

2026 B2B AI Citation Benchmark

AI engines know who these companies are. They just never recommend them.

We asked ChatGPT, Perplexity and Gemini to recommend a B2B vendor 54 times across six industries. Not one of the 30 companies we tested was named — though all three engines described them accurately the moment we asked by name.

Gregg Kell, Founder and Editor · Spotlight on Startups

Original research · Published 25 August 2026 · Data collected 17–19 August 2026

0 of 54

Number of times an AI engine, when asked to recommend a provider, named one of the 30 companies in this study

Executive summary

We asked ChatGPT, Perplexity and Gemini to recommend B2B vendors — "the best companies for X", "who should I consider", "the top-rated vendors this year" — across six industry categories. Across 54 independent runs, the three engines named none of the 30 companies in our sample. Not one, not once.

This is not a recognition problem. When we named the same companies directly, the engines described them in specific, accurate detail — founding years, head-office cities, named founders, service lines — citing them in 97.5% of runs (96.7% if we exclude the one prompt that cannot fail by design).

These companies are not invisible to AI engines. They are invisible only at the moment a buyer asks who to hire — cited 97% of the time when named, and 0% of the time when not.

That gap is the finding. Being known and being recommended are different capabilities, and the second is where small and mid-sized B2B firms are currently absent. Meanwhile 28 of the 30 companies scored below 50 out of 100 on an AI-readiness audit, with the weakest layer — monitoring whether engines mention you at all — averaging 5.3, with eleven companies scoring zero.

We also set out to test whether higher AI-readiness scores predict more citations. We could not answer that question, and we explain why in full below rather than dressing up an inconclusive result.

What we tested

Thirty B2B companies, five in each of six categories: business and management consulting, B2B financial services, HR/staffing and recruiting, IT services and managed service providers, earned media and digital PR, and B2B SaaS (project management tools).

Each company was put through eight fixed prompts on each of three engines — ChatGPT, Perplexity and Gemini — producing 720 coded observations from 504 independent runs. Every response was coded Cited, Mentioned or Absent against a written rule, and every code is backed by a recorded excerpt.

The eight prompts fall into two groups, and that division turned out to be the study's central axis:

Prompt group	Company named in the question?	Prompts
Discovery	No	3
Named-company	Yes	5

Discovery prompts are how a buyer actually searches when they do not yet know who to call. Named-company prompts test whether the engine knows a company once it has been handed the name.

Finding 1 — Asked to recommend, the engines named no one in our sample

Across 54 independent discovery runs — six categories × three discovery prompts × three engines — not one of the 30 companies was named. The exact one-sided 95% upper bound on the true citation rate is 5.4%.

What the engines returned instead was consistent and, in hindsight, obvious: national and global brands. Asked about staffing they named Randstad, Robert Half, ADP, Adecco, ManpowerGroup, Insperity and TriNet. Asked about consulting they named McKinsey, BCG, Bain, Deloitte, Accenture and EY-Parthenon. Asked about bookkeeping they named Pilot, inDinero, Decimal, Bookkeeper360 and Kruze Consulting. They also leaned heavily on a small set of third-party rankings — Forbes, Vault, G2 and ClearlyRated.

None of that is surprising on its own. What is striking is the absence of any middle ground. There was no partial credit, no occasional mention, no single category where a smaller firm broke through. **The result is a clean zero.**

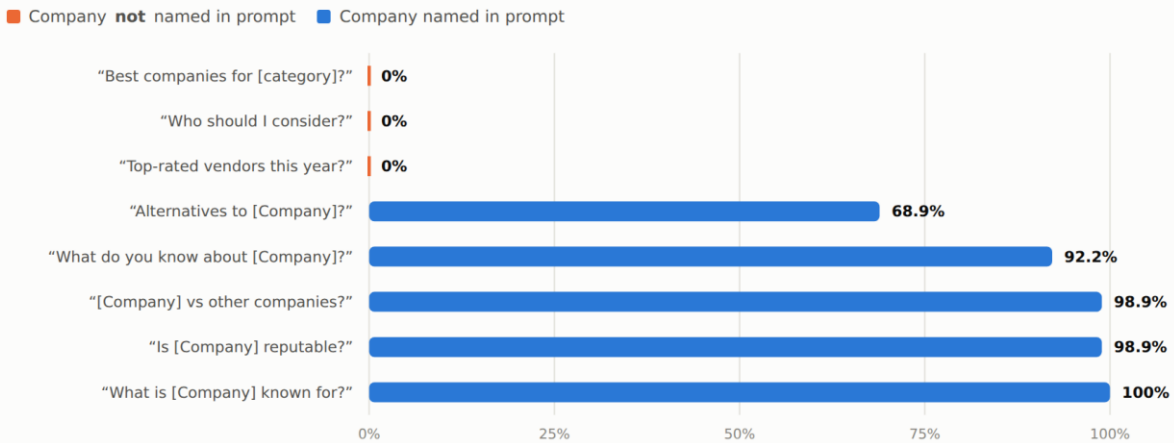
We call this a discovery blackout, and we use the term in a specific, measurable sense: **a company or set of companies returns zero citations across every discovery-prompt run in a test, while remaining recognizable to the same engines when named directly.**

