

The AI Search Field Reference

The full mechanics behind the Signal Note “How AI Search Actually Works.” Every crawler, every control, and the test you can run this week.

-
- 01 The five checks
 - 02 The words, in plain English
 - 03 The crawler directory
 - 04 The robots.txt cheat sheet
 - 05 The baseline test
 - 06 What the research says
 - 07 What to hand your web person
-

Jim Haney

AI advisor and fractional CMO · haneystrategy.com

Every crawler, control, and figure in this document was verified against the publisher's own documentation on August 26, 2026. Where studies disagree, the disagreement is shown rather than resolved.

01 The five checks

Work them in order. Each one is worth very little until the one above it is true. Print this page, fill in the last column honestly, and you have a real starting position instead of a feeling.

THE QUESTION	WHAT A PASS LOOKS LIKE	THE TELL IT IS BROKEN	YOUR READ
01 Can they read you?	Every page a buyer needs is reachable, indexed, allowed to show a snippet, and readable as text without running JavaScript.	You rank respectably on Google, no AI engine ever names you, and nobody at your company can tell you what your robots.txt file says.	
02 Do they know who and where you are?	Every engine can tell that you are one company, in one set of locations, selling one clear set of things. Your website, listings, and profiles agree.	Ask three AI engines what your company does and where it operates, and you get three meaningfully different answers.	
03 Do you answer the question?	Real material sits under several branches of the question at once, with answers under headings written in your buyer's words.	Your best pages are built around two or three phrases somebody picked years ago, and nobody has written down what buyers actually ask on a first call.	
04 Does the world vouch for you?	Reviews, trade press, directories, association listings, and earned coverage corroborate the claims on your own site.	Search your company name and everything on the first two pages is something you published or paid for.	
05 Can you prove it is working?	A dated, logged-out baseline across four engines, the same prompts re-run on a schedule, and a line back to pipeline.	Your last report on this was a screenshot of a chat window, taken while signed in.	

PASS · NEEDS WORK · NO IDEA. ALL THREE ARE USEFUL ANSWERS.

02 The words, in plain English

Every term you will hear in a conversation about AI search, defined the way I would define it on a call. Nothing here requires a technical background.

Crawler

An automated program that visits web pages and reads them. Also called a bot or a spider. Every search engine and every AI company runs several.

Crawling

The visit. A crawler arrives at your page and reads what is there. Crawling alone does nothing for you.

Indexing

The filing. The engine decides a page is worth keeping and puts it in the library it searches later. A page that was never indexed cannot be quoted.

robots.txt

A plain text file at `yourcompany.com/robots.txt` listing which automated visitors are welcome. A request, not a lock. Most declared crawlers honor it, but compliance has been publicly contested, and a real block lives on the server rather than in a text file.

User agent

The name a crawler announces itself by, and the name you use in `robots.txt` to allow or block it. GPTBot and Googlebot are user agents.

Snippet

The short preview of page text shown under a search result. Google requires snippet eligibility before a page can be cited in AI Overviews or AI Mode.

nosnippet / max-snippet

Page-level settings that suppress or shorten the snippet. Often switched on by a plugin default nobody has looked at. They make a page ineligible to be quoted while leaving it able to rank.

noindex

A page-level instruction telling engines not to file the page at all. Correct for thank-you pages and internal utilities, expensive anywhere else.

Structured data (schema)

A small block of code stating plainly what a page is about: this is a company, this is what it sells, this is where it operates. It helps engines understand, but Google says it is not required to appear in AI features.

Entity

The engine's single mental index card about your business. One company, one place, one set of services. Conflicting information across the web makes the card unreliable.

Query fan-out

Google's term for how AI Mode works: it breaks one question into subtopics and runs many searches at once, then assembles a single answer from what comes back.

Passage retrieval

Engines pull the paragraph that answers the question, not the whole page. A good answer buried under a vague heading is worth less than the same answer under a clear one.

AI Overview

The written answer Google places above the traditional results, with a handful of supporting links.

AI Mode

Google's fully conversational search experience, built on the same index as Google Search.

Answer engine

Any system that answers a question in sentences instead of returning a list. ChatGPT, Claude, Perplexity, Google's AI Mode, Microsoft Copilot.

AEO, GEO, LLMO

Three names for the same young discipline: answer engine optimization, generative engine optimization, large language model optimization. No standard has settled. Treat the acronym as noise and the work as real.

llms.txt

A proposed file some vendors pitch as required for AI visibility. Google states in its own documentation that no AI text file is needed to appear in its AI features.

Zero-click search

A search that ends without the person clicking any result. Roughly two thirds of US Google searches now end this way.

Google Search Console

Google's free tool showing which queries bring you traffic and how you are indexed. The authoritative source for your indexation state, but it does not break out AI Overviews separately and knows nothing about ChatGPT, Claude, or Perplexity.

