually saying

Reasoned. It is saying either that the supplier has measured time to effect and can show you the study, or that they have not. There is no third option. If they have, the study is the most valuable thing they own and they should be delighted to show it. If they have not, the number was chosen because it fits a contract term. Neither answer is disqualifying on its own, but the difference between them is the difference between a measured claim and a sales convention, and you are entitled to know which you are buying.

The compounding argument, and its limits

Reasoned. Chapter 9 at /category-door/ notes that the largest change in absence mix in the measured data sits in the middle of the visibility range rather than at the bottom, which is consistent with recognition accumulating quietly and then tipping. If that reading is right, effort early in an engagement produces little visible movement and the movement, when it arrives, is not proportional to the effort in the month it appears. That is a hypothesis drawn from a cross-section and it is exactly the kind of claim that a longitudinal study would confirm or destroy.

What would settle this

Reasoned. One study would do it: a fixed question set, a named engine set, repeat runs, a cohort of companies measured at regular intervals across a year, with both naming and citation columns published and the raw answers released. It is not technically difficult and it is expensive in patience rather than in engineering. Nobody has published it. Until somebody does, every timeline claim in this category, including every one in this chapter, is inference.

The one timing statement this manual will make

Reasoned. Do not judge an engagement on a single-run measurement taken less than two measurement cycles after it began, because before that point the instrument cannot distinguish the work from its own variance. That is a statement about measurement, not about how quickly evidence propagates, and it is the only timing claim in this manual that follows from something measured rather than from judgement about the world.

The asymmetry that should make you sceptical of fast claims

Reasoned. A supplier benefits from a short promised timeline at the point of sale and from a long one at the point of review. That asymmetry means quoted timelines drift toward whatever the current conversation rewards, unless they are anchored to something published. Asking for the anchor is not adversarial. It is the only way to tell a considered estimate from a number that adjusts itself to the room.

What this chapter does not claim

Reasoned. It does not claim that AI search visibility work is slow, or fast. It does not claim that any particular cadence is optimal. It does not claim that the leading-indicator reading of the absence mix is correct, only that it is consistent with the measured cross-section and testable. And it does not claim that the absence of a published longitudinal study means the work does not function, which would be an argument from ignorance in the opposite direction.

What this means for your buying decision

Reasoned. Ask a supplier for the study behind any timeline they quote, and accept "we have not measured it" as a good answer if it comes with a measurement plan. Require repeat runs at every measurement and a fixed question set from the first day, because retrofitting either destroys the baseline. Ask for the absence shape breakdown at each measurement, not just the percentage. And treat any contract that promises a specific score by a specific date as a promise about an instrument nobody has calibrated. Chapter 12 at /how-to-buy-geo/ has the full list.

Where to go next

Reasoned. Chapter 6 at /measurement-noise/ is the measured basis for the detection constraint that dominates this chapter. Chapter 8 at /valid-measurement/ sets out what a measurement has to carry before a cadence built on it means anything. Chapter 12 at /how-to-buy-geo/ turns all of it into a purchase decision.

Sources

The measured figures cited in this chapter are the within-engine repeat agreement values from The 2026 State of GEO, Volume III, and the absence shape distribution across tiers from Volume II. Neither volume measures elapsed time to effect, and no figure in this chapter is presented as doing so. External work is attributed to Semrush and to Aggarwal and colleagues with publication dates and identifiers, reproduced from Volume III's reference list. Everything is at github.com/Broadcastwell/state-of-geo-2026.

Chapter 12. How to buy GEO without getting sold a number

By the end of this chapter you will have twelve questions to put to any supplier, a description of what a defensible deliverable contains, a way to read a proposal, and a clear account of the conditions under which the right decision is to hire nobody.

The claim this chapter defends

Reasoned. Almost every failure mode in buying AI search visibility work reduces to accepting a number without its denominator. **Measured, Volume III.** The measured facts that make that failure expensive are specific: engines agree with roughly half their own previous shortlist on a repeated identical question, 14 of 32 companies were visible on some engines and invisible on others, and one engine declined to answer 22 of 280 questions. **Reasoned.** Each of those makes a confident single number wrong in a different direction, and each has a question a buyer can ask that exposes it.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

The twelve questions

Reasoned. Put these in writing before signing anything. Every one is answerable in a sentence by a supplier doing the work properly, and the pattern of which ones cause difficulty is more informative than any individual answer.

1. Which engines does the score cover, and is that stated on the front page of the report rather than in a footnote?
2. On what date, and in what collection window?
3. What exactly are the questions, in full text, and will they be identical next quarter?
4. How many runs did each question get, and what was the run-to-run variance?
5. How many questions produced no AI answer at all?
6. Is the headline percentage computed over questions asked or over answers returned?
7. Are naming and citation reported as separate columns?
8. What is the absence list, with each question's shape?
9. What was excluded from the denominators, on what rule, and are the excluded rows still in the delivered data?
10. Will the raw answers be delivered, so the number can be recomputed without you?
11. What is the scoring rule for deciding a brand was named, and is it deterministic?
12. Do you sell the remedy for the problem you are measuring, and is that disclosed in the report?

Why question four is the sharpest

Measured, Volume III. Within-engine agreement on repeated identical questions was 0.499 for Google AI Overviews over 62 repeat pairs and 0.442 for Claude over 74. **Reasoned.** A supplier who runs each question once has sold you a point estimate from a distribution and has not told you the width. That is not a minor methodological preference, it means the difference between two of their reports may be entirely instrument. If the answer to question four is that every question was run once and the variance is unknown, the rest of the report cannot be read as movement.

Why question twelve is not rhetorical

Reasoned. Nearly every supplier of AI visibility measurement also sells the remedy, and the conflict is structural rather than a matter of anybody's integrity. The relevant question is not whether the conflict exists but whether it is disclosed and whether the data is open enough for a sceptic to recompute the result. A supplier who declines to state the conflict on the report has chosen the presentation that suits them over the one that lets you check them, which tells you something about the rest of the report.

What a defensible deliverable contains

Reasoned. Six things, all cheap to produce at collection time and impossible to reconstruct afterwards. The question set in full text. The raw answers, one row per engine per question per run. Three counts per question set: asked, answered, and answers naming you. Naming and citation as separate columns. The absence list with each question's assigned shape. And the run-to-run variance on a repeat subsample. Anything beyond that is presentation. Anything less than that is a claim you cannot check.