Two things that definition deliberately excludes. It is not a claim about intent — we have no evidence that any engine suppresses anyone, and "blackout" here describes an observed outcome, not a decision by anybody. And it is not a claim about the whole internet: it is a statement about what three engines returned for eighteen specific questions on two days in August 2026. A blackout in this sense is a measurement, and it can be checked, disputed, or overturned by re-running the same prompts.

FIGURE 1

Citation rate is set by the question, not by the company

Share of runs where a sample company was cited, by prompt. The single largest effect in the study is simply whether the company's name appears in the question.



Each prompt run once per company across ChatGPT, Perplexity and Gemini; n=90 observations per prompt (30 companies × 3 engines). Discovery prompts are category-level, so their 90 observations derive from 18 independent runs.

Source: 2026 B2B AI Citation Benchmark — Spotlight on Startups

Figure 1 — Citation rate by prompt, all three engines pooled.

Finding 2 — Recognition is not the constraint

The same companies that were invisible in discovery were described accurately and in detail the moment we named them. Across the four prompts that name the company and ask the engine about that company, engines cited it in 351 of 360 observations, or 97.5%. One of those four, “What is [Company] known for?”, returned a citation in every single run, which is close to guaranteed by its phrasing. Excluding it entirely, the figure is 261 of 270, or 96.7%. We report both so the number survives a skeptical recount.

The fifth prompt that names the company, “what are some alternatives to [Company]?”, is the study’s one band with real variation, and it is reported separately throughout. It is worth being precise about why we separate it, because we ran a second comparison-style prompt as well: “How does [Company] compare to other companies in [category]?” returned 98.9%. Both name the company and both invite comparison. The difference is what they ask the engine to produce. One asks it to talk about the company it was handed. The other asks it to name other companies, and it is the only named-company prompt an engine can answer completely without citing the company at all. That is where the variation lives, and it is why we keep it in its own band.

The detail was specific and checkable. Gemini identified TMC Financing as a non-profit Certified Development Company and a top SBA 504 lender based in Oakland, California. It placed Bytagig in

Clackamas, Oregon, founded in 2009 by John Jackson. It named CadenceSEO's two founders. ChatGPT knew H2R-Solutions was founded in Austin in 2016 and worked in fractional HR and managed payroll — while noting, pointedly, that its public documentation was thin.

The engines are not failing to recognize these companies. They are failing to retrieve them.

Those are different problems with different solutions, and almost all published advice addresses the first one.

One caveat belongs here rather than in a footnote. Our strongest named-company prompt — "What is [Company] known for?" — was cited in 90 of 90 runs. A prompt that hands the engine a name and asks it to elaborate cannot really fail, and we would not present that 100% as evidence of anything except that the engines will repeat a name given to them.

