

RETRIEVAL, CITATION, AND THE END OF RANK

The Day Your Website Died

What the empirical data says about being named by a retrieval-augmented system, and which parts of the pipeline you can actually reach

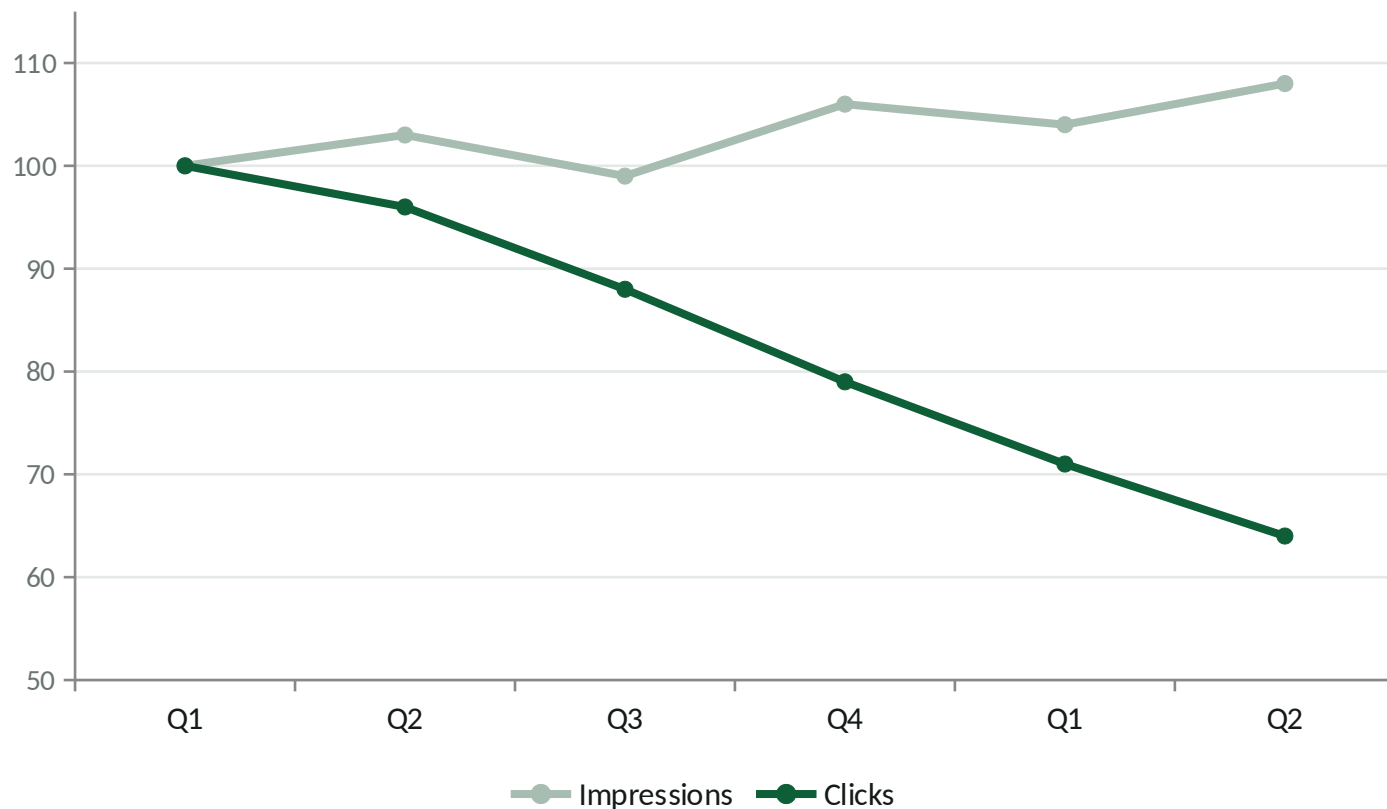
Richard Lowe · 33 years enterprise IT · 113+ books

Figures verified against their sources as of August 2026. This field moves fast; see the errata page on the closing slide for anything that has changed since.