How to read a proposal

Reasoned. Read it backwards, starting from what will be delivered rather than from what will be done. If the deliverable is a dashboard, ask what is behind it and whether you get the rows. If the proposal describes activity in volume terms, count how many of the twelve questions it answers without prompting. Then look for the word "AI search" used as though it were one place, because a proposal that has not distinguished the engines has not engaged with the thing it is proposing to move. Chapter 5 at /engine-divergence/ is why that distinction matters.

The diagnosis has to come before the prescription

Measured, Volume II. Companies named in zero of ten answers lose category-level questions 55.3% of the time, while companies named seven or more times lose head-to-head comparisons 68.8% of the time. **Reasoned.** Those two conditions need different work, and the difference is in kind rather than in degree. A proposal that prescribes the same programme regardless of which gate you are at has not diagnosed you. Ask which gate your absence data says you are at, and ask to see the classified absence list that supports the answer. Chapter 2 at /absence-ladder/ is the model.

What a good answer to "what will you do" sounds like

Reasoned. It names the gate, it says what evidence about you will exist at the end of the engagement that does not exist now, and it says where that evidence will live, specifically including how much of it will be somewhere you do not own. It does not promise a score. Chapter 9 at /category-door/ and Chapter 10 at /comparison-gate/ set out the standards the evidence has to meet at each gate, and a proposal can be read directly against them.

What a bad answer sounds like

Reasoned. A promised percentage by a promised date. A guarantee of position, which does not exist in a system that selects rather than ranks. A plan expressed entirely as content volume on your own domain, which addresses neither gate directly. A blended cross-engine score with no per-engine breakdown. And any use of the phrase "AI SEO" as though the transfer of practice from search engine optimisation were settled rather than the open question Chapter 0 at /selection-not-ranking/ treats it as.

On timelines in a proposal

Reasoned. Ask for the study behind any quoted timeline. As Chapter 11 at /how-long-it-takes/ sets out, no published open dataset measures elapsed time to effect in this category, so a specific promise is either backed by private evidence the supplier should be eager to show, or it is a contract convention. "We have not measured it, here is our measurement plan" is a better answer than a confident number, and a supplier who gives it is telling you something reliable about how they treat evidence generally.

How to price the work without a benchmark

Reasoned. There is no published benchmark for what this work should cost, so any figure quoted as market rate is an assertion. What a buyer can do instead is ask which parts of a proposal are fixed and which scale, and with what. Measurement scales with questions, engines and runs, and those three are declared. Evidence work scales with how many independent places the material has to reach. A proposal that cannot be decomposed that way is priced on the outcome it is implying rather than on the work it contains, which is the shape that makes overruns invisible.

The renewal conversation is where the criteria pay off

Reasoned. Everything in this chapter is easy to require before signing and nearly impossible to retrofit. A fixed question set from day one makes the second measurement comparable. Repeat runs from day one make a change distinguishable from noise. Raw rows from day one mean the renewal decision can be made on evidence rather than on a narrative about a chart. Buyers who skip these at the start almost always discover them at renewal, at the point where the information would have been worth most.

What to do when the answers are good but the score is flat

Measured, Volume II. Absence shape shifts across visibility tiers before the tiers themselves change, moving from 55.3% category-level at zero visibility to 20.8% in the middle band. **Reasoned.** If a supplier meets every criterion in this chapter and the headline

percentage has not moved, the next thing to look at is whether the composition of the absence list has changed, since the two-gate model predicts that shifts before the score does. That is a genuine possibility rather than an excuse, and the way to tell the difference is that a supplier offering it should have been reporting the shape breakdown from the first measurement rather than introducing it when the number disappoints.

On tools, as distinct from suppliers

Reasoned. A tool is buying you collection and presentation, not judgement, so the twelve questions apply to it unchanged and the answers are usually easier to establish because the methodology is documented or it is not. The specific things to check are whether the question set is yours and fixed, whether repeat runs are available at all rather than as a premium tier, whether the export contains raw answers, and whether the engine list is stated. A tool that cannot export the rows is a tool whose numbers you cannot audit.

The case for hiring nobody

Reasoned. The diagnostic in this manual requires no supplier. Write the ten questions a real buyer in your category would ask, put each to an engine several times, record whether you were named and whether your own domain was cited, and keep the text of the questions you lost. That gives you a baseline, a gate diagnosis and a variance estimate for the cost of an afternoon. Doing that first is strictly better than buying it, because it tells you what you are buying before you buy it.

When hiring nobody is the right permanent answer

Reasoned. If you are at the category door and have no route to third-party corroboration, no supplier can manufacture recognition you have no basis for. If your category is one an engine does not model cleanly, the work may be category definition rather than visibility work, which is a different discipline. And if you cannot commit to a fixed question set and repeated measurement, you will not be able to tell whether anything worked, in which case the money buys activity and not information. None of those is a reason to buy carefully. They are reasons not to buy.

What this manual's own self-audit shows about the category

Measured, self-audit of 18 August 2026. On a re-measure of ten published questions across four engines at one run per question, this firm was named in 1 of 40 answers and cited once among 674 citations, and the naming and citation were the same answer, in which an engine quoted its own research page while recommending other agencies. **Reasoned.** That is published because a supplier asking you to demand evidence should be measurable on the same terms. It is also a live example of Chapter 3 at /cited-not-recommended/, and anybody quoting the 1 of 40 as a visibility result would be making the error this manual is about.

What this chapter does not claim

Reasoned. It does not claim that a supplier meeting all twelve criteria will produce a commercial result, because no dataset here links measurement quality to outcomes. It does not claim that suppliers failing some of them are acting in bad faith, since most of the practices criticised here are industry conventions rather than deceptions. And it does not claim that these are the only criteria that matter, only that a number failing them cannot be interpreted whatever else is true of it.

What this means for your buying decision

Reasoned. Run the diagnostic yourself before you take a meeting, so you arrive knowing which gate you are at. Put the twelve questions in writing and treat the pattern of difficulty as the signal. Require the six deliverable items in the contract rather than the report. And be prepared for the answer to be that you should do nothing yet, which is a legitimate outcome of an honest evaluation and one that no supplier will reach on your behalf.

Where to go next

Reasoned. Chapter 8 at /valid-measurement/ is the technical basis for most of the twelve questions. Chapter 6 at /measurement-noise/ is the evidence behind the repeat-run requirement. Appendix E at /references/ carries the full reference list and the complete self-audit disclosure.

Sources

The measured figures in this chapter come from The 2026 State of GEO, Volume II for the absence ladder tiers, and Volume III for the within-engine repeat agreement, the company-level 14 of 32 result and the trigger rate. The self-audit figures are the operator's own and are published with their dates and run structures in the block at the foot of every page of this manual. The criteria themselves are inference from those measurements and are labelled as such. Data and code are at github.com/Broadcastwell/state-of-geo-2026.

