RESEARCH

Every AI crawler user agent, and what blocking each one actually does

Most robots.txt files I open are blocking the wrong half. The agent that trains a model and the agent that puts you in the answer are usually two different names, and two of them ignore robots.txt on purpose.

Written by Ali Kamran
SEO & GEO Consultant · Published August 22, 2026 · 8 min read

Three kinds of agent, and why the difference matters

A line in robots.txt only does something if you know what the name on it is for. There are three jobs happening, and companies use separate names for each one.

Training crawlers collect content that may go into training a model. **Search crawlers** build the index an assistant searches when someone asks a question. **User fetchers** go and get one specific page right now, because a person asked something and the assistant decided to open a link.

Blocking the training crawler is a copyright and control decision. Blocking the search crawler is the one that takes you out of the answers. People mix these up constantly, and

the usual result is a site that has opted out of being recommended while still being perfectly available for training, which is the opposite of what almost everyone wants.

OpenAI

OpenAI documents four agents, and they are genuinely separate.

GPTBot crawls content that may be used in training OpenAI's foundation models. **OAI-SearchBot** exists to surface websites in search results inside ChatGPT's search features, and OpenAI says it is not used for training. **ChatGPT-User** fetches a page when a user's question requires it. **OAI-AdsBot** checks ad landing pages.

The one worth reading twice: OpenAI's own documentation says ChatGPT-User does not follow robots.txt rules, because those actions are initiated by a user. GPTBot and OAI-SearchBot do follow it.

So the common pattern of blocking GPTBot and leaving OAI-SearchBot open is coherent. You are declining to be training data while staying eligible to be recommended. Blocking both is also coherent. Blocking OAI-SearchBot alone, which I have found on client sites more than once, is almost never what anyone meant to do.

Anthropic

Same three way split, different names. **ClaudeBot** collects web content that could contribute to training. **Claude-SearchBot** indexes content to improve the relevance and accuracy of search responses. **Claude-User** handles real time retrieval when someone asks Claude a question.

Anthropic publishes a list of its crawler IP addresses at claude.com/crawling/bots.json, which is the only reliable way to tell a genuine ClaudeBot hit from something wearing its name in your logs. Anyone can put any string in a user agent header. Verifying against a published IP range is the check that survives contact with reality.

Google

Google is the one people get wrong most often, and the mistake costs the most.

Google-Extended is a control over whether content Google crawls may be used to train Gemini models and ground answers in some of Google's other systems. Google states plainly that it does not affect a site's inclusion in Google Search and is not used as a ranking signal.

What it does not do is turn off AI Overviews or AI Mode. Those are served out of Google Search itself. Google's documentation says a page must be indexed and eligible to be shown in Google Search with a snippet in order to appear as a supporting link in AI Overviews or AI Mode, and that there are no additional technical requirements beyond that.

Which means the controls for AI Overviews are the same snippet controls that have existed for years: `nosnippet`, `data-nosnippet`, `max-snippet` and `noindex`. There is no separate switch. If you want out of AI Overviews, you are trading away normal snippet display to get it, and that is the whole decision.

GoogleOther is a general purpose crawler used across product teams for research and development. **Google-CloudVertexBot** crawls sites on request from site owners building Vertex AI agents, so it is not something arriving uninvited.

Perplexity

PerplexityBot exists to surface and link websites in Perplexity's search results, and Perplexity states it is not used to crawl content for AI foundation models. It obeys robots.txt.

Perplexity-User visits a page when a user's question calls for it, and Perplexity's own documentation says this fetcher generally ignores robots.txt rules, on the same reasoning OpenAI gives.

That is now two vendors stating on the record that a robots.txt block will not stop a user triggered fetch. If your reason for blocking is legal or contractual rather than commercial, that distinction matters, because robots.txt is not going to deliver what you think you bought.

Microsoft and Bing

There is no separate Copilot crawler to block. Copilot answers are grounded in the Bing index, which bingbot builds, so the controls are meta directives rather than user agent lines.

Microsoft's stated behaviour: content tagged `NOCACHE` may still appear in a Copilot answer, but only the URL, title and snippet are displayed, and only those elements may be used in training Microsoft's foundation models. Content tagged `NOARCHIVE` is not included in Copilot answers, is not linked to in them, and is not used for training.

Both tags still allow the page to rank normally in Bing search results. This is the cleanest opt out any of the big providers offer, in the sense that it separates the search listing from the AI answer without making you choose between them.

Apple and Meta

Applebot is the crawler behind Siri and Spotlight. **Applebot-Extended** never requests a page at all. It is a robots.txt token that governs whether content Applebot has already collected may be used to train Apple's foundation models. Disallowing it has no effect on whether you show up in Apple's search features.