THE OBSERVATION

Impressions held. Clicks did not.

ILLUSTRATIVE



The shape, not the volume

Flat impressions with decaying CTR is the signature of an answer layer intercepting the query.

No penalty. No ranking change. The surface is answering instead of routing.

Illustrative shape, not measured data. The divergence between the two lines is the diagnostic; the magnitudes are drawn for clarity.

The pipeline, and your two access points



You can influence what enters the candidate set and what a passage looks like when it is scored. You cannot touch query expansion, grounding, or citation policy.

Everything sold as AI SEO operates on two of five stages, and mostly on one.

Two distinct failures, two different fixes

NOT RETRIEVED

Never in the candidate set

- Crawler blocked or render-dependent
- No passage matches the sub-query
- Behind JS, auth, or a slow origin
- Entity not resolvable to a known thing

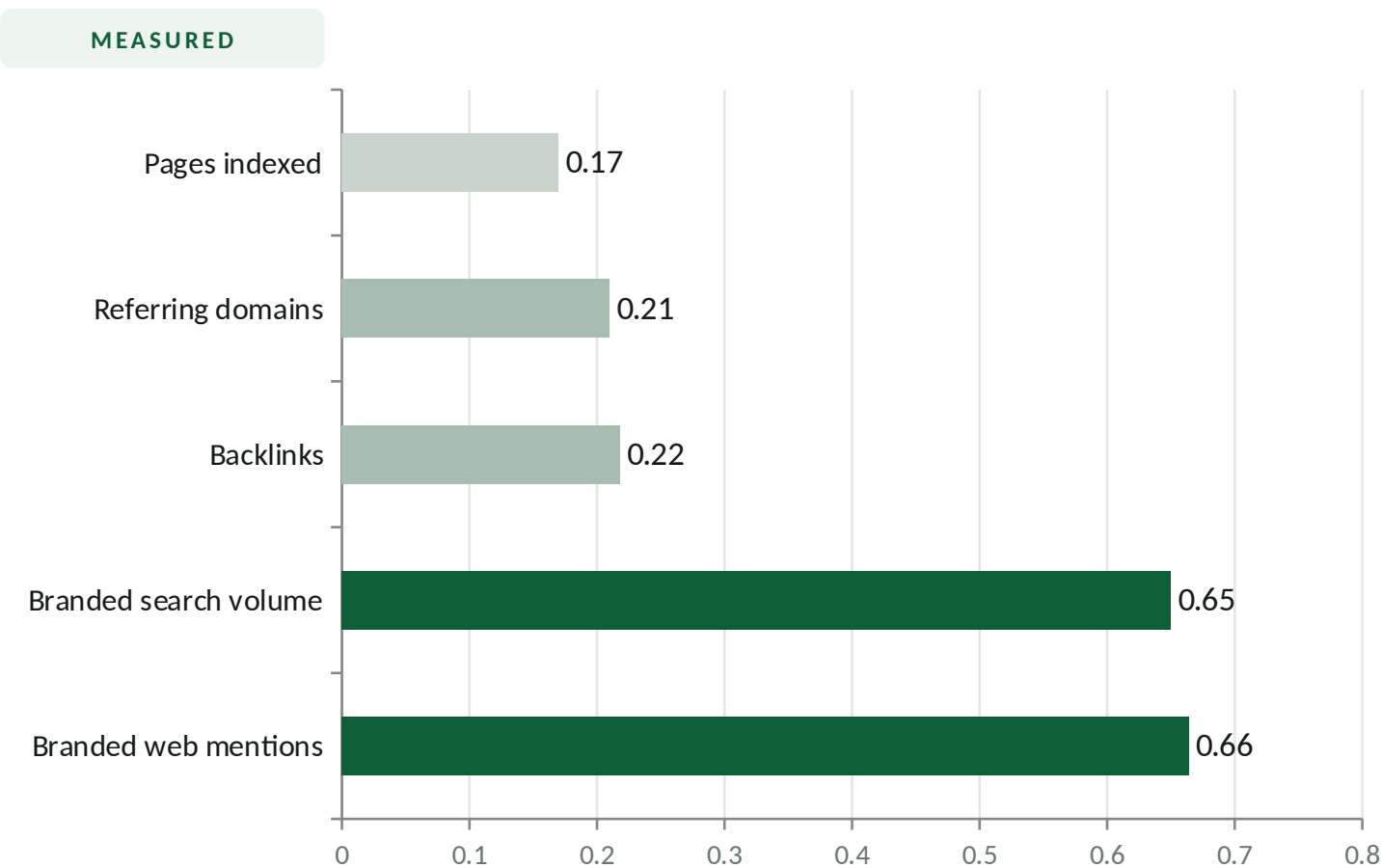
RETRIEVED, NOT CITED

In the set, absent from the answer

- Passage is hedged, so it grounds nothing
- Claim unconfirmed by any other source
- Contradicts a higher-confidence record
- A corroborated competitor answered it

Almost all published advice addresses the left box. Most real losses are in the right one.

Correlation with AI visibility, n = 75,000 brands