Appendix A. The ten-question buyer bank

Reference material for Chapter 8 at /valid-measurement/. This appendix reproduces the question bank design at the level the source volumes publish it, and no deeper.

What a question bank is for

Reasoned. A visibility measurement is a measurement of performance on a set of questions. The bank therefore defines what "visibility" means for that measurement, and two banks over the same company on the same engine in the same window will produce different scores. That makes the bank the single most consequential design decision in the instrument, and the one most often left undeclared.

Rule one: questions are phrased the way a buyer types them

Measured, Volume I. The Volume I bank used ten standardised buyer questions per category, a mix of best-of, direct comparison and use-case questions, phrased the way a buyer types them. **Reasoned.** The alternative, questions containing the subject's brand name, measures whether the engine knows the company exists rather than whether it competes. A bank of brand-name questions will produce a flattering score that has no relationship to whether a buyer researching the category encounters the vendor.

Rule two: the bank is fixed before collection and never regenerated

Measured, Volume III. Of 280 questions, 276 came from the study's existing question bank and 4 from the published Volume I dataset. No question was regenerated. **Reasoned.** Regeneration between measurements introduces a degree of freedom that can be tuned, deliberately or otherwise, toward questions the subject performs well on. Fixing the bank removes it, and publishing the bank makes the fixing checkable.

Rule three: every category covers the shapes that matter

Measured, Volume III. Each category was required to carry at least one best-of, one alternatives, one comparison, one use-case and one pricing or evaluation question. 33 of 40 categories met that in full. The 7 that did not are listed explicitly in the published sample manifest rather than quietly padded, because the question bank for those categories did not contain the missing shape.

Rule four: the realised mix is reported, not just the intended one

Measured, Volume III. The realised shape mix across the 280 questions was 80 alternatives, 80 best-of, 47 comparison, 38 evaluation, 33 use-case and 2 other. **Reasoned.** Intended coverage and realised coverage differ whenever a bank is drawn from a finite pool. Reporting both is what allows a reader to judge whether a score is weighted toward shapes that are easy or hard for the subject, and Chapter 7 at /overview-trigger-rate/ shows that the mix also moves the trigger rate and therefore the denominator.

The six shapes

Measured, Volume II. The classification scheme, its rules and its precedence order are reproduced in full in Appendix B at /absence-rules/. In summary the shapes are head-to-head comparison, alternatives-to-incumbent, best-of shortlist, evaluation criteria, use case or problem, and definitional.

A ten-question template, offered as construction

Reasoned. The volumes publish the shapes and the coverage requirement but do not publish a fixed ten-question allocation, so what follows is an explicit construction rather than a reproduction. A ten-question bank that satisfies the published coverage requirement and reflects the realised mix would carry roughly three best-of or shortlist questions, three alternatives-to-incumbent questions, two head-to-head comparisons naming real rivals, one evaluation or pricing question, and one use-case question. Adjust the weighting to your category, declare the weighting you used, and then never change it.

Slot	Shape	What it tests
1 to 3	Best-of shortlist	Whether you are assembled into the candidate set at all
4 to 6	Alternatives-to-incumbent	Whether you are reachable from the incumbent's name
7 to 8	Head-to-head comparison	Whether you win preference against a named rival
9	Evaluation or pricing	Whether you appear when the buyer is comparing on criteria
10	Use case or problem	Whether you appear when the buyer describes a problem rather than a category

Rule five: the questions name real rivals, not aspirational ones

Measured, Volume II. For every answer the study recorded which competitor was named most often. **Reasoned.** The rivals a buyer pairs you with are an observable fact rather than a positioning choice, and comparison questions built around a rival nobody pairs you with will measure nothing. Build the comparison slots from the pairings that appear in your own absence list, which Chapter 10 at /comparison-gate/ discusses.

Rule six: the bank is published with the score

Reasoned. Without the question text, a reader cannot tell whether a high score reflects broad presence or a bank weighted toward definitional questions where almost everyone appears. Publication is what makes a score checkable by somebody who did not run it, and it costs nothing.

The limitation the source volumes state about their own bank

Measured, Volume II. The ten questions per category were generated to reflect realistic buyer research rather than sampled from real search logs. Volume II states that a different question set would produce a different absence mix, and that the shapes are drawn from

that set. **Reasoned.** So the absolute proportions in any published result are set-dependent, while the tier comparison is internally consistent because every tier faced the same bank. Read the ordering as the finding and the decimals as properties of the instrument.

What is deliberately not in this appendix

Reasoned. This appendix publishes the design rules and the shape taxonomy at the level the source volumes already publish them. It does not describe how any commercial measurement is operated, scheduled or delivered, and a reader looking for that will not find it in this manual.

Sources

The bank design, provenance and coverage requirement, the 33 of 40 disclosure and the realised shape mix are from The 2026 State of GEO, Volume III. The ten-questions-per-category design and its phrasing rule are Volume I. The shape taxonomy and the limitation on question generation are Volume II. The ten-slot template is an explicit construction by this manual and is labelled as such. Everything is at github.com/Broadcastwell/state-of-geo-2026.

Appendix B. Absence classification rules and precedence

Reference material for Chapter 2 at /absence-ladder/, Chapter 9 at /category-door/ and Chapter 10 at /comparison-gate/. Reproduced as published in Volume II.

What is being classified

Measured, Volume II. For every question where the company was not named, the full question text was retained. That produced 616 absence records across 85 companies in 60 categories. Each record was then assigned exactly one shape.

How completeness was validated

Measured, Volume II. For all 85 companies, the absence count equals ten minus the visibility score, with zero mismatches. **Reasoned.** That check is what makes the denominator of every percentage in the ladder the complete set of losses rather than a convenience subset, and it is the single cheapest integrity check available on an absence dataset.

The six shapes and their rules

Measured, Volume II. Each absence question was assigned one shape by regular-expression rules applied in fixed precedence.

Shape	Rule	Example
Head-to-head comparison	contains "vs", "versus", or "comparison"	"Whitespace vs Artificial Labs: which is better?"
Alternatives-to-incumbent	contains "alternative"	"What are the best Sequel alternatives?"
Best-of shortlist	contains "best" or "top N"	"What is the best resource management software?"
Evaluation criteria	evaluate, choose, select, criteria, pricing model	"How should I compare pricing models when switching?"
Use case or problem	opens "how can/do/does" or contains "use case"	"How can this software reduce manual data entry?"
Definitional	opens "what is/are" without "best"	"What is reinsurance placement software?"

