

How answer engines choose which sources to cite

Four providers publish documents about how their products reach web pages and link to them. This briefing reads those documents again, records what each says about choosing and citing sources, and records what each leaves unstated, with the source and date for every claim.

Author	Amit Govil, Founder and CEO, XONIK
Published	October 22, 2026
As of	October 20, 2026
Reading time	10 min read
Online	https://xonik.com/insights/market-intelligence/how-answer-engines-choose-sources/

IN BRIEF

As of 20 October 2026, four providers document which pages can be used as a source, and none says how it picks among eligible pages. Google requires an indexed page eligible for a snippet; OpenAI, Anthropic and Perplexity document crawlers a site owner can allow or block. None states criteria or weights. This briefing records what each says.

KEY FINDINGS

1. At Google and at OpenAI, the setting that controls training is separate from the one that controls search.

Google-Extended does not affect inclusion in Google Search and is not a ranking signal. (Sources: Google, AI features and your website (updated 10 December 2025), as of 20 October 2026; Google's common crawlers (last updated 15 October 2026), as of 20 October 2026.) OpenAI documents OAI-SearchBot and GPTBot as settings that are independent of each other. (Source: OpenAI, Overview of OpenAI crawlers (no date shown), as of 20 October 2026.)

2. Fetches made on a person's request may ignore robots.txt.

OpenAI says robots.txt rules may not apply to ChatGPT-User. (Source: OpenAI, Overview of OpenAI crawlers (no date shown), as of 20 October 2026.) Perplexity says its user fetcher generally ignores them. (Source: Perplexity, Perplexity Crawlers (no date shown), as of 20 October 2026.)

3. Blocking Anthropic's robots by IP address may not work as an opt-out.

Anthropic says its robots honor robots.txt directives and that blocking by IP address may not work as an opt-out. (Source: Anthropic Help Center, Does Anthropic crawl data from the web (dated April 8, 2026), as of 20 October 2026.)

4. Google comes closest to saying what it wants, and the other three say less.

Google states an aim of a broader and more varied group of links. OpenAI says only that multiple factors are used, and Anthropic and Perplexity describe a process without criteria. (Sources: Google, AI features and your website (updated 10 December 2025); OpenAI Help Center, Searching the web with ChatGPT (page showed an update on the day it was read); Anthropic Help Center, Enable and use web search (shown as updated over a week before it was read); Perplexity Help Center, How does Perplexity work (last modified August 28, 2026); each as of 20 October 2026.)

5. The work that can be acted on is access.

Beyond access, the providers publish no rule a firm can follow to be chosen, so advice that claims to know the rule is not supported by these documents. (Sources: Google, AI features and your website (updated 10 December 2025); OpenAI, Overview of OpenAI crawlers (no date shown); Perplexity, Perplexity Crawlers (no date shown); each as of 20 October 2026.)

6. A test of your own visibility is an observation, not a ranking.

OpenAI's own help page says citations can be incomplete, outdated or incorrect. (Source: OpenAI Help Center, Searching the web with ChatGPT (page showed an update on the day it was read), as of 20 October 2026.)

What does Google say?

A page can be a supporting link in AI Overviews or AI Mode when it is indexed and eligible to be shown in Google Search with a snippet, and Google states that there are no additional technical requirements. It also states that meeting every requirement does not mean the page will be crawled, indexed or served. Both features may issue several related searches across subtopics and data sources, which Google says lets it show a broader and more varied group of links than a classic search, and AI Overviews appear only when Google's systems judge them additive to classic Search. No new machine-readable file, AI text file or special schema.org markup is needed. (Source: Google, AI features and your website (updated 10 December 2025), as of 20 October 2026.)