Meta runs five. **meta-externalagent** crawls for training foundation models and for indexing content directly. **meta-webindexer** supports Meta AI search results. **meta-externalfetcher** fetches individual links at a user's request. **meta-externalads** serves advertising products. **facebookexternalhit** is the old one that generates link previews, and blocking it breaks the preview card every time someone shares you on a Meta platform, which is a real cost people discover late.

You will also see **CCBot**, **Amazonbot** and **Bytespider** in logs. CCBot belongs to Common Crawl, a public archive that third parties have used as training data for years, so blocking it is a decision about a downstream population you cannot enumerate. Public documentation on what the other two feed is thinner, and I would rather say that than guess.

The full table

EVERY AGENT, WHAT IT CONTROLS, AND WHAT BLOCKING IT COSTS

AGENT	JOB	OBEYS ROBOTS.TXT	WHAT BLOCKING IT COSTS YOU
OPENAI			
GPTBot	Training	Yes	Nothing in ChatGPT search. You leave the training set.
OAI-SearchBot	Search	Yes	Eligibility to be surfaced in ChatGPT search. This is the expensive one.
ChatGPT-User	User fetch	No	Nothing, because the block is not honoured.
OAI-AdsBot	Ads	Yes	Landing page checks for ads you are running.
ANTHROPIC			
ClaudeBot	Training	Yes	You leave the training set.
Claude-SearchBot	Search	Yes	Eligibility to be used in Claude's search responses.
Claude-User	User fetch	Yes	Live retrieval when a person asks about you.
GOOGLE			
Googlebot	Search, and AI	Yes	Everything. Do not do this.

AGENT	JOB	OBEYS ROBOTS.TXT	WHAT BLOCKING IT COSTS YOU
	Overviews by extension		
Google-Extended	Gemini training and grounding	Yes	Nothing in Search or AI Overviews. Google says so directly.
GoogleOther	Internal research	Yes	Little that is visible to you.
Google-CloudVertexBot	Vertex AI agents, on request	Yes	Only relevant if you asked for it.
PERPLEXITY			
PerplexityBot	Search	Yes	Eligibility to be surfaced and linked in Perplexity. Not training.
Perplexity-User	User fetch	No	Nothing, because the block is not honoured.
APPLE AND META			
Applebot	Search, Siri, Spotlight	Yes	Apple search surfaces.
Applebot-Extended	Training permission token, not a crawler	Yes	Nothing in search. Your content leaves Apple's training set.
meta-externalagent	Training and indexing	Yes	Meta's training set and some indexing.
meta-webindexer	Meta AI search	Yes	Eligibility inside Meta AI search results.
meta-externalfetcher	User fetch	Yes	Live retrieval of individual links.

AGENT	JOB	OBEYS ROBOTS.TXT	WHAT BLOCKING IT COSTS YOU
facebookexternalhit	Link previews	Yes	Your share cards. Blocking this is almost always an accident.
OTHER			
CCBot	Common Crawl public archive	Yes	An unknown number of downstream users, since the archive is public.
Amazonbot	Alexa and Amazon services	Yes	Documentation on downstream use is thin.
Bytespider	ByteDance	Reported as inconsistent	Documentation on downstream use is thin.

What blocking does not do

Three things worth being clear about, because the sales pitch around AI crawler control tends to skip all of them.

It does not retrieve anything already collected. A block applied today governs future requests. Content gathered last year is gathered.

It does not stop the two user fetchers that say in writing that they do not follow robots.txt. If your motivation is legal, robots.txt is a request, not a fence.

It does not change anything in Google Search. Google-Extended is not an AI Overviews switch, and there is no separate one to find.

If you want to see what your own robots.txt currently says to each of these agents, I built a free checker that reads it back to you in plain language. It does not measure whether you are actually cited anywhere, and I would not trust anything that claims to do both in one click.

FAQ

Does blocking GPTBot remove my site from ChatGPT?

Which AI crawlers ignore robots.txt?

Does Google-Extended turn off AI Overviews?

About this PDF

This is the PDF edition of a page published at ali-kamran.com/research/ai-crawler-user-agents. 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, **fusoziw7097**, 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: OpenAI crawler documentation (developers.openai.com/api/docs/bots) · Anthropic support documentation on web crawling · Perplexity bot documentation (docs.perplexity.ai/guides/bots) · Google Search Central, Google common crawlers and AI features documentation · Bing Webmaster Blog, "Announcing new options for webmasters to control usage of their content in Bing Chat," September 2023 · Apple Support, About Applebot · Meta for Developers, web crawlers documentation.