The precedence order

Measured, Volume II. The rules are applied in the order shown, top to bottom. A question containing both "best" and "vs" is counted as a comparison, not a shortlist. **Reasoned.** Precedence is what makes the scheme deterministic. Without a stated order, the same question can land in two shapes depending on implementation, and two people classifying the same absence list will not agree.

The residual category

Measured, Volume II. 2.8% of records fell to *other*, meaning they matched none of the six rules. **Reasoned.** A published residual is a useful signal about a taxonomy. A scheme with no residual has usually forced every case into a bucket, and a scheme with a large one is not describing its material. 2.8% is small enough to be ignorable and non-zero, which is what an honest rule set looks like.

The collapsed buckets used in the ladder

Measured, Volume II. Category-level means best-of shortlists plus alternatives-to-incumbent queries, combined. Head-to-head comparison is reported on its own. **Reasoned.** The collapse is what turns six shapes into a two-gate diagnostic. Best-of and alternatives-to-incumbent both ask who belongs in a set without adjudicating between two named vendors, which is why they group. Comparison is separated because it tests preference rather than membership.

The visibility tiers

Measured, Volume II. Records were grouped by the subject company's visibility score into four tiers: named 0 of 10, named 1 to 3, named 4 to 6, and named 7 to 10, carrying 300, 199, 101 and 16 absence records respectively across 30, 25, 21 and 8 companies. **Measured, Volume II.** Nine companies scored 7 or above but one scored 10 of 10 and produced no absence records, so the top tier rests on 8 companies, and company counts sum to 84 in the tier table against 85 in the sample.

The stated limitation of the scheme

Measured, Volume II. Question shapes were assigned by regular expression, not human annotation, and Volume II publishes the rules and their precedence so anyone can reclassify and check. **Reasoned.** A human annotator would classify some questions differently, most plausibly at the boundary between the evaluation-criteria and use-case rules, which both key on how a question opens. **Reasoned.** The trade is accuracy for reproducibility, and it is the right trade for a published dataset because the rules and the 616 classified questions are both published. Anyone who disagrees with a rule can reclassify the whole set and recompute the table rather than argue about individual calls.

How to apply this to your own absence list

Reasoned. Take the questions you were not named in, keep the full text, and apply the six rules in the order given: comparison first, then alternatives, then best-of, then evaluation, then use case, then definitional. Anything matching none of them is other. Then collapse best-of and alternatives into a single category-level bucket and keep comparison separate. Those two counts are the diagnostic described in Chapter 2 at /absence-ladder/.

Sources

The classification scheme, its precedence, the worked examples, the residual share, the tier definitions and their record counts, the completeness validation and the stated limitation are all from The 2026 State of GEO, Volume II. The 616 classified questions are published

with their verbatim text, assigned shape and visibility tier in `absence_questions_classified.csv`, and the aggregate table in `absence_shape_by_tier.csv`, at github.com/Broadcastwell/state-of-geo-2026.

Appendix C. Scoring and matching specification

Reference material for Chapter 8 at /valid-measurement/. Reproduced at exactly the level Volume III publishes it and no deeper.

Why the rules matter more than they look

Reasoned. A visibility score is the output of a decision procedure applied to answer text. If that procedure is a judgement call, the score depends on who ran it. If it is a stated rule, two independent parties get the same number from the same answers. Everything in this appendix exists to make the second true.

Two layers, with a strict ordering

Measured, Volume III. Layer 1 is deterministic and authoritative for every company in the study universe. Layer 2 is a language-model extractor that exists to catch vendors outside that universe, and it may only add out-of-universe names. Any Layer 2 claim about a universe company is discarded unless Layer 1 also finds the string in the answer.

Reasoned. That ordering is the safeguard. It means the model-based layer can widen coverage but cannot inflate or deflate a measured company's numbers in either direction, so every universe-only statistic in the study is independent of it.

Text normalisation, applied before any matching

Measured, Volume III. Answer text is normalised for non-breaking and thin spaces, curly quotes and the dash family, then whitespace-collapsed. **Reasoned.** Without normalisation, a brand rendered with a typographic apostrophe fails to match the same brand rendered with a straight one, and the failure is silent. Normalisation is unglamorous and it is where a large share of avoidable scoring errors live.

Brand matching

Measured, Volume III. Brand matching is word-boundary. It is case-sensitive when the brand is a single plain alphabetic word, so a company called Pitch does not match the ordinary verb. A brand carrying a dot, digit, hyphen or space matches case-insensitively.

Reasoned. The asymmetry is deliberate. Single common words carry a high false-positive risk and need the extra constraint of case. Names containing punctuation or digits are unlikely to appear by accident, so the constraint can be relaxed without cost. Volume III states these are Volume I's rules, unchanged, which is what makes the two volumes comparable.

Domain matching

Measured, Volume III. Domain matching is at hostname level, with multi-domain values split on pipes and commas, so subdomains count and near-misses do not. **Reasoned.** Hostname level means a citation of a subdomain counts as a citation of the company, which

is correct because companies publish across subdomains. It also means a domain that merely contains the company name as a substring does not count, which prevents a class of false positive that would otherwise inflate citation counts.

Citation classification

Measured, Volume III. Every cited URL is reduced to a hostname and classified as vendor-owned, review platform, editorial, analyst, community or other, reusing Volume I's classification so the vendor-authored figure from that volume is comparable per engine. **Measured, Volume I.** In Volume I the same scheme carried the additional labels encyclopaedia and social, and the 100 most-cited domains were classified by hand.

Reasoned. Reusing a classification across volumes is what allows a like-for-like comparison. Changing a taxonomy between studies while keeping the same headline metric is a common and quiet way to make two incomparable numbers look like a trend.

The two scored outcomes

Measured, Volume I. Each answer was scored on two independent binary outcomes: whether the company's brand was named, and whether its own domain was cited as a source. Every cited URL was logged. **Reasoned.** Keeping these separate is what makes Chapter 3 at /cited-not-recommended/ possible. A measurement that blends them into one visibility figure cannot recover the distinction later.

Completeness and exclusion rules

Measured, Volume III. The primary sample is the set of questions where every engine in the relevant set returned a usable answer, and pairwise metrics additionally use every question both engines in that pair answered, which is why pair sample sizes differ and are reported. Attrition is reported per engine rather than absorbed.