Read the split

Two off-site identity signals sit near 0.65. Every link-graph metric sits near 0.21. Index size is close to noise.

The link graph was a proxy for reputation. These systems have a more direct one.

Ahrefs, 75,000 brands, AI Overview visibility. Observational. Ahrefs state explicitly this is correlation and not causation; the confound with brand size is not controlled for.

Training crawlers and retrieval crawlers are different

TRAINING

GPTBot ClaudeBot CCBot Google-Extended

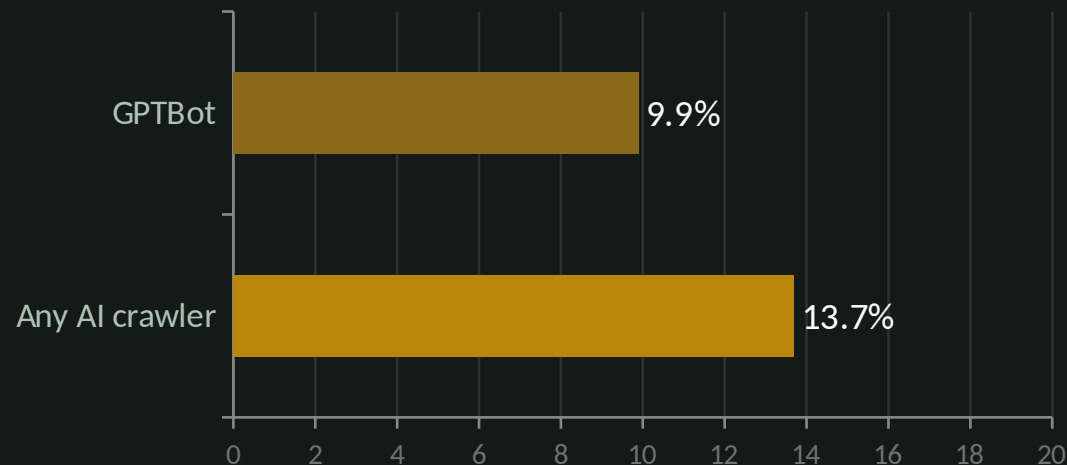
Blocking removes you from future weights. No effect on live citation.

RETRIEVAL

OAI-SearchBot PerplexityBot Claude-User

Blocking removes you from answers being composed right now.

MEASURED



Most of it is unintentional. Somebody edited robots.txt once, blocked whatever existed that year, and never revisited it.

Four systems, four retrieval strategies

Grounded live fetch

Fetches at query time. Freshness matters. Robots and render behavior decide everything.

Index-mediated

Reads somebody else's index. You inherit that index's crawl policy and its staleness.

Parametric-first

Answers from weights, retrieves only on uncertainty. Training cutoff is your ceiling.

Hybrid, undocumented

Switches strategy by query class, with no published rule for which.

