

AI CITATIONS SERIES · NO. 3

Can your own site get cited by AI?

Yes — about a third of the time. We set out to document a "self-published penalty" and the data refused to cooperate. So instead we tested four theories about what separates the brand-owned pages that get cited from the third-party pages that do. Three of the four were noise. This report includes them anyway.

39 top-cited pages · 12 self-published vs 27 third-party · Fisher's exact, two-tailed · July 2026 · **Perplexity only** — see scope note

31%

of top-cited pages are the brand's own

12 of 39 — your site is a real lane

1 of 4

theories survived significance testing

the other three are published too

p=0.004

the one that survived: FAQ blocks & tables

75% vs 22%

Scope, before anything else. Every number here is **Perplexity, n=39**. ChatGPT and Gemini don't expose citation URLs at all, so there's no page to analyse for them. And n=12 vs n=27 is a small sample — we ran the significance tests precisely because eyeballing percentages at this size is how you fool yourself.

FINDING 01

Your own site is cited about a third of the time

31%

Of 39 top-cited pages, **12 were self-published** — a brand's own page, ranking for a question that brand commercially cares about. **27 were third-party**: listicles, directories, review sites, comparison posts.

So the pessimistic story — "AI won't cite you, only the listicles that outrank you" — is wrong. It cites you roughly a third of the time. The interesting part is what's different about those pages.

FINDING 02 · THE ONLY ONE THAT SURVIVED

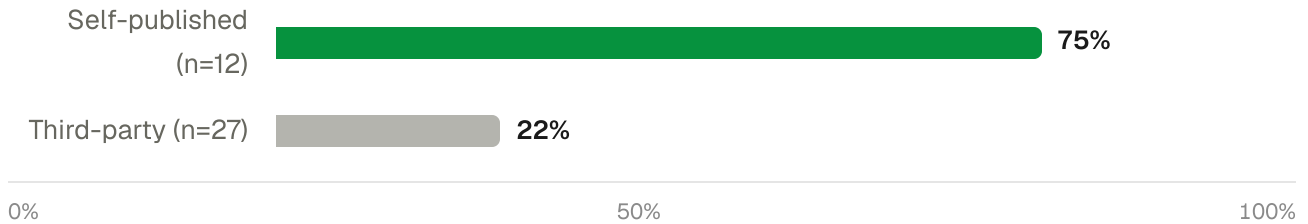
Structure was the difference. Nothing else was.

We tested four hypotheses. Here is every result, including the failures:

TRAIT	SELF-PUBLISHED	THIRD-PARTY	P (FISHER)	
FAQ blocks or data tables	75%	22%	0.004	SIGNIFICANT
Contains original data	42%	26%	0.455	NOISE
Listicle / "best-of" format	75%	78%	1.000	NOISE
Has a named author	42%	44%	1.000	NOISE

Pages carrying FAQ blocks or data tables

the only trait that separated cited brand pages from cited third-party pages



One trait clears the corrected threshold: **structured content**. Brand-owned pages that get cited are more than 3× as likely to carry an FAQ block or a data table as the third-party pages that get cited.

We think the mechanism is boring and mechanical. An FAQ block is a question with its answer directly underneath. A table is a set of facts with labels attached. Both are trivially extractable. A retrieval engine looking for something quotable finds it faster in a table than in a paragraph. Third-party listicles don't need that help — they win on being a listicle. Your page apparently does.

Structure isn't decoration. It's how the engine finds the quote.

The honest caveat. n=12. This is a strong signal in a small sample. We ran four comparisons, so the Bonferroni-corrected threshold is $p < 0.0125$, not 0.05 — this clears it, but we wouldn't bet the company on the exact multiple. We'd bet on the direction.

FINDING 03 · THE SECTION WE'D HAVE CUT IF WE WERE SELLING

Three theories we had to throw away

Original data didn't separate them. 42% vs 26% looks like a gap. At n=12 vs n=27 it's $p = 0.455$ — comfortably inside the range you'd get from coin flips. Elsewhere in our research original data is a live advantage. On *this* question at *this* sample size we can't claim it, so we aren't claiming it.

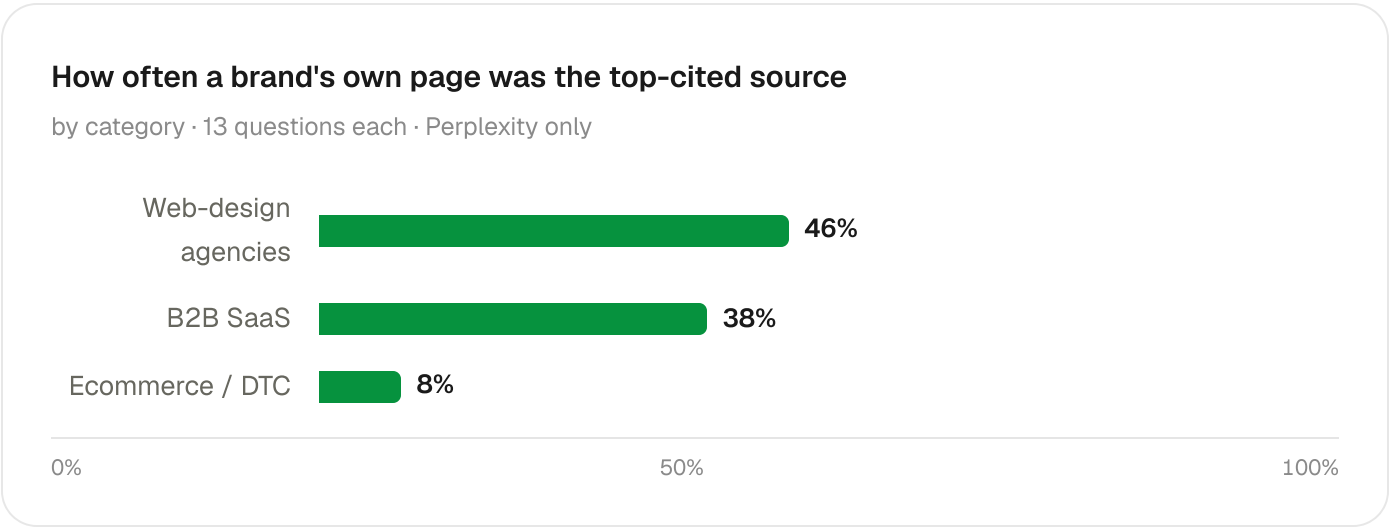
Format didn't separate them. 75% vs 78% listicle. $p = 1.000$. Cited pages are overwhelmingly listicles regardless of who publishes them. Being a "best-of" post is table stakes, not an edge.