Measured, Volume III. A Google query that returns no AI Overview is recorded with its own status rather than as an error, because Google declining to generate is a result about the engine. Rows carrying a billing error are excluded from every denominator and retained in the raw data with an error status, so the exclusion is checkable.

Repeat handling

Measured, Volume III. A repeat pair whose two runs used different model strings measures a model change rather than run-to-run noise and is excluded by construction. Volume III reports that no pair triggered that rule. **Measured, Volume III.** An engine needs at least 12 repeat pairs before a verdict is asserted for it.

Reproducibility

Measured, Volume III. The bootstrap uses 2000 replicates with a fixed seed, so the published confidence intervals reproduce exactly, and figures are generated from the computed results file rather than typed, so a caption cannot drift from the dataset. **Reasoned.** A fixed seed is the difference between an interval a reader can verify and one they have to trust.

What is deliberately not in this appendix

Reasoned. This reproduces the scoring and matching specification at the level the source volumes publish it. It does not describe how any commercial measurement is operated, scheduled or delivered, and nothing here constitutes an operable procedure for running a sweep.

Sources

The two-layer design, normalisation, brand and domain matching rules, citation classification, completeness and exclusion rules, repeat handling and bootstrap settings are from The 2026 State of GEO, Volume III, sections 3.4 to 3.6 and 7. The two scored outcomes and the hand classification of the top 100 domains are Volume I. The extractor prompt and the analysis scripts are published in full at github.com/Broadcastwell/state-of-geo-2026 under `volume-iii/analysis/`.

Appendix D. Glossary

One definition per term. Where a term comes from a source volume the volume is named in the entry, which stands in place of an evidence label: a definition is a statement about usage rather than a claim about the world, so the three-level scheme used everywhere else in this manual does not apply here. Where a term is this manual's own coinage that is stated explicitly.

The evidence levels

Measured. A claim taken from one of the three State of GEO volumes, with the volume named, recomputable from the published data.

Reported. A claim taken from a source outside this research programme, cited with its publisher or authors and a date or identifier.

Reasoned. An inference drawn from measured or reported material. Argument rather than measurement, marked so it can be disputed without disputing the data. This three-level scheme is this manual's own convention.

The unit of analysis

Answer. One response from one engine to one question on one run. The unit everything else is counted over.

Named. Volume I and II. A binary outcome recording that the subject company's brand string appeared in the answer text under the matching rules in Appendix C at /scoring-spec/.

Cited. Volume I and II. A separate binary outcome recording that the subject company's own domain appeared among the sources the engine cited for that answer.

Absence record. Volume II. One question on which the subject company was not named, retained with its full question text.

Absence list. This manual's term for the set of a company's own absence records.

The model

Selection, not ranking. This manual's framing, Chapter 0 at /selection-not-ranking/. The observation that an AI answer is assembled by choosing a small number of vendors from a candidate set rather than by ordering a full list.

Candidate set. The group of vendors an engine assembles before writing an answer. Not directly observable from outside; inferred from which vendors appear.

Category door. This manual's term for the first gate: whether an engine treats a company as a member of its category at all.

Comparison gate. This manual's term for the second gate: whether an engine prefers a company over a specific named rival.

Absence ladder. Volume II. The finding that the mix of question shapes a company loses shifts systematically as its visibility rises.

Visibility tier. Volume II. One of four bands grouping companies by score: named 0 of 10, 1 to 3, 4 to 6, and 7 to 10.

Question vocabulary

Question bank. The fixed set of buyer questions a measurement is computed over. See Appendix A at /question-bank/.

Question shape. Volume II. One of six categories a question is assigned to by rule: head-to-head comparison, alternatives-to-incumbent, best-of shortlist, evaluation criteria, use case or problem, definitional.

Precedence. Volume II. The fixed order in which shape rules are applied, so a question matching two rules lands deterministically in one shape.

Category-level. Volume II. Best-of shortlist plus alternatives-to-incumbent, collapsed into one bucket because both ask who belongs in a set rather than adjudicating between two named vendors.

Agreement and variance

Jaccard similarity. Volume III. The size of the intersection of two vendor sets divided by the size of their union. 1.0 means identical shortlists.

Overlap coefficient. Volume III. An agreement measure normalised by the smaller of the two sets, used as a length control because it cannot be depressed by one side simply being longer.

Repeat pair. Volume III. Two runs of the same question on the same engine, compared with each other.

Within-engine agreement. Volume III. Mean agreement across repeat pairs on one engine. The engine's agreement with itself.

Between-engine agreement. Volume III. Mean agreement between one engine and the others on the same questions.

Noise floor. This manual's term for an engine's within-engine agreement, used as the threshold any claimed difference has to exceed to be interpretable.

Fleiss kappa. Volume III. A chance-corrected agreement statistic treating engines as raters making a binary mention decision.

Bootstrap. Volume III. Resampling used to produce confidence intervals, run at 2000 replicates with a fixed seed so the intervals reproduce exactly.

Denominator vocabulary

Trigger rate. Volume III. The share of questions on which an engine returned a generated answer at all.

No AI Overview. Volume III. A recorded status meaning Google returned no AI Overview for that question, kept distinct from an error because it is a result about the engine.

Answer rate. Volume III. Answers returned divided by questions attempted, per engine.

Unstated denominator. This manual's term for a reported percentage whose base has not been declared.

Scoring vocabulary

Study universe. Volume III. The set of companies, domains, competitors and aliases the deterministic matching layer knows about.

Layer 1. Volume III. The deterministic matching layer, authoritative for every company in the study universe.

Layer 2. Volume III. A language-model extraction layer that may only add vendors from outside the study universe and can never override Layer 1 about a universe company.

Vendor-authored. Volume I. A citation classification for a source published on a software vendor's own domain.

Evidence layer. This manual's term for the set of sources engines cite when composing answers about a category.

Sample vocabulary

Challenger-skewed. Volume I and II. A sample selected as plausible non-leaders in categories with an identifiable incumbent, which lowers average named rates relative to a random draw of all vendors.

Category leader. Volume II. The brand scoring highest in a category's sweep. Leaders enter the data through mentions rather than through selection.

Winner-take-most. Volume I. The observed shape in which a small group of names is selected repeatedly while a large group is selected almost never.

Collection window. The period during which answers were gathered. Part of the value of any figure, not metadata on it.

Disclosure vocabulary

Operator disclosure. The statement, carried on every page of this manual, that the firm running the measurement sells services in the category it measures, is excluded from the sample and from every ranking, and publishes the raw data and code so the result can be recomputed.