The same query against all four returns four different citation sets. There is no shared ranking function to optimize against, and no published contract that any of it stays put.

Entity resolution is the actual gate

Before anything can cite you, it has to decide you are one resolvable entity and not three partial ones.

HIGH CONFIDENCE

n independent records agree on the same attributes.
Confidence compounds with source diversity, not with source count.

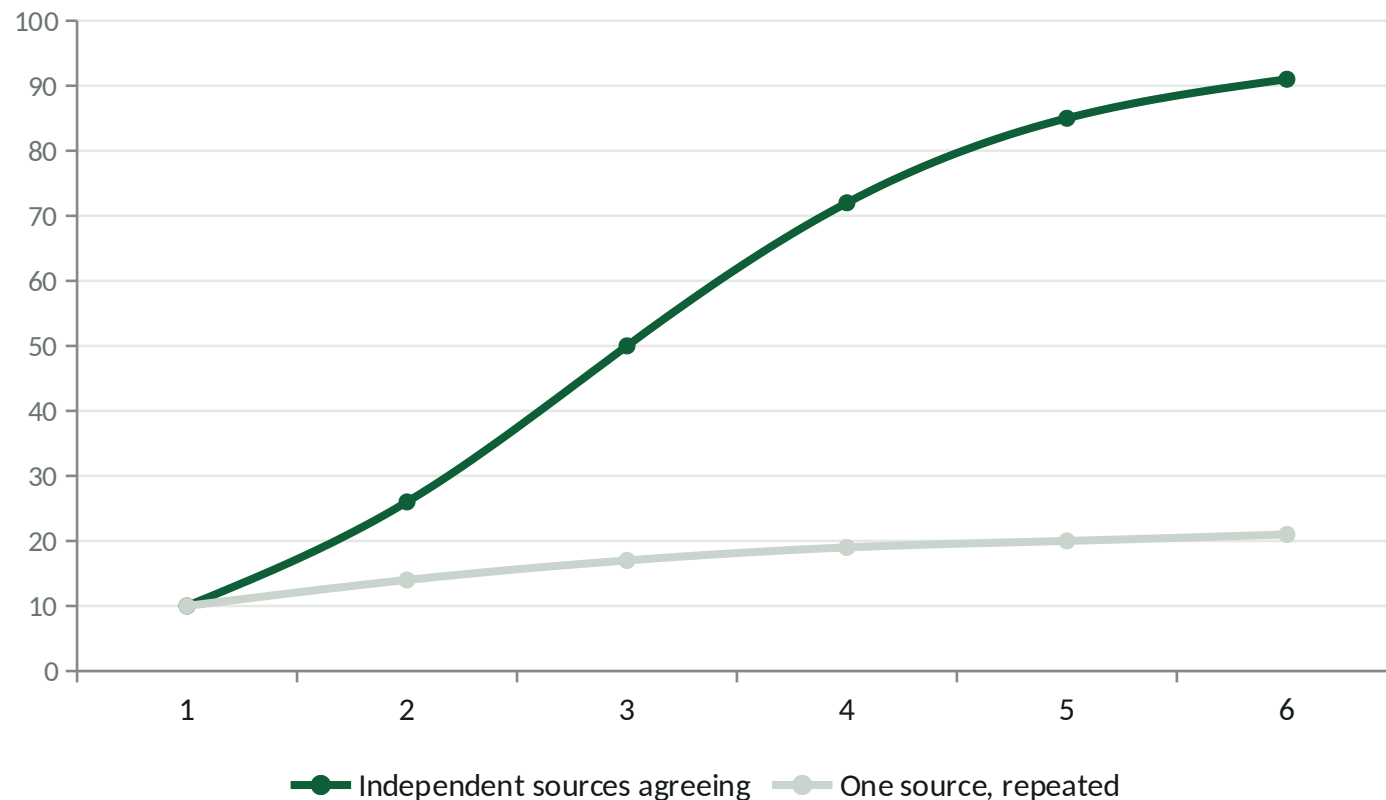
CONTRADICTION

One disagreeing attribute does not fail locally. It lowers confidence across every claim attached to the entity, including the correct ones.

