RESEARCH

# What each AI provider actually documents about how it picks sources

Almost every piece of GEO advice you will read cites nothing. So here is the opposite: every official statement I could find, quoted, linked and dated, plus the much longer list of things none of them say.

**Written by Ali Kamran**
SEO & GEO Consultant · Published August 22, 2026 · 9 min read

## Why this page exists

Pick any ten GEO tips off LinkedIn and check how many carry a link to something a provider actually said. In my experience it is close to none. The advice circulates because it circulates.

So this is the inverse of a tips post. Every claim below is a quote from an official source, with a link and a date. Where a provider has said nothing, the entry says nothing, which is the most useful part of the page.

## Google

Google documents more than everyone else combined, and most of what it documents is a list of things you do not need to do.

The eligibility rule is the concrete part. Google states that "to be eligible to be shown in generative AI features on Google Search, a page must be indexed and eligible to be shown in Google Search with a snippet" (AI optimization guide). Ordinary indexing is the gate.

On mechanism, Google describes retrieval augmented generation as "relying on our core Search ranking systems to retrieve relevant, up-to-date web pages." So the retrieval layer is Search, not a separate system with separate rules.

Then four explicit denials, all from the same document, all of which contradict something currently being sold:

- "You don't need to create new machine readable files ... Google Search itself doesn't use them," said of llms.txt directly
- "There's no requirement to break your content into tiny pieces for AI"
- "You don't need to write in a specific way just for generative AI search"
- "Structured data isn't required for generative AI search"

That last one deserves a pause, because schema is one of the most confidently sold GEO tactics in the market. Google says it is not required. Not that it is useless for Search features, where it genuinely does things, but that it is not a requirement for generative AI search.

## OpenAI

OpenAI documents access thoroughly and selection barely at all.

On ranking, the entire published position is one sentence: "ChatGPT ranks search results using multiple factors intended to help users find relevant, reliable information.

Placement is not guaranteed" (Searching the web with ChatGPT). The factors are not listed anywhere.

On access, it is specific and actionable. "Any public website can appear in ChatGPT search." Do not block OAI-SearchBot, and confirm your host or CDN allows traffic from OpenAI's published searchbot IP addresses. That second half catches people out, because a robots.txt can say yes while a bot protection rule says no.

One nuance worth knowing, from the publisher FAQ: if OpenAI obtains the URL of a disallowed page from a third party search provider, it may surface just the link and page title. To prevent that you use `noindex`, and for the tag to be read the crawler has to be allowed to crawl the page. Blocking harder makes you less controllable, not more.

Also concrete and useful: ChatGPT appends `utm_source=chatgpt.com` to referral URLs, which is why ChatGPT referrals are measurable in analytics at all.

## Anthropic

Anthropic documents its three crawlers clearly and publishes an IP list for verification. On how Claude chooses which sources to cite, there is developer facing documentation describing that Claude generates a targeted search query, retrieves results and cites the material it used, plus domain allow and block lists for organisations building on the API.

None of that is publisher guidance. There is no statement about what makes one page more likely to be cited than another, and I would rather write that plainly than infer a set of factors from a tool description.

## Perplexity

Perplexity's public documentation covers its crawlers and WAF configuration. PerplexityBot exists to surface and link sites in results and is not used for foundation model training. Perplexity-User handles user triggered fetches and generally ignores robots.txt.

On selection or ranking, nothing. There is no published guidance for publishers on what gets cited.

## Microsoft

There is no separate Copilot crawler, because Copilot is grounded in the Bing index that bingbot builds. What Microsoft documents precisely is the effect of two meta directives.

Content tagged `NOCACHE` may appear in a Copilot answer, but only the URL, title and snippet are displayed, and only those elements may be used in training. Content tagged `NOARCHIVE` is not included in Copilot answers, is not linked in them, and is not used for training. Both still rank normally in Bing search results.

That is the most granular opt out any provider offers, and almost nobody uses it.

## The full table

WHAT EACH PROVIDER DOCUMENTS, AND WHAT IT DOES NOT

| PROVIDER | DOCUMENTED ABOUT SELECTION | EXPLICITLY DENIED | NOT ADDRESSED |
|---|---|---|---|
| **Google** | Indexed and snippet eligible is the gate. Retrieval runs on core Search ranking. | llms.txt, content chunking, AI specific writing, structured data as a requirement | Why one indexed page is chosen over another |
| **OpenAI** | "Multiple factors", placement not guaranteed. Allow OAI-SearchBot and its IP ranges. | Nothing denied outright | The factors themselves. |
| **Anthropic** | Claude writes a targeted query, retrieves, cites what it used. Developer facing. | Nothing denied outright | Any publisher facing guidance at all |
| **Perplexity** | Crawler roles only. PerplexityBot surfaces and links, not training. | Nothing denied outright | Selection, ranking, citation criteria |

| PROVIDER | DOCUMENTED ABOUT SELECTION | EXPLICITLY DENIED | NOT ADDRESSED |
|---|---|---|---|
| **Microsoft** | Copilot grounds in the Bing index. NOCACHE and NOARCHIVE effects specified exactly. | Nothing denied outright | How grounding picks among indexed pages |

## What none of them document

Read the right hand column again. Every provider leaves the same hole, and it is the hole every GEO pitch claims to fill.

Nobody publishes how the internal query is written when a system decides to search. Nobody publishes how candidate pages are ranked once retrieved. Nobody publishes how much weight a source carries, whether being mentioned elsewhere on the web feeds into it, or whether personalisation changes which sources appear.

> WORTH NOTICING
>
> Google is the only provider that has published anything resembling a negative result, and all four of its denials point the same way: the thing that works is being indexable and worth reading. Everything sold on top of that is inference. Some of it is reasonable inference. None of it is documented.

I am not arguing the undocumented tactics are worthless. Clear headings and short answerable passages are probably good for retrieval, and I still recommend them. I am arguing that "probably" is the honest word, and that a table with an empty column is more useful to you than a confident list.

> FAQ

**Does Google say structured data helps with AI Overviews?** +

**Has OpenAI published its ranking factors?** +

## About this PDF

This is the PDF edition of a page published at ali-kamran.com/research/what-ai-providers-document-about-sources. 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, **mudutiq5839**, 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, Google's guide to optimizing for generative AI features on Search · Google Search Central, AI features and your website · OpenAI Help Center, Searching the web with ChatGPT · OpenAI Help Center, Publishers and Developers FAQ · Anthropic support documentation on web crawling, and Claude platform documentation on the web search tool · Perplexity bot documentation · Bing Webmaster Blog, September 2023.