Self-audit. A measurement the operator runs against itself on its own published question set, reported with its date and run structure. Both self-audit figures appear in the block at the foot of every page and in Appendix E at /references/.

Concept DOI. A Zenodo identifier covering all versions of a record, which resolves to the latest. Not a version and not cited in place of one.

Version DOI. A Zenodo identifier for one specific published version. The three volumes are cited by version DOI throughout.

Sources

Terms attributed to a volume are used in the sense that volume defines. Terms marked as this manual's own are conventions adopted here for clarity and carry no external authority. The underlying data and code are at github.com/Broadcastwell/state-of-geo-2026.

Appendix E. References and self-audit disclosure

The three source volumes

Volume	Title	Version DOI
Volume I	The 2026 State of Generative Engine Optimization, v1.0	10.5281/zenodo.21537014
Volume II	The Absence Ladder. The 2026 State of GEO, v2.0	10.5281/zenodo.21586091
Volume III	Divergence Survives a Length Control on Google AI Overviews and Vanishes on Claude. The 2026 State of GEO, v3.0	10.5281/zenodo.21789120

Measured. All three are published under CC BY 4.0 with their data and analysis code at github.com/Broadcastwell/state-of-geo-2026. Zenodo also issues a concept identifier covering all versions of the series, which resolves to the latest version. It is not a volume and it is cited nowhere in this manual.

Cite Volume III as: Sivakumar, S. (2026). Divergence Survives a Length Control on Google AI Overviews and Vanishes on Claude. Measuring Vendor-Set Agreement Across Four AI Search Engines. The 2026 State of Generative Engine Optimization, v3.0. Zenodo.

External work cited in this manual

Reported. The entries below are reproduced from Volume III's reference list, which states that every reference was checked against the arXiv listing or the publisher page before publication, and that author lists are as printed by the source. This manual has not independently re-verified them.

Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., Deshpande, A. (2024). GEO: Generative Engine Optimization. KDD 2024. arXiv:2311.09735.

Bagga, P. S., Farias, V. F., Korkotashvili, T., Peng, T., Wu, Y. (2025). E-GEO: A Testbed for Generative Engine Optimization in E-Commerce. arXiv:2511.20867.

BrightEdge (2026). Why AI Engines Cite Different Sources but Recommend the Same Brands. Weekly AI and Search Insights, 24 April 2026.

Chu, X., Hou, Y. (2026). Incumbent Advantage: Brand Bias and Cognitive Manipulation Dynamics in LLM Recommendation Systems. arXiv:2606.17443.

DerivateX (2026). ChatGPT and Google AI Overviews agree on tools, not sources. Open benchmark dataset, June 2026.

Harsel, L., Yudina, A., Skopec, C. (2026). AI Overviews are expanding across commercial intent search. Semrush, 2 July 2026.

Jack, W., Lehman, N., Maloney, K., Xu, S. (2026). Prominence-Stratified Failure Modes in Retrieval-Augmented Commercial Recommendation: A 37,000-Run Audit. arXiv: 2605.27439.

Jack, W., Lehman, N., Maloney, K., Xu, S. (2026). Divergent Recommendations, Convergent Diagnoses: Cross-Provider Failure-Mode Convergence in AI Commercial Recommendation. arXiv:2606.26116.

Khromova, Y. (2025, updated 2026). ChatGPT vs Perplexity vs Google vs Bing: AI Search Engine Comparison. SE Ranking. Data collected 26 February to 3 March 2025.

Lafferty, N. (2025). AI Platform Citation Patterns. Profound.

McDonald, T., Cooley, H., Williams, M. (2026). AIO Impact on Google CTR: 2026 Update. Seer Interactive, 24 April 2026. Data through February 2026.

Schulte, J., Bleeker, M., Kaufmann, P. (2026). Don't Measure Once: Measuring Visibility in AI Search (GEO). arXiv:2604.07585.

Zatuchin, D. (2026). Who Owns the AI Recommendation? arXiv:2606.23057. Dataset released on Zenodo under CC BY 4.0.

Reported. Volume III notes of Jack and colleagues (arXiv:2606.26116) that the same-prompt rerun baseline quoted in that paper is cited by it from a separate industry report rather than measured within it. This manual repeats that qualification wherever the work is used.

Open flags

Reasoned. `FLAGS.md` at the root of this manual's repository records six open items where a published source is incomplete or two published sources disagree. They are disclosures rather than defects. Where a chapter touches one, it quotes the source as published, asserts no relation between conflicting figures, and does not attempt reconciliation. The current items concern Volume III's answer counts, Volume I's answer and question counts, a duration statement in Volume III, the location of the unique-domain figure, the date on the published self-audit, and the run structure of the August re-measure.

Self-audit disclosure, in full

Measured, self-audit of July 2026. In the four-engine, five-run self-audit published alongside Volume II in July 2026, Broadcastwell was named in 0 of 200 answers and cited 0 times among 663 citations. The structure was four engines, ten questions, five runs per question. That published figure stands with its date and is never replaced. The July date is taken from the published citation metadata rather than from the disclosure sentence itself, which gives no month, and that is recorded as an open flag.

Measured, self-audit of 18 August 2026. Re-measured on the same ten published questions across the same four engines, at one run per question rather than five, between 03:51 and 04:03 UTC on 18 August 2026: Broadcastwell was named in 1 of 40 answers and cited once among 674 citations. Forty of forty calls returned a scored answer and no engine failed.

Measured, self-audit of 18 August 2026. The single naming and the single citation are the same answer. One engine, answering a comparison question about two other agencies, quoted Broadcastwell's own published visibility page as a source and named Broadcastwell as the publisher of that page.

Reasoned. That is a citation of the research and not a recommendation of the firm. Reported as a visibility result it would look like an agency appearing in an answer, and it is instead a live instance of the pattern in Chapter 3 at /cited-not-recommended/. Anybody quoting the 1 of 40 as evidence of visibility, including this firm, would be quoting a number whose content contradicts its headline.

Reasoned. The two lines are not directly comparable, because one rests on five runs per question and the other on one. Neither replaces the other and both are published on every page of this manual with their dates and run structures attached. The reason for publishing both is that a manual asking buyers to demand evidence should be measurable on the same terms it sets.

Licence and status

Attribution

Sivakumar, S. (2026). The Absence Manual. Broadcastwell. docs.broadcastwell.com

Sources

The volume identifiers and citation formats are from the published Zenodo records and the source repository. The external reference list is reproduced from Volume III. The self-audit figures are the operator's own, published with their dates and run structures. Everything is at github.com/Broadcastwell/state-of-geo-2026.