Finding 3 — All three engines produced the same blackout independently

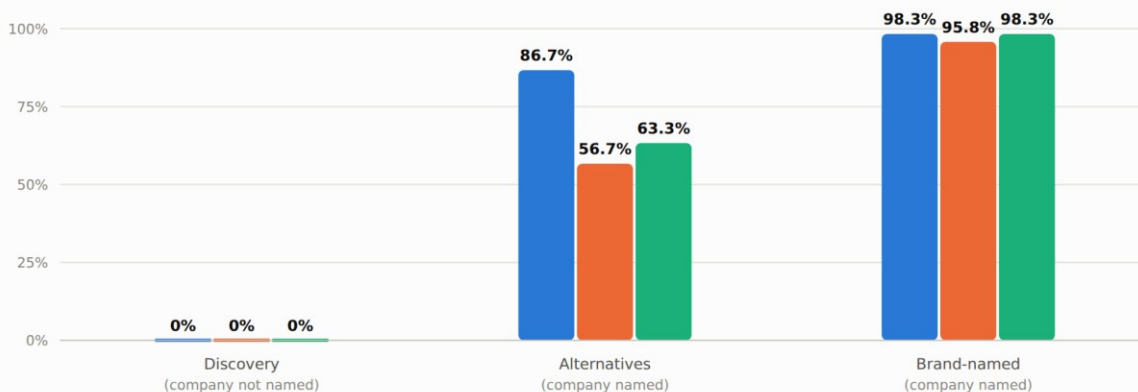
The three engines were tested separately, on different days, under different account conditions. They agreed unanimously on 100% of discovery observations and 97.5% of named-company observations.

Where they diverged was the "alternatives to [Company]" prompt — the only band in the study with meaningful variance — where unanimity fell to 40%. ChatGPT cited a sample company in 86.7% of those runs, Gemini in 63.3%, Perplexity in 56.7%. **That spread is itself a finding: engine choice matters for borderline visibility, and not at all for the blackout.**

FIGURE 2**All three engines agree — including on zero**

The discovery blackout replicates independently on every engine tested. Engines diverge only on the “alternatives” prompt, which is also the only band with meaningful variance.

■ ChatGPT ■ Perplexity ■ Gemini



Tested 18–19 Aug 2026, signed out (ChatGPT, Perplexity) or on a neutral history-disabled account (Gemini). Engines agree unanimously on 100% of discovery and 97.5% of brand-named observations, but only 40% of “alternatives”.

Source: 2026 B2B AI Citation Benchmark — Spotlight on Startups

Figure 2 — Citation rate by engine and prompt group.

Finding 4 — 28 of 30 companies scored below 50 on AI readiness

Each company was scored 0–100 on a five-layer AI-readiness audit covering earned media, entity clarity, citation architecture, distribution and monitoring. The mean was 35.1, the median 38, and the range 18 to 62. Only two companies cleared 50: Arketi Group at 62 and Hire With Near at 54.

The layer scores locate the weakness precisely:

Layer	Mean score
Entity clarity	47.0
Distribution	36.5
Earned media	29.2
Citation architecture	28.3
Monitoring	5.3