03 The crawler directory

Every major AI company gives you two separate controls: one governs whether they can train on you, the other governs whether they can find you and quote you in an answer. Blocking the wrong one makes you invisible. Blocking the right one only opts you out of training. Note that these are not always the same kind of thing. OpenAI, Anthropic, and Perplexity use two crawlers. Google and Apple each run one crawler plus a control token that does no crawling at all. Microsoft has no AI crawler to block, and governs Copilot with page-level meta tags instead.

OpenAI

USER AGENT	JOB	WHAT IT DOES	BLOCKING IT MEANS
GPTBot	Training	Improving OpenAI's foundation models.	You opt out of model training. Your eligibility to be cited in ChatGPT search is unaffected.
OAI-SearchBot	Search and citation	Surfacing websites in search results in ChatGPT's search features.	You remove yourself from ChatGPT's search answers. This is the one most people block by accident.
ChatGPT-User	User-initiated	Certain user actions in ChatGPT and Custom GPTs.	Triggered when a person asks for a page, so standard robots.txt rules may not apply.
OAI-AdsBot	Safety review	Validating the safety of pages submitted as ads on ChatGPT.	Relevant only if you advertise on ChatGPT.

Anthropic

USER AGENT	JOB	WHAT IT DOES	BLOCKING IT MEANS
ClaudeBot	Training	Collecting web content that may contribute to model training.	You opt out of model training only.
Claude-SearchBot	Search and citation	Improving the relevance and accuracy of search responses.	You remove yourself from Claude's search answers.
Claude-User	User-initiated	Accessing sites when a person asks Claude a question.	Person-driven fetch rather than a scheduled crawl.

Perplexity

USER AGENT	JOB	WHAT IT DOES	BLOCKING IT MEANS
PerplexityBot	Search and citation	Surfacing and linking websites in Perplexity results. Explicitly not used for foundation model training.	You remove yourself from Perplexity entirely. There is no training tradeoff to weigh here.
Perplexity-User	User-initiated	Visiting a page when a person asks a question that needs it.	Perplexity states these fetches generally disregard robots.txt because a person requested them.

Google

USER AGENT	JOB	WHAT IT DOES	BLOCKING IT MEANS
Googlebot	Search, AI Overviews, AI Mode	The one crawler behind ordinary search results and Google's AI features alike.	You remove yourself from Google search and from every Google AI answer at the same time.
Google-Extended	Training opt-out only	NOT a crawler. Google states it “doesn't have a separate HTTP request user agent string” and that the token is “used in a control capacity.” It governs whether already-crawled content trains Gemini.	You opt out of that training. Google states it does not affect your inclusion in Google Search, and AI Overviews and AI Mode are unaffected because those run on Googlebot.

Apple

USER AGENT	JOB	WHAT IT DOES	BLOCKING IT MEANS
Applebot	Search	Powering Spotlight, Siri, and Safari search.	You remove yourself from Apple's search surfaces.
Applebot-Extended	Training opt-out only	Does not crawl. It signals whether already-crawled content may train Apple's foundation models.	You opt out of training. Apple states pages disallowing it can still appear in search results.

Microsoft

USER AGENT	JOB	WHAT IT DOES	BLOCKING IT MEANS
bingbot	Search	Bing's crawler, which also feeds Copilot's answers.	You remove yourself from Bing and from Copilot.

USER AGENT	JOB	WHAT IT DOES	BLOCKING IT MEANS
No separate AI agent	Page-level controls instead	Microsoft controls Copilot usage with the nocache and noarchive meta directives on the page, not with a robots.txt user agent.	See the cheat sheet: this one is not in your robots.txt at all.

04 The robots.txt cheat sheet

Three files, and the seven traps that put companies in the wrong one. Read yours at yourcompany.com/robots.txt before you change anything.

Open to everything

Where most companies should sit. Nothing stands between you and being found.

```
User-agent: *  
Allow: /  
  
Sitemap: https://yourcompany.com/sitemap.xml
```

Opt out of training, stay quotable

The considered middle. You decline to feed the models while remaining eligible to be recommended by them.

```
# Opt out of model training  
User-agent: GPTBot  
Disallow: /  
  
User-agent: ClaudeBot  
Disallow: /  
  
User-agent: Google-Extended  
Disallow: /  
  
User-agent: Applebot-Extended  
Disallow: /  
  
# Left open on purpose. These are the crawlers  
# that find you and cite you in an answer:  
#   OAI-SearchBot, Claude-SearchBot,  
#   PerplexityBot, Googlebot, Applebot
```

The accidental blackout

Almost never what anyone intended, and more common than you would think.

```
User-agent: *  
Disallow: /
```

Seven traps

- ◆ A blanket `User-agent: *` with `Disallow: /` blocks everything, including Googlebot and every crawler that would have cited you. It is the most expensive two lines in this document.
- ◆ Blocking `GPTBot` does not remove you from ChatGPT's search answers. That is `OAI-SearchBot`, a different line in the same file.
- ◆ Blocking `Google-Extended` does not remove you from AI Overviews or AI Mode. Those run on `Googlebot`.
- ◆ Blocking `Applebot-Extended` does not remove you from Apple search results. Apple says so in its own documentation.
- ◆ Microsoft Copilot is not controlled in robots.txt at all. It uses `nocache` and `noarchive` meta directives on the page, and neither removes you from Bing's ordinary search results.
- ◆ User-initiated agents such as `ChatGPT-User` and `Perplexity-User` may fetch a page regardless of robots.txt, because a person asked for it.
- ◆ robots.txt is a public request, not a lock. Anyone can read yours, and bad actors ignore it. Never use it to hide anything sensitive.