Half its own shortlist

The finding

Measured, Volume III. Asked the identical buyer question again inside the same collection window, Google AI Overviews agreed with roughly half of its own previous vendor shortlist, at a mean set agreement of 0.499 across 62 repeat pairs, and Claude at 0.442 across 74 repeat pairs, on a stratified subsample of 25 questions. **Reasoned.** Every single-run AI visibility number sold in this industry carries that variance and cannot show it. The same company measured twice can produce two different scores, and a single-run number cannot tell you how far apart the two could have been.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

Why anyone measured this

Reasoned. These systems are not deterministic. Retrieval is live, ranking is probabilistic, and the answer is generated rather than looked up. Ask the same engine the same question twice and the shortlist moves.

Reasoned. That has a consequence for every cross-engine comparison ever published. Any measured difference between two engines contains an unknown amount of the same variation you would get by asking one engine twice. Unless the noise floor is measured on the same questions in the same window, a divergence headline is not interpretable at all. It is a number whose denominator has not been stated.

Measured, Volume III. The 2026 State of GEO, Volume III, put 280 B2B software buyer questions across 40 categories to four production AI search engines in a single window, 05:28 to 16:29 UTC on 2026-08-04. A stratified subsample of 25 questions, balanced across question shapes, was run three times on all four engines specifically so the engine-to-engine difference could be read against each engine's own run-to-run noise.

What went wrong, and why it is disclosed

Measured, Volume III. Three of the four API accounts ran out of credit during collection. One was topped up and finished, two were not. The study as executed holds 18 ChatGPT, 280 Claude, 175 Perplexity and 280 Google AI Overviews answers, against 280 planned for each. Rows carrying a billing error are excluded from every denominator and retained in the raw data so the exclusion is checkable.

Measured, Volume III. The two short engines were deliberately not refilled. A refill collected in a second window would make every cross-engine pair span two windows, so engine drift between windows would read as engine divergence and inflate the exact quantity the study was measuring. **Reasoned.** Staying short was the more honest option, and the four-engine intersection is reported at its true size of 18 questions rather than dressed up.

Measured, Volume III. Only the two engines that completed contributed repeat pairs. ChatGPT and Perplexity contributed none and therefore get no verdict.

The numbers

Engine	Repeat pairs	Same engine, repeated	Different engines	Gap (95% CI)
Claude	74	0.442 (0.389 to 0.494)	0.277 (0.222 to 0.333), n = 37	0.165 (0.082 to 0.242)
Google AI Overviews	62	0.499 (0.441 to 0.556)	0.240 (0.186 to 0.295), n = 35	0.258 (0.178 to 0.338)

Reasoned. The left-hand figure is the one that matters commercially. The best case here is an engine agreeing with itself on roughly half its own shortlist across runs of the identical question.

The split, so this cannot be read as overclaiming

Reasoned. The gap between the two columns looks like clean evidence that engines genuinely differ. On one engine it is. On the other it is not, and reporting only the survivor would have been the easy mistake.

Measured, Volume III. Engines name very different numbers of vendors per answer: 4.77 for Google AI Overviews, 8.83 for Claude, 10.69 for Perplexity, 14.72 for ChatGPT. **Reasoned.** Set overlap between two similar-length lists runs mechanically higher than between two lists of very different length. Same-engine repeat pairs are length-matched by construction. Different-engine pairs are not. So the comparison was rerun three more ways: restricted to the study universe, truncated to each answer's first five named vendors, and with the measure swapped for an overlap coefficient normalised by the smaller set.

Measured, Volume III. On Google AI Overviews the gap survives all three, at 0.325, 0.167 and 0.193 respectively, all positive at 95%. On Claude it does not. The raw gap of 0.165 falls to 0.075, with an interval of −0.014 to 0.177, under truncation, and to 0.002, with an interval of −0.112 to 0.113, under the overlap coefficient. Both include zero.

Reasoned. On Claude, what looks like cross-engine divergence cannot be distinguished from the fact that the engines being compared name lists of different lengths.

Why the pooled figure is not the headline

Measured, Volume III. Pooled across engines, same-engine agreement is 0.468 (0.427 to 0.505) over 136 repeat pairs against different-engine agreement of 0.247 (0.201 to 0.298) over 51 pairs. That looks decisive and it is not, because 54.4% of those same-engine pairs are Claude. **Reasoned.** A pooled mean is dominated by whichever engine contributed the most pairs and can report a separation only one engine actually has. That is an unstated denominator, which is exactly the failure the study exists to name.

What this means if someone sells you a score

Reasoned. Anyone selling a number described as "your AI search visibility" without naming the engine, the date and the question set is selling a number whose denominator is unstated.

Reasoned. Four questions settle it. Which engine produced this. On what date, in what window. What were the questions, in full text. How many runs did each question get, and if one, what is the run-to-run variance on those questions. A supplier doing the work properly answers all four in a sentence each. A supplier who cannot answer the fourth has sold you a point estimate from a distribution without telling you its width.

Reasoned. You can measure your own noise floor in an afternoon. Take five of your buyer questions, put each to the same engine three times on the same day, write down the vendors named each time, and compare the runs pairwise. That average is the number every other visibility figure you are shown should be read against.

What this does not claim

Reasoned. It does not claim that variance makes measurement pointless. Variance is a property of the system that can be quantified and reported, and this study quantified it. It also says nothing about the run-to-run stability of ChatGPT or Perplexity, which contributed no repeat data. **Measured, Volume III.** Everything here was collected in a single window on one day, on products under continuous change, Google AI Overviews is scraped rather than API-served, and the engines measured are production endpoints rather than the consumer applications people type into.

The full argument, the complete robustness table, the pooled figures and every limitation are in Chapter 6 of The Absence Manual, Your Score Is Noisier Than Your Vendor Admits.

Source data: The 2026 State of GEO, Volume III, 10.5281/zenodo.21789120. Data and code at github.com/Broadcastwell/state-of-geo-2026.

Same score, different problem

The finding

Measured, Volume II. Across 616 recorded absences from 85 B2B software companies in 60 categories, what a company loses to an AI answer changes shape as its visibility rises: companies named in zero of ten answers lose category-level questions 55.3% of the time, while companies named in seven or more of ten lose head-to-head comparisons 68.8% of the time. The pattern holds across all four visibility tiers at chi-square 61.8, df 15, $p = 1.3 \times 10^{-7}$. **Reasoned.** Two companies can hold the identical visibility score and be at opposite ends of what they need to do next.