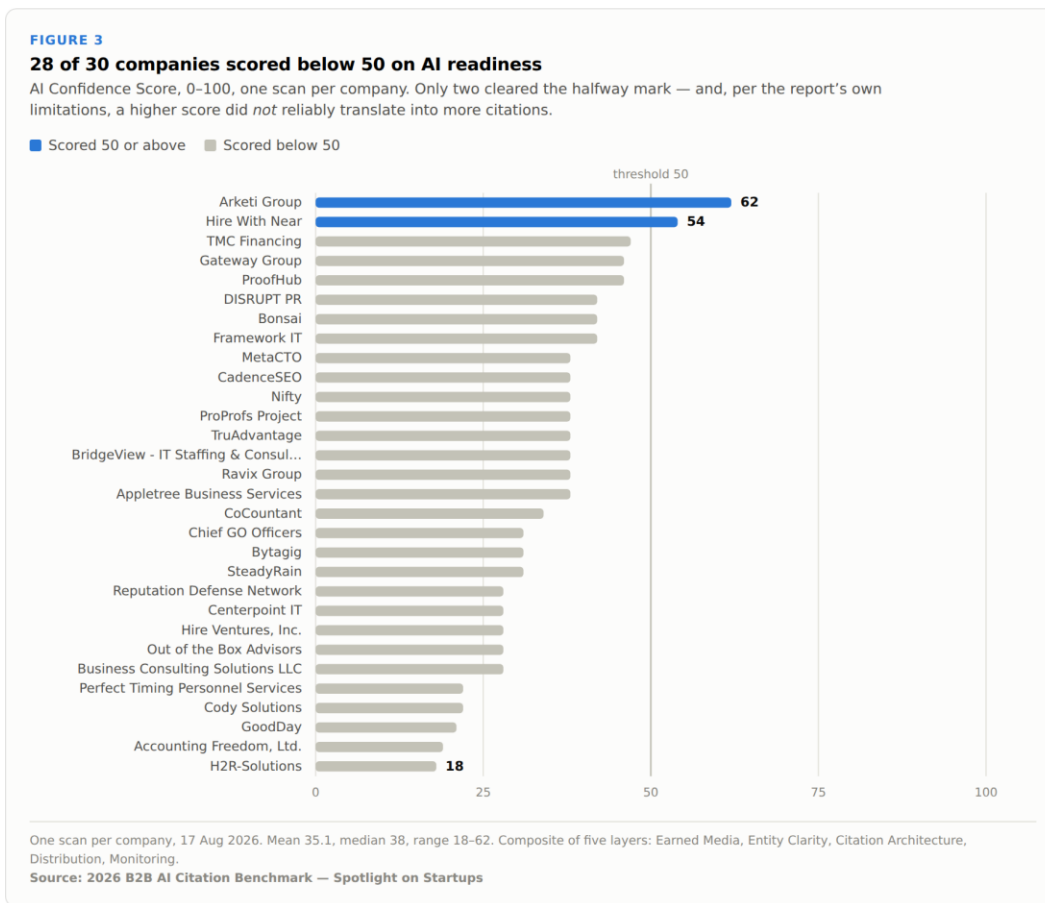


Figure 3 — AI Confidence Score by company, sorted.

Monitoring is the outlier, and the most telling number in the table. Eleven of the thirty companies scored zero — meaning they had no mechanism of any kind for detecting whether AI engines mention them. **A blackout of exactly the sort this study documents would be, for those eleven firms, entirely invisible.**

Finding 5 — Two companies lost their own names to a stock index and a houseplant

Asked the bare question, "What do you know about [Company]?", all three engines independently misidentified the same two companies.

Nifty, a project-management platform, was answered as the NIFTY 50 — the benchmark stock index of the National Stock Exchange of India. Bonsai, a business-management platform for agencies and freelancers, was answered as bonsai, the East Asian art of cultivating miniature trees; one engine closed by offering advice on selecting a beginner-friendly species.

Three engines, tested separately on different days, all answered the same two brand-name questions with a stock index and a houseplant.

Both companies were correctly identified the moment the prompt carried any category context. The failure is narrow and entirely fixable, but it is worth stating plainly: for these two firms, on the most basic possible brand query, the engine returned an answer about something else. A third company, Gateway Group, collided partially — one engine of the three could not resolve which Gateway Group was meant.

Engines also flagged name ambiguity on their own initiative. All three noted that two different firms operate as "Business Consulting Solutions LLC"; Gemini separately distinguished the Newport Beach investor-relations firm Gateway Group from an unrelated software company of the same name, and separated Cody Solutions from a similarly-named AI product.

Finding 6 — Perplexity had 133 firms to offer, and none of them were ours

A zero is only interesting if there was room to be included. So, we went back through the discovery answers and counted who the engines actually recommended.

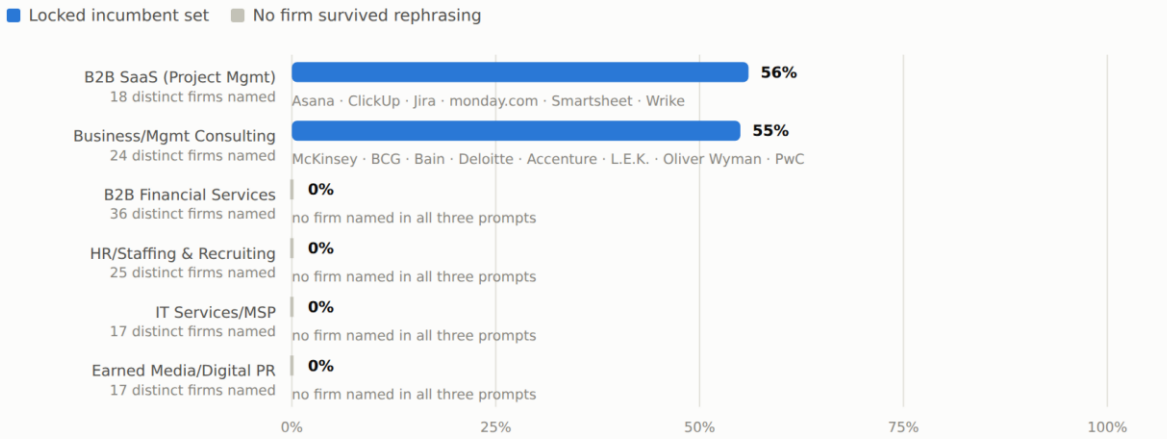
We did this on the 18 Perplexity discovery runs, because Perplexity is the one engine that issues a permanent URL for every result — meaning any reader can re-open the exact answers we counted and check our arithmetic. Across those 18 runs Perplexity named 133 distinct companies. Not one was in our sample.