Crawling for Search is controlled with robots.txt rules for Googlebot, and what is shown from a page with nosnippet, data-nosnippet, max-snippet or noindex. Google-Extended is a separate token that controls use of crawled content for training future Gemini models and for grounding answers in Gemini Apps and in Vertex AI's search grounding; Google says it does not affect inclusion in Google Search and is not a ranking signal. Traffic from AI features is reported in the Performance report of Search Console, within the Web search type. (Sources: Google, AI features and your website (updated 10 December 2025), as of 20 October 2026; Google's common crawlers (last updated 15 October 2026), as of 20 October 2026.)

What does OpenAI say?

OpenAI documents two crawler settings that are independent of each other: OAI-SearchBot, which surfaces sites in ChatGPT's search features, and GPTBot, which crawls content that may be used to train its models. For a site that opts out of OAI-SearchBot, OpenAI says ChatGPT search answers will not show it, though it can still appear as a navigational link; if both are allowed, OpenAI may use one crawl for both purposes; and changes to robots.txt can take about a day to take effect for search. ChatGPT-User acts on a person's

request, robots.txt rules may not apply to it, and it is not used to decide whether content appears in Search. (Source: OpenAI, Overview of OpenAI crawlers (no date shown), as of 20 October 2026.)

OpenAI's help page says ChatGPT ranks search results using multiple factors intended to help people find relevant, reliable information, and that placement is not assured. To be eligible a site must allow OAI-SearchBot and allow traffic from OpenAI's published IP addresses. The page also says that ChatGPT search sometimes partners with other search providers, that it may rewrite a question into several queries, and that results and citations can be incomplete, outdated or incorrect. It does not list the factors. (Source: OpenAI Help Center, Searching the web with ChatGPT (page showed an update on the day it was read), as of 20 October 2026.)

What does Anthropic say?

Anthropic documents three robots. ClaudeBot collects content that could contribute to training. Claude-SearchBot is the search robot: disabling it stops indexing for search, which Anthropic says may reduce a site's visibility in what users see when they search. Claude-User retrieves pages when a person asks a question, and disabling it may reduce visibility for user-directed web search. Anthropic says its robots honor robots.txt directives and support the Crawl-delay extension, and that blocking by IP address may not work as an opt-out. (Source: Anthropic Help Center, Does Anthropic crawl data from the web (dated April 8, 2026), as of 20 October 2026.)

Anthropic's help page on web search says that Claude invokes a search tool to ground its answers in content from the live web and that every response includes citations. It does not say how the sources behind a citation are chosen. (Source: Anthropic Help Center, Enable and use web search (shown as updated over a week before it was read), as of 20 October 2026.)

What does Perplexity say?

Perplexity documents PerplexityBot, which is designed to surface and link websites in its search results and is not used to crawl content for AI foundation models, and Perplexity-User, which supports user actions: when a person asks a question it might visit a page and include a link to it in the answer. Because a person requested the fetch, Perplexity says this fetcher generally ignores robots.txt rules. Changes can take up to a day to be reflected. (Source: Perplexity, Perplexity Crawlers (no date shown), as of 20 October 2026.)

Perplexity's help page says each answer includes numbered citations linking to the original sources, and describes the process in general terms: it searches the web, gathers information from sources it calls authoritative and compiles the most relevant insights. It gives no criteria for what it treats as authoritative or relevant. (Source: Perplexity Help Center, How does Perplexity work (last modified August 28, 2026), as of 20 October 2026.)

EXHIBIT 1

Crawler names and what a robots.txt rule for each controls, as of 20 October 2026