Bylines didn't separate them. 42% vs 44%. $p = 1.000$. There's a great deal of GEO advice built on author/E-E-A-T signals. In this data a named author does nothing to distinguish a cited brand page from a cited third-party page. That's not proof it doesn't matter. It's proof we didn't measure it mattering.

And length ran the wrong way. Cited brand-owned pages had a median of **3,158 words**; third-party pages ran **3,901**. If length were the lever, self-published pages would be longer, not shorter.

FINDING 04 · THE BEST FINDING WE CAN LEAST DEFEND

Category may matter more than anything you do



An agency's own site was cited nearly half the time. An ecommerce brand's own site was cited **once in thirteen questions**. If that holds at scale it's worth more than every on-page tactic in this report combined.

But we have to be straight with you: agency vs ecommerce is $p = 0.073$. It does not clear significance at 13 questions per category. It's the single most interesting thing in this dataset and the single best argument for running the study bigger. We're running it bigger.

WHAT TO DO ABOUT IT

Four things this data actually supports

- 01

Structure your pages

FAQ blocks and data tables are the one thing that separated cited brand pages from cited third-party pages at significance. They're also cheap.
- 02

Don't over-invest in bylines on this evidence

They may well matter. They didn't measure as mattering here, and we're not going to pretend otherwise.
- 03

Know your category's ceiling

If you sell DTC, your own blog cleared 8% in this sample — third-party placement is likely the higher-leverage play. If you're an agency, your own site is a real lane.
- 04

Don't add words

Cited brand pages were *shorter* than cited third-party pages. Length is not the lever.

THE DATA

Every number, in one place

METRIC	SCOPE	VALUE
Top-cited pages analysed	Perplexity, 39 questions	39
Self-published	Perplexity, 39 questions	12 / 39 = 31%

METRIC	SCOPE	VALUE
Third-party	Perplexity, 39 questions	27 / 39 = 69%
FAQ/tables — self-published	n=12	9 = 75%
FAQ/tables — third-party	n=27	6 = 22%
FAQ/tables — significance	Fisher, 2-tailed	p = 0.004
Original data — self-pub vs third-party	n=12 / n=27	42% vs 26% · p = 0.455
Listicle — self-pub vs third-party	n=12 / n=27	75% vs 78% · p = 1.000
Named author — self-pub vs third-party	n=12 / n=27	42% vs 44% · p = 1.000
Median word count — self-published	n=12	3,158
Median word count — third-party	n=27	3,901
Self-published rate — agencies	n=13	6 = 46%
Self-published rate — B2B SaaS	n=13	5 = 38%
Self-published rate — ecommerce	n=13	1 = 8%
Agencies vs ecommerce	Fisher, 2-tailed	p = 0.073

Limits, plainly. One engine (Perplexity) — the only one that shows its sources. **n=39 pages**, split 12 / 27. One result clears the corrected threshold; the rest are reported as failures on purpose. **13 questions per category**, which is why the most interesting finding is the one we can least defend. **Coding is judgement** — "has original data" and "self-published" were coded by rule, but edge cases exist. The dataset ships with this report so you can disagree with our specific calls rather than in general. **One snapshot, July 2026**. And `verify-numbers.py` recomputes every figure above straight from the CSVs.

QUESTIONS THIS REPORT ANSWERS

Quick answers, straight from the data

Can a brand's own website get cited by AI?

Yes — 12 of the 39 top-cited pages (31%) were the brand's own page, not a third-party listicle (Perplexity, July 2026).

What makes a brand page citable?

Structure. Cited brand pages carried FAQ blocks or data tables 75% of the time versus 22% for cited third-party pages — $p=0.004$, the only significant trait of the four we tested.

Do bylines or word count help a page get cited?

Not measurably here. Named authors: 42% vs 44% ($p=1.000$). And cited brand pages were shorter than third-party ones — median 3,158 vs 3,901 words — so length is not the lever.

Is your own site getting cited?

Same method, pointed at your domain. We'll ask the engines your buyers' questions and show you whether your pages — or someone else's — come back.

[Run the free audit →](#)

The 2026 AI Citations series · No. 3

By HarperFlow · July 15, 2026 · every figure recomputes from the released dataset

Download the dataset

All six reports →