Across 18 attempts to get a recommendation, Perplexity produced 133 different companies. None of them was one of the thirty we were testing.

Counting who gets named also revealed something we did not expect. **Concentration is not a property of AI recommendation in general — it is a property of the category.** In two of the six categories, a fixed set of incumbents appeared in every single differently-phrased discovery prompt. In the other four, no firm at all survived rephrasing.

FIGURE 4**Perplexity had 133 firms to offer. None of them were ours.**

Across 18 discovery runs Perplexity recommended 133 distinct companies, none from our sample. In two categories a fixed set of incumbents appeared in every differently-phrased prompt. In the other four, no firm at all survived all three rephrasings.



Measured on the 18 Perplexity discovery runs — the only subset with permanent per-run URLs a third party can re-open and re-count. "Locked" = named in all three differently-phrased discovery prompts for that category.

Source: 2026 B2B AI Citation Benchmark — Spotlight on Startups

Figure 4 — Share of recommendation slots taken by firms named in all three discovery prompts, by category.

In B2B project-management software, six vendors (Asana, ClickUp, Jira, monday.com, Smartsheet and Wrike) took 56% of every recommendation slot regardless of how the question was worded. In management consulting, eight firms took 55%. But in financial services, staffing, IT services and PR, no firm at all survived all three rephrasings. B2B financial services returned 36 distinct firms across the three prompts, and the firms that did recur appeared in only two of them.

For a company trying to be recommended, those two situations are not the same problem. In a locked category the recommendation set is effectively closed and stable. In a churning category the set is open — the engines clearly have no settled view — and yet our sample companies were still absent from all of it.

What this means

The instinct on reading a result like this is to treat it as a visibility problem and reach for the familiar remedies — more content, more keywords, more presence. Our data does not support that reading, and we want to be careful not to imply it does.

What the data shows is narrower and stranger. These companies are already in the models. The engines returned founding years, head-office cities, named founders and accurate service descriptions for firms with a few dozen employees. The information is there. It is simply not reachable through the question a buyer actually asks. A prospective client who types "who are the best managed IT providers for my company" is served a list assembled from national brands and a

handful of third-party rankings, and nothing a mid-sized MSP has published appears to enter that list at all.

That reframes the problem. The task is not becoming known to an AI engine — for these thirty companies, that has already happened. The task is becoming retrievable at the moment of recommendation, and those are different capabilities with different solutions.

What we cannot tell you

We cannot tell you what fixes this, and we are not going to pretend otherwise. We built this study around an AI-readiness score precisely because we expected it to predict citation, and it did not do so reliably. Any confident prescription about how to appear in AI recommendations, including one from us, is currently running ahead of the evidence.

Two things in the data are at least suggestive, and we offer them as such rather than as advice. The first is that citation rates on comparison-style prompts varied more by **industry category** than by anything we measured about the individual companies. We should be careful about why, though. Our own first guess was wrong. We assumed the categories with the most entrenched incumbents would be the hardest for a smaller firm to enter. The data says otherwise. Project-management software and consulting are the only two categories with a locked recommendation set, and they are locked to almost exactly the same degree, at 56% and 55% of slots. Yet they sit at opposite ends of our sample's comparison-prompt performance: project management best at 93%, consulting at 47%. Whatever is driving the category effect, a closed incumbent set at the top is not it.

The second is that entity ambiguity is a real and unusually concrete failure: two of our thirty companies lost their own brand query to an unrelated meaning, on all three engines, and both were resolved correctly as soon as the question carried any category context. That is a narrow problem with an obvious shape, and it is the one issue in this report we would feel comfortable calling straightforwardly fixable.

What we will be watching

This is a single snapshot taken over two days of prompt runs in August 2026. It cannot tell you whether the blackout is widening, narrowing or stable, and that is the question we most want answered. We intend to re-run this study annually against the same categories and the same prompt structure, with a larger sample designed for the statistical power this edition lacked.