Provider	Crawler name	What the provider says it controls	Source and date
Google	Googlebot	Search, including its AI features, and other Google products	Google's common crawlers, last updated 15 October 2026; as of 20 October 2026
Google	Google-Extended	Training future Gemini models and grounding in Gemini Apps and Vertex AI; not Search inclusion	Google's common crawlers, last updated 15 October 2026; as of 20 October 2026
OpenAI	OAI-SearchBot	Whether a site is shown in ChatGPT search answers	Overview of OpenAI crawlers, no date shown; as of 20 October 2026
OpenAI	GPTBot	Whether content may be used to train OpenAI's models	Overview of OpenAI crawlers, no date shown; as of 20 October 2026
OpenAI	ChatGPT-User	User-requested visits; robots.txt rules may not apply	Overview of OpenAI crawlers, no date shown; as of 20 October 2026
Anthropic	Claude-SearchBot	Indexing for search results	Anthropic Help Center, dated April 8, 2026; as of 20 October 2026
Anthropic	Claude-User	User-directed retrieval	Anthropic Help Center, dated April 8, 2026; as of 20 October 2026
Anthropic	ClaudeBot	Collection of content that could contribute to training	Anthropic Help Center, dated April 8, 2026; as of 20 October 2026
Perplexity	PerplexityBot	Surfacing and linking sites in Perplexity search	Perplexity Crawlers, no date shown; as of 20 October 2026
Perplexity	Perplexity-User	User-requested visits; generally ignores robots.txt	Perplexity Crawlers, no date shown; as of 20 October 2026

The names a robots.txt file must address. Search and training have separate names, and the user-requested fetchers are treated differently from the crawlers.

What do the providers leave unsaid?

In the pages read, none of the four providers gives criteria, weights or signals for choosing one eligible page over another. Google comes closest, stating an aim of a broader and more varied group of links. OpenAI says only that multiple factors are used, and Anthropic and Perplexity describe a process without criteria. The documents state who can be used and what a site owner can control; they do not state how a source is selected, and they do not describe how any provider decides which of several suitable pages to cite. Exhibit 2 sets this out. (Sources: Google, AI features and your website (updated 10 December 2025); OpenAI Help Center, Searching the web with ChatGPT (page showed an update on the day it was read); Anthropic Help Center, Enable and use web search (shown as updated over a week before it was read); Perplexity Help Center, How does Perplexity work (last modified August 28, 2026); each as of 20 October 2026.)

EXHIBIT 2

What is documented and what is not, as of 20 October 2026

Question	What is documented	What is not documented	Source and date
Who can be used as a source?	Google: indexed and eligible for a snippet. OpenAI, Anthropic and Perplexity: pages their search or retrieval crawlers can reach	Whether other unpublished conditions apply	Google, AI features, updated 10 December 2025; OpenAI, Anthropic and Perplexity crawler pages; as of 20 October 2026
How is one eligible page chosen over another?	Google: an aim of a broader and more varied group of links. OpenAI: multiple factors	The criteria, weights and signals, for every provider read	Same pages and the OpenAI and Perplexity help pages; as of 20 October 2026
Can a site owner control it?	Yes, with robots.txt rules per crawler; for Google also snippet controls and noindex	How fast a change takes effect in every product; OpenAI and Perplexity say about a day for robots.txt	Google, AI features; OpenAI and Perplexity crawler pages; as of 20 October 2026
Is markup or a special file needed?	Google: no special markup or AI text file	Whether any other provider requires one; none of the pages read says so	Google, AI features, updated 10 December 2025; as of 20 October 2026
Is it measured for the site owner?	Google: traffic from AI features is in Search Console, within the Web search type	A comparable report from the other three providers, in the pages read	Google, AI features, updated 10 December 2025; as of 20 October 2026

Every row ends in the same place: access is documented, selection is not.

The documents state who can be used and what a site owner can control; they do not state how a source is selected.

(Sources: Google, AI features and your website (updated 10 December 2025); OpenAI Help Center, Searching the web with ChatGPT (page showed an update on the day it was read); Anthropic Help Center, Enable and use web search (shown as updated over a week before it was read); Perplexity Help Center, How does Perplexity work (last modified August 28, 2026); each as of 20 October 2026.)

What does this mean for a firm that wants to be cited?