05 The baseline test

The test almost everyone runs first is the wrong one. Typing your own company name into an account you use daily returns a flattering answer that means nothing, because the model remembers what you have asked before. Run it this way instead.

- 01 Get out of your own history, and note it works differently on each engine. ChatGPT has a temporary chat. Perplexity and Google AI Mode run in a private window. Claude has no signed-out mode, so turn memory off in settings and open a new conversation.
- 02 Write down three questions in the words a buyer would use. No company names in them.
- 03 Ask all three questions of ChatGPT, Claude, Perplexity, and Google's AI Mode.
- 04 Record which companies each engine named, in order, and whether you appeared at all.
- 05 Save the answers word for word, with the date. Screenshots are fine. Undated ones are worthless.
- 06 Open Google Search Console for the search side. It will not show AI Overview performance separately, and nothing about the other engines.
- 07 Re-run the identical prompts on a schedule. Changed prompts measure nothing.

Record it here

ENGINE	QUESTION 1	QUESTION 2	QUESTION 3
ChatGPT			
Claude			
Perplexity			
Google AI Mode			

DATE RUN SIGNED OUT ☐ MEMORY OFF ☐ PRIVATE WINDOW ☐

06 What the research says

Five figures worth knowing, each with its source and that source's interest in the answer. Three of the five are published by companies selling into this category, which is stated rather than hidden. They disagree on the size of the gap between ranking and being cited. They have agreed on its direction for a year.

FIGURE	WHAT IT MEASURES	SOURCE AND INTEREST
37.9%	of pages cited in Google's AI Overviews also ranked in the top ten for that question. 31.2% ranked 11 to 100. 31.0% did not rank in the first hundred at all. The top-ten figure was about 76% in the previous study eight months earlier.	Ahrefs, 863,000 keyword SERPs and 4M cited URLs, March 2, 2026. Sells AI-visibility tooling.
16.7%	of AI Overview citations came from top-ten results across nine industries over sixteen months, with most overlap growth coming from positions 21 to 100.	BrightEdge Generative Parser, May 2024 to September 2025. Sells AI-search tooling.
~52%	of AI Overview citations came from pages that do not appear in the top 100 ranked results at all.	Originality.AI, November 18, 2025. Sells AI detection.
68.01%	of US Google searches ended without a click on anything, January through April 2026.	SparkToro analysis of Similarweb's US clickstream panel, June 2026. Independent analyst.
+120%	more clicks for brands cited in Google's AI Overviews, counted per time they appeared. Correlation, not a demonstrated cause: cited brands skew better known, and better known brands earn higher click rates anyway.	Seer Interactive, 5.47M queries, 2026. Agency that sells the service.

One number deliberately absent

How many sub-questions a query fan-out actually runs. Google has never published it, and outside estimates range from eight to more than twenty depending on the question. A number nobody can source is not a number worth planning against.

07 What to hand your web person

Copy this list into an email. Nothing on it requires a budget, a new vendor, or a rebuild. It requires somebody to open the file and look.

- 01 Send me our robots.txt file and tell me which crawlers we block and who decided that. I am not asking anyone to edit it yet: one wrong line removes us from Google.
- 02 Confirm none of our commercial pages carry `noindex`, `nosnippet`, or a short `max-snippet` value.
- 03 Confirm our most persuasive copy exists as real text, not only inside images, sliders, or components that build after the page loads.
- 04 Show me our organization structured data and confirm the company name, address, and description match our listings.
- 05 List every place our company name, address, and phone number appear online, and flag the ones that disagree.
- 06 Confirm we have one location profile per real location, not two.
- 07 Give me our indexed page count from Google Search Console and the list of pages Google chose not to index.
- 08 Show me the twenty questions buyers ask on a first call, and which pages answer them under a heading in the buyer's own words.
- 09 Set a calendar reminder to re-run the logged-out baseline on the same schedule every month.

If the answers come back ugly

That is the good outcome. Most of what this document covers is cheap to fix once somebody has actually looked, and the companies that fix it early are the ones getting named while their competitors are still arguing about whether any of this matters.

If you want to walk through what came back, bring the screenshots. That is a good hour.

JIM HANEY · HANEYSTRATEGY.COM/BOOK