We would change our minds about the central finding if a repeat run produced discovery citations at any meaningful rate, if a substantially larger sample showed readiness scores predicting citation, or if the pattern turned out to be an artifact of our category selection rather than a

general property of how these engines assemble recommendations. Those are the results we will be looking for, and we will publish them either way.

What we could not determine

We designed this study to answer one question: does a company's AI-readiness score predict how often engines cite it? We could not answer it — and the reason is worth stating plainly.

On discovery prompts, no company was cited at all. Every company's citation rate was zero, which leaves nothing for the score to explain: a correlation needs two quantities that vary, and here only one of them did. So the central question went untested. That is a limit of the design, not evidence that readiness doesn't matter.

One prompt band did produce real variation, and there the relationship is unresolved rather than absent. Across 30 companies, higher scores went with more citations at $r = +0.33$ ($p = 0.072$) — just short of the conventional significance threshold. But falling short of a threshold is not the same as finding nothing, and this particular result shifts depending on choices we had to make in coding the data. We would rather show you where it bends than have a critic find it first:

Method	Effect size	p-value	Significant?
Pearson (our primary coding)	+0.333	0.072	No
Spearman rank	+0.368	0.046	Yes
Counting partial mentions as hits *	+0.517	0.0035	Yes
Binomial GLM	β +0.052	0.039	Yes

* Mentioned recoded as a citation rather than as Absent — a [sensitivity analysis](#) on our coding rule, re-run with Pearson.

The 95% confidence interval runs from -0.03 to $+0.62$. At $n=30$, this study had roughly a 44% chance of detecting an effect of that size even if it were real; around 69 companies would have been needed for 80% power. Three of four defensible analytical choices return a significant positive relationship, and one does not.

The honest conclusion is that we do not know. A larger sample would settle it. We are reporting the point estimate, the interval and the sensitivity so that anyone re-running our numbers arrives where we did.

We also examined whether individual score layers performed better than the composite. Distribution and entity clarity both showed stronger associations, but neither survived controlling for industry category, and with more than a dozen tests run across the dataset, roughly one apparent result is expected by chance. We report these as exploratory only and draw no conclusion from them.

Methodology

Sample

Thirty companies, five in each of six B2B categories, selected for industry diversity rather than by size, revenue or prominence. Categories: business and management consulting; B2B financial services; HR/staffing and recruiting; IT services and MSPs; earned media and digital PR; and B2B SaaS, restricted to project-management tools.

Prompts and coding

Eight fixed prompts per company. Three are discovery prompts that do not name the company; five name it directly. Each prompt was run once per engine — the study is a single snapshot, not a longitudinal average.

Every response was coded against a written three-way rule. Cited requires the response to demonstrate knowledge beyond the prompt itself — specific services, a location, a founding year, named people, review data, or an explicit recommendation. Mentioned means the name appears but only with framing inferable from the name alone. Absent means the name does not appear, appears only as a restatement of the question, or the engine misidentified the entity. Unrecognized companies code as Absent.

Prompt wording was regenerated partway through the study to replace machine-generated category labels with natural phrasing; all 720 observations use the corrected wording.

Test conditions

ChatGPT and Perplexity were signed out. Gemini cannot run signed out, so it was run on a neutral account with Gemini Apps Activity and Web & App Activity disabled, no chat history and no stored notebooks. Account state was verified from the page itself before each session rather than assumed. Perplexity was tested 18 August 2026, Gemini 19 August 2026, ChatGPT across both days.

Counting runs

The three discovery prompts are category-level: one query serves all five companies in a category. We therefore report 54 independent discovery runs, which expand to 270 company-level observations. Across the whole study, 720 observations derive from 504 independent runs. Where we state a rate, we give the observation count; where we state the strength of the discovery finding we use the independent-run count, which is the more conservative figure.

Audit trail

Every coded row carries a recorded excerpt of the response it was coded from. Perplexity generates a permanent per-result URL, and every Perplexity row carries one, making that engine fully re-checkable by a third party. ChatGPT provides no per-run link. Gemini conversation URLs are not durable — with activity history disabled, starting a new conversation deletes the previous one — so Gemini and ChatGPT rows are evidenced by their recorded excerpts rather than by live links. We regard this asymmetry as a limitation and state it rather than obscuring it.

