RESEARCH

How a page reaches an AI answer, and where it usually breaks

Four things have to happen in order before an assistant can say your name. Three of them are documented. The one in the middle is invisible from outside, and most advice you will read quietly pretends otherwise.

Written by Ali Kamran
SEO & GEO Consultant · Published August 22, 2026 · 7 min read

People ask me why an AI does not mention their product, and the question usually skips several steps. Four separate things have to happen before an assistant can say anything about your page, each one run by a different system, each one able to fail on its own without telling you.

Step 1: something has to fetch the page

A crawler has to request the URL and get real content back. That is the only step you can see directly, because it lands in your server logs.

It fails for boring reasons. robots.txt says no. A firewall rule blocks the user agent. The page returns a 200 with an empty shell and fills itself in with JavaScript, so the fetcher gets markup with no words in it. The URL redirects somewhere unhelpful.

Worth knowing which fetcher you are talking about, because the companies run several with different jobs, and blocking the wrong one is the single most common mistake I find. I wrote a full breakdown of every agent separately.

Step 2: something has to keep it

Fetched and kept are two different things, and this is where most of the confusion in this field lives.

There are two separate destinations. One is a search index, the thing an assistant queries live when someone asks a question. The other is a training corpus, which shapes what a model knows in general but does not give it your page on demand.

You can tell these are separate pipelines because the companies say so with their own naming. OpenAI runs GPTBot for training and OAI-SearchBot for search. Anthropic runs ClaudeBot for training and Claude-SearchBot for search. Perplexity says outright that PerplexityBot is not used to crawl content for foundation models. Three vendors, three explicit splits.

Google is the clearest of all on what the requirement is. To be eligible to appear as a supporting link in AI Overviews or AI Mode, Google says a page must be indexed and eligible to be shown in Google Search with a snippet, and that there are no additional technical requirements beyond that. Ordinary indexing is the gate. Nothing exotic sits behind it.

So a bot hit in your logs proves a fetch happened. It proves nothing about whether anything was kept, and nobody publishes a way to check.

Step 3: something has to retrieve it, for a query you never see

Here is the step that breaks most of the advice in this field.

When a person asks an assistant a question, the assistant does not usually search for what they typed. It writes its own query, often several, in wording nobody outside the company ever sees. Your page has to be findable for those queries, not for the sentence the human wrote.

That is why keyword thinking transfers badly here. You are optimising for a search you cannot read, generated by a system that does not publish its rewriting rules, ranked by a retrieval layer nobody has documented.

I can measure the input and the output. The middle is closed. Anyone telling you precisely why one page got pulled and another did not is reasoning backwards from a result, which is exactly the habit that made a satirical file about office cats pass four separate industry tests for whether a standard works.

Step 4: the model has to use it, and decide whether to say so

Retrieved is still not cited. The model receives a set of candidate sources and writes an answer. It may use your page and attribute it. It may use your page and attribute a different one. It may mention your brand with no link at all, which is worth something to you commercially and shows up in no analytics tool anywhere.

This is why the vocabulary matters. A mention, a citation and a referral are three different events with three different measurement methods, and treating them as one number is how AI visibility reporting becomes fiction.

Which step each tactic actually touches

A useful sanity check when someone sells you something. Ask which of the four steps it operates on, and whether that claim is documented anywhere.

COMMON TACTICS, MAPPED TO THE STEP THEY AFFECT

TACTIC	STEP IT TOUCHES	DOCUMENTED BY A PROVIDER
robots.txt rules	1, fetch	Yes, by every major provider
Server side rendering, real text in the HTML	1, fetch	Partly. Google documents rendering. Others do not.
Being indexed in ordinary search	2, keep	Yes, Google states it is the eligibility gate
Schema markup	2, keep	For Search features, yes. For AI answers, no provider confirms a direct effect.
llms.txt	Claimed 1 and 2	No. No major provider has confirmed reading it.
Clear headings, short answerable passages	3, retrieve	No. Reasonable, widely believed, unconfirmed.
nosnippet, max-snippet, data-nosnippet	4, generate	Yes, Google documents these as the AI Overviews controls
NOARCHIVE and NOCACHE	4, generate	Yes, Microsoft documents the exact effect on Copilot
Being talked about on other people's sites	2 and 3	No. Strongly suspected, not published.

Notice how much of that column says no. That is the honest state of this field in 2026, and a table like this is more useful to you than another confident list of best practices.

Where it actually breaks, from the audits I run

In practice the failures cluster at step 1, which is the least interesting step and the easiest to fix. A robots.txt line somebody added years ago. A page that needs JavaScript to show a single word of text. A bot protection rule that returns a challenge page to every fetcher that is not a browser.

The second cluster is at step 2, and it is almost always a plain indexing problem wearing an AI costume. The page is not in Google's index, so it cannot be a supporting link in AI Overviews, because Google has said that is the requirement.

Steps 3 and 4 are where the interesting work is, and where I am least willing to promise anything specific, because the mechanism is not public. What I can do there is measure: ask the questions your buyers ask, record what comes back, and repeat it on a schedule so the movement means something.

What nobody outside these companies knows

Writing this down because most articles on this topic skip it.

Nobody outside these companies knows how the internal query is written, how candidate sources are ranked, how much weight a source gets, whether personalisation or chat history changes retrieval, or how often the answer would change if you asked the identical question an hour later. That last one is measurable from outside, so I am currently measuring it.

Everything on this page above that line is either documented by a provider or observable in a server log. Below it is honest ignorance. Both are worth publishing.

> FAQ

Does being crawled by an AI bot mean my page is in the index? +

What does Google require for a page to appear in AI Overviews? +

About this PDF

This is the PDF edition of a page published at ali-kamran.com/research/how-ai-systems-reach-your-page. The web version carries test markers for a published study on AI content ingestion. Those markers are removed here, because a marker that exists in two places tests nothing.

This file carries one marker of its own, **xopuguq1421**, which appears in this PDF and nowhere else on the internet. It is meaningless. It exists to find out whether a PDF hosted on a domain gets ingested the way an HTML page does.

Primary sources, all accessed August 22, 2026: Google Search Central, AI features and your website · Google Search Central, Google common crawlers · OpenAI crawler documentation (developers.openai.com/api/docs/bots) · Anthropic support documentation on web crawling · Perplexity bot documentation (docs.perplexity.ai/guides/bots) · Bing Webmaster Blog on NOCACHE and NOARCHIVE, September 2023 · Mark Williams-Cook, "How cats.txt showed llms.txt evidence is GEO astrology."