Our reading of the documents is that two things can be acted on. A page that cannot be crawled, or is not eligible, cannot be used, and the conditions for both are published, so the cheapest work is access: read robots.txt for each crawler name in Exhibit 1, check that the host or content delivery network allows the providers' published IP addresses, and check indexing and snippets for key pages. Beyond access, the providers publish no rule a firm can follow to be chosen, so advice that claims to know the rule is not supported by these documents. (Sources: Google, AI features and your website (updated 10 December 2025); OpenAI, Overview of OpenAI crawlers (no date shown); Perplexity, Perplexity Crawlers (no date shown); each as of 20 October 2026.)

A firm can still write clear pages that answer a buyer's question, describe itself consistently, and test its visibility by asking the questions, and it should treat the result as an observation, not a ranking, because

OpenAI's own help page says citations can be incomplete, outdated or incorrect. (Source: OpenAI Help Center, Searching the web with ChatGPT (page showed an update on the day it was read), as of 20 October 2026.)

What should you watch for?

The documents change without notice. As read, Google's AI features page was last updated on 10 December 2025 and its crawler page on 15 October 2026, Anthropic's article is dated April 8, 2026, and Perplexity's help page was last modified on August 28, 2026. OpenAI's and Perplexity's crawler pages show no date. The address of Google's crawler page and of Perplexity's crawler page each redirected to a new address when opened, so links to them may move. Watch also for changes to the products that the documents do not describe, and for changes to what Search Console or other reports show about AI features. (Sources: Google, AI features and your website (updated 10 December 2025); Google's common crawlers (last updated 15 October 2026); Anthropic Help Center, Does Anthropic crawl data from the web (dated April 8, 2026); Perplexity Help Center, How does Perplexity work (last modified August 28, 2026); Perplexity, Perplexity Crawlers (no date shown); OpenAI, Overview of OpenAI crawlers (no date shown); each as of 20 October 2026.)

Method, as-of date and limits

This briefing describes the position on 20 October 2026. Each page was opened that day in a web browser and read in full: Google's AI features page and common crawlers page, OpenAI's crawler overview and its help page on searching the web with ChatGPT, Anthropic's article on its crawlers and its help page on web search, and Perplexity's crawler page and its help page on how it works. Only what a page states is recorded. No assistant was tested for this briefing. The next quarterly review will move the date.

- Four providers are covered. Others, such as Microsoft, are not.
- The pages are the providers' own descriptions. They may omit how the products really behave.
- Where a page shows no date, it may have changed since it was read.
- A statement that a page does not say something means only that we did not find it in the pages listed.
- Nothing here predicts which firm any assistant will name.

Sources

Each of these pages was read on 20 October 2026.

1. [Google, the page AI features and your website, updated 10 December 2025: eligibility, query fan-out, crawler controls and reporting of traffic](#)
2. [Google Crawling Infrastructure, the page on Google's common crawlers, last updated 15 October 2026: Googlebot and Google-Extended](#)
3. [OpenAI, the crawler overview, no date shown: OAI-SearchBot, GPTBot and ChatGPT-User](#)
4. [OpenAI Help Center, the article on searching the web with ChatGPT; the update was shown relative to the day it was read: ranking factors, eligibility and caveats](#)

5. [Anthropic, the article on whether Anthropic crawls data from the web and how site owners can block the crawler, dated April 8, 2026: ClaudeBot, Claude-SearchBot and Claude-User](#)
6. [Anthropic Help Center, the article on enabling and using web search; shown as updated over a week before it was read: citations in every response](#)
7. [Perplexity, the crawlers page, no date shown: PerplexityBot and Perplexity-User](#)
8. [Perplexity Help Center, the article on how Perplexity works, last modified August 28, 2026: numbered citations and how it gathers sources](#)

USE THIS

[Assess your AI search visibility](#)

[Read the white paper on answer engines and buyers](#)

Read your robots.txt file against the four providers' crawler names this week.

This document was made from the article at

<https://xonik.com/insights/market-intelligence/how-answer-engines-choose-sources/>. The online article is the current version.