Verification of the headline finding

Company-name detection was performed by a token-split matcher that splits each company name at spaces, hyphens and camel-case seams and allows either separator between tokens — so "H2R-Solutions" also matches "H2R Solutions", and "GoodDay" also matches "Good Day". This matters: an earlier, naive substring match produced a false positive ("Arketi" matching inside "marketing"), and an over-strict word-boundary match then produced a false negative. Both were caught and corrected before analysis.

Because the zero-citation finding is the report's central claim, we re-verified it directly. All 18 independent Perplexity discovery runs were re-opened at their permanent URLs and re-scanned in full against all 30 company names: zero matches. The 18 Gemini discovery runs were scanned the same way at collection. The 18 ChatGPT discovery runs cannot be re-opened, because ChatGPT issues no per-run link, and were re-checked against their recorded excerpts only. In total, 36 of the 54 independent discovery runs have been re-verified against complete source text, and 18 against excerpts. We found no missed mention in any of them.

Disclosure

Spotlight on Startups is a d/b/a of Kell Web Solutions, Inc. We have published Founder Spotlight profiles of four of the thirty companies in this sample: Chief GO Officers, H2R-Solutions, TMC

Financing and MetaCTO. One of those four, H2R-Solutions, is also a paying client of Kell Web Solutions — the only client relationship anywhere in the sample. The other three are editorial only, and none of the remaining twenty-six companies has any commercial relationship with us. All four were selected on the same basis as the rest and were not exempted from any finding — H2R-Solutions and MetaCTO both appear among the small number of cases where an engine explicitly stated it did not recognize the company. That subsample is four companies, far too small to support any comparison between them and the rest of the sample, and we make none.

Limitations

Correlation, not causation. Nothing here establishes that any readiness measure causes citation. The study observes; it does not intervene.

One run per cell. AI outputs are stochastic and drift between model versions. A single pass cannot estimate variability, and individual coding could plausibly differ on a re-run.

A single moment in time. Readiness audits were run on 17 August 2026 and every prompt run on 18 or 19 August 2026. This report is a snapshot, not a trend. It cannot say whether the pattern it documents is stable, worsening or already improving, and any of those is possible. The 2027 edition is designed to answer that.

The "alternatives" prompt names the company. A citation there means the engine echoed a name it was handed, which is weaker evidence than a discovery citation. Because discovery had no variance and named-company prompts sat near a ceiling, most of our statistical analysis necessarily rests on this one prompt. We think that is a real weakness and flag it prominently.

Engines are not independent replicates. The three engines differ significantly from one another on the alternatives prompt, so pooled percentages across engines understate the true uncertainty.

Engine is confounded with test date. Because each engine was tested on a particular day, we cannot fully separate differences between engines from day-to-day drift in their indexes.

Sample size. Thirty companies supports the discovery finding, which is an unbroken zero, but is underpowered for the correlational questions. Roughly 69 would have been needed.

Score range restriction. Observed scores span 18–62 on a 0–100 scale. Correlations computed over a compressed range understate the relationship that might exist across a fuller spread.

The readiness score is a judgment instrument. The AI Confidence Score is produced by a structured but manual audit: a single assessor evaluates five layers against defined criteria and assigns each a 0–100 score. No second assessor scored the sample, so we cannot report inter-rater reliability. Judgment calls — how much weight an audio interview carries against a written trade feature, for example — were made consistently in intent but not verified for consistency. This

affects the readiness half of every correlation in this report and is one more reason we report those results as unresolved.

The SaaS category is narrow. B2B SaaS was fixed to project-management tools and is not representative of B2B SaaS generally.

Category overlap. Earned media and digital PR is the category Spotlight on Startups itself operates in. It was retained for coverage but not weighted more heavily than any other.

Vendor bias. This study was produced by Kell Web Solutions, Inc., which sells services related to AI visibility. Our mitigation is method transparency, publication of an inconclusive central result, and inclusion of our own paying client among the failure cases.

Citing this report

Spotlight on Startups (2026). The 2026 B2B AI Citation Benchmark. Data collected 17–19 August 2026 across ChatGPT, Perplexity and Gemini; 30 companies, 8 prompts, 720 coded observations from 504 independent runs.

Journalists and researchers may reproduce any figure in this report with attribution. The underlying dataset, including every coded response excerpt, is published in full: 720 coded observations, 168 permanent result URLs, and the working behind every figure in this report. Download it at spotlightonstartups.com/ai-citation-benchmark-dataset.