Which is why an outdated job title on a conference page from 2019 costs more than a missing one.

Corroboration compounds. Repetition does not.

ILLUSTRATIVE



The practical read

The same claim on ten pages of your own domain is one source repeated. Three unaffiliated records is three sources.

Which is why index size correlated at 0.17.

The unit of retrieval is a passage

POOR CHUNK

“There are many factors to consider, and the right approach depends on your circumstances, which we explore below.”

Survives standalone. Asserts nothing. Grounds nothing.

RETRIEVABLE CHUNK

“A first PCI audit usually returns findings for controls nobody was assigned, not for controls that failed. Those are different problems.”

Self-contained, falsifiable, answers a question alone.

Chunking is done to you, not by you. Write so that any 200 words lifted out of context still assert something checkable.

Cited content skews fresher than organic results

25.7%

fresher than the organic top ten,
on average

What follows from it

- A dated page is a weaker candidate than the same page undated
- Revising a strong page beats publishing a new weak one
- Parametric-first engines invert this, since their cutoff is fixed
- So freshness is a per-engine lever, not a universal one

Anchor claims to mechanisms and date the specifics. A mechanism does not decay. A figure does.

There is no rank, so what do you measure?

1

Sampled citation rate

Fixed query set, run per engine on a schedule. Record whether you appear and where in the answer.

2

Share of voice

Of n competitors named across the query set, what fraction of slots are yours. Noisy, directionally useful.

3

Referral fingerprints

Traffic from assistant domains, plus enquiries citing things you never told them.

All three sample a non-stationary, non-deterministic system. Run identical prompts, log raw answers, and treat single readings as noise.

Measured, inferred, and unknown



Measured

Crawler access. Whether you are named for a query at a point in time. Correlations across large brand samples.



Inferred

How confidence is computed. Why one source is preferred. What triggers retrieval versus a parametric answer.



Unknown

Citation policy. The ranking function over passages. Whether any of it is stable month to month.

Anybody selling a guaranteed method is making claims about rows two and three.

Where the effort actually goes

● Audit crawler access	Hours	Binary. Gates everything else.
● Deduplicate your facts	Hours	Removes the contradiction penalty across every claim.
● Rewrite for standalone passages	Weeks	Directly changes what is retrievable.
● Structured identity data	Days	Removes ambiguity in entity resolution.
● Third-party corroboration	Months	Highest correlation, slowest instrument.
● Publishing volume	Ongoing	$r = 0.17$. Do it for readers, not for this.

The top two are free, take an afternoon, and are the two almost nobody has done.

Every figure in this deck, and what it rests on

Correlation with AI visibility

Ahrefs study of 75,000 brands against AI Overview visibility. Observational and uncontrolled for brand size, which Ahrefs state themselves. Slide 5.

AI crawler blocking rates

Crawl of 1,744 sites: 13.7% blocking at least one AI crawler, 9.9% blocking GPTBot. Other samples report figures near 22%. Slide 6.

Freshness of cited content

Meta-analysis across 54 studies. Aggregate across engines; per-engine variance is large and unreported. Slide 11.

Impressions and clicks, corroboration curve

Illustrative, not measured. Drawn to show a shape and an ordering. Do not cite either as data. Slides 2 and 9.

Pipeline stages, entity confidence

Inferred from observed behavior and public documentation. No vendor publishes a citation policy or a confidence function. Slides 3 and 8.

Nothing in this deck is proprietary or under NDA. Reuse any of it with attribution. If a figure has moved since August 2026, the errata page will say so.

An answer has no page two.

Ranking was a list you could hold a position in. Citation is a set membership decision made per query, by a system with no published contract and no feedback channel.

The Day Your Website Died

368 pages · 42 chapters · sources dated

Living errata page for the numbers

thewritingking.com/aeo-updates