> **Operator disclosure.** Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

Two companies, one number

Reasoned. A visibility score of 2 out of 10 tells you that eight answers went by without you. It does not tell you what those eight answers were about.

Reasoned. Picture two companies that both score 2. The first is losing every best-of shortlist in its category. When a buyer asks an AI engine who the players are, the answer lists four vendors and this company is not one of them. It is not being rejected. It is not being considered.

Reasoned. The second company appears on those shortlists. It shows up when the buyer asks who exists. What it loses is the eight questions of the form "this company versus that one," where a buyer has already narrowed to two names and wants a verdict. It is in the set and losing inside it.

Reasoned. Those are different problems. They have different fixes, different content requirements and different costs. The single percentage hides the difference completely.

Inverting the measurement

Measured, Volume II. The usual approach counts the answers where a company appeared. The 2026 State of GEO, Volume II, did the opposite: it kept the full text of every buyer question where the company was not named, and asked what those questions had in common.

Measured, Volume II. The method was 85 B2B software companies across 60 categories, ten buyer questions each, put to one AI engine with live web search, one run per question. Every question where the company was not named was retained with its verbatim text. That produced 616 absence records, and completeness was validated rather than assumed: for all 85 companies, the absence count equals ten minus the visibility score, with zero mismatches.

Measured, Volume II. Each absence question was then assigned one shape by regular-expression rules applied in fixed precedence, so a question containing both "best" and "vs" counts as a comparison rather than a shortlist. The rules and their order are published, so the classification is reproducible by anyone who wants to check it. 2.8% of records fell to a residual "other" category.

The ladder

Visibility	Companies	Absences	Category-level	Head-to-head
Named 0 of 10	30	300	**55.3%**	20.0%
Named 1 to 3	25	199	48.2%	24.6%
Named 4 to 6	21	101	20.8%	41.6%
Named 7 to 10	8	16	12.5%	**68.8%**

Measured, Volume II. Category-level means best-of shortlists plus alternatives-to-incumbent queries.

Measured, Volume II, and reasoned from it. Read the two right-hand columns downward. They move in opposite directions, monotonically, across every tier. That is the shape of a sequential process rather than a single wall.

Measured, Volume II. Two things about the bottom row should travel with it wherever it is quoted. Nine companies scored 7 or above, but one of them scored 10 of 10 and therefore produced no absence records at all, so the top tier rests on the remaining 8 companies and their 16 records. Company counts across the four tiers sum to 84 here and to 85 in the sample, for that reason. **Reasoned.** The 68.8% figure is the most quotable number in the study and it sits on the smallest base in it.

Measured, Volume II. The relationship survives that small base being removed. Visibility against absence-is-comparison runs at $r = 0.234$, $p = 4.1 \times 10^{-9}$ across all 616 records, and excluding the top tier entirely it still holds at $r = 0.184$, $p = 5.9 \times 10^{-6}$, $n = 600$.

Two gates

Measured, Volume II, and reasoned from it. A company named zero times is failing at the category door. The engine does not consider it a member of the set when a buyer asks who the players are. More than half of its losses are questions that never mention a competitor by name.

Measured, Volume II, and reasoned from it. A company named seven or more times has cleared that door. It is a recognised member of the category, and what it loses is the second gate: direct comparison against a specific named rival.

Reasoned. The middle tiers sit exactly where you would expect if those two gates are sequential rather than simultaneous.

What to do with this

Reasoned. The work implied by each stage is different. A company at the category door needs entity presence and inclusion in the third-party sources engines consult when assembling a set. Its problem is membership. A company at the comparison gate needs comparison content a model can quote against a specific named competitor. Its problem is preference.

Reasoned. Prescribing the second to a company stuck at the first is a common and expensive mistake, and a single visibility percentage cannot tell you which one you are.

Reasoned. The diagnostic costs nothing. Write down the ten questions a real buyer in your category would type. Put each to an AI engine with web search on. Record only whether you were named. Keep the full text of the questions where you were not. Sort those by shape, collapse best-of and alternatives into one category-level bucket, keep head-to-head comparison separate, and count. Whichever bucket dominates tells you which gate you are at.

What this does not prove

Reasoned. The ladder describes where a company sits, not why it got there. Nothing in the dataset is an intervention study: no company was measured, changed and measured again. The sample is challenger-skewed by construction, so the zero-visibility rate is a property of this sample and not of B2B software generally. All answers came from one engine, one run per question, and run-to-run variance was not measured here, so treat the direction of the finding as the result rather than the decimal.

The full argument, the six shape rules with examples, the complete statistics and every limitation are in Chapter 2 of The Absence Manual, The Absence Ladder.

Source data: The 2026 State of GEO, Volume II, 10.5281/zenodo.21586091. Data and code at github.com/Broadcastwell/state-of-geo-2026.

About this manual

Author. Sairam Sivakumar, Broadcastwell.

Operator disclosure. Broadcastwell ran this measurement and sells services in the category it measures. Broadcastwell is excluded from the measured sample and from every ranking. The mitigation is not that the conflict is absent, it is that the raw data and the code are public and the result can be recomputed by anyone who disagrees.

Self-audit, July 2026. In the four-engine, five-run self-audit published alongside Volume II in July 2026, Broadcastwell was named in 0 of 200 answers and cited 0 times among 663 citations. That published figure stands with its date and is never replaced.

Self-audit, 18 August 2026. Re-measured on the same ten published questions across the same four engines, at one run per question rather than five, between 03:51 and 04:03 UTC on 18 August 2026: named in 1 of 40 answers, and cited once among 674 citations. The single naming and the single citation are the same answer, in which the engine quoted Broadcastwell's own published visibility page as a source. The two lines are not directly comparable, because one rests on five runs per question and the other on one.

Not peer reviewed. This is an independent industry study published as an open dataset with the analysis code that produced every figure in it. It has not been through academic peer review. Read it as measurement, and check the measurement. If you disagree with a number here, recompute it from the public data and publish what you get.

The three volumes.

Volume	What it covers	DOI
Volume I	85 companies, 61 categories, 860 scored answers and 5,160 citations, one engine held constant. The dataset README additionally records 1,753 unique domains cited	10.5281/zenodo.21537014
Volume II	The Absence Ladder. All 616 absence records classified by question shape	10.5281/zenodo.21586091
Volume III	Cross-engine divergence. 280 questions, 40 categories, four engines, 853 answers	10.5281/zenodo.21789120

Data and analysis code for all three volumes: github.com/Broadcastwell/state-of-geo-2026.

Version 1.0, August 2026.