Scott Tischler

FOUNDER & CHAIRMAN · AIRECOMMEND.AI

TECHNICAL

# What Your Server Logs Reveal: Tracking the AI Crawlers Deciding Whether You Get Cited

By Scott Tischler · October 11, 2026 · 9 min read

If you want to know whether AI engines can actually read your site, stop guessing and read your server logs. They are the single most honest record you have of which AI crawlers are visiting, which pages they fetch, and what happens when they knock — a clean 200, or a 404, a 403, a redirect loop, or a timeout. Before ChatGPT, Gemini, Perplexity or Google's AI answers can ever cite or recommend you, one of their bots has to successfully pull the page. Your logs are the only place that proves it happened. Most people optimizing for AI visibility have never looked, and it shows.

I spend a lot of my time at AIrecommend.ai in exactly this layer — the unglamorous plumbing underneath the strategy decks. And the pattern is consistent: a business worries about whether its content is good enough to get recommended, while its logs quietly show the crawlers getting a 403 on the one page that would have won the answer. This is a fixable problem, and it is invisible unless you look.

## Why the log is the ground truth

Every other tool in the AEO stack is an inference. A rank tracker infers. An "AI visibility" dashboard infers. Even asking ChatGPT a question and watching whether you appear tells you the outcome, not the mechanism. The server log is different: it is a line-by-line record of reality. When GPTBot requests */pricing* at 2:14am and your server hands back a 200 with the full HTML, that is not a model of what happened. That *is* what happened.

That distinction matters because the failures that kill AI visibility are usually mechanical, not creative. A robots.txt rule added during a redesign. A security plugin that treats an unfamiliar user-agent as a threat and returns a 403. A migration that 301-redirects half the site into a chain the crawler gives up on. A CDN challenge page that serves a bot a wall of JavaScript instead of your content. None of these show up in a content audit. All of them show up in the log, in plain text, with a timestamp.

My read, not a guarantee: for most sites, the gap between "we published great answers" and "AI never cites us" is more often a fetching problem than a quality problem. The log is where you find out which one you have.

## Learn the names before you read the lines

You can't read a log you can't decode, and the AI crawler landscape has its own vocabulary. The critical thing to understand is that the major AI companies run **more than one** bot, and they do different jobs. Lumping them together is how people make bad robots.txt decisions.

On the OpenAI side you'll see **GPTBot**, which gathers content that may be used for training; **OAI-SearchBot**, which builds the index behind ChatGPT's search features; and **ChatGPT-User**, which is the fetch that fires when a live user's conversation needs to pull a specific page right now. Anthropic runs **ClaudeBot** and a **Claude-SearchBot** for its search-grounded answers. Perplexity runs **PerplexityBot** for indexing and a live **Perplexity-User** fetcher for in-the-moment requests. Google uses the familiar **Googlebot**, but exposes **Google-Extended** as a separate token you can allow or disallow to control Gemini and AI training independently of regular search.

The useful mental split is *training and indexing crawlers* versus *live-answer fetchers*. The live fetchers — the ones with "User" or "Search" in the name — are the ones pulling your page in the exact moment a customer asks a question an engine thinks you can answer. When those start showing up on a specific URL, that URL has become part of how the engine answers something. That is a signal worth watching.

## What to actually look for

Open the log and ask four questions in order. They go from "are they here at all" to "are they getting what they need."

**One: are the AI crawlers visiting at all?** Filter your log by those user-agent strings over the last 30 days. If you see nothing — no GPTBot, no ClaudeBot, no PerplexityBot — that is your headline problem. Either you are blocking them (check robots.txt and any firewall or CDN bot rules), or your site is new or low-authority enough that they haven't prioritized it yet. Either way, everything else is moot until the bots can get in.

**Two: what status codes are they getting?** This is the highest-value thing in the whole exercise. A page that returns 200 to GPTBot is readable. A page that returns 403 is being actively refused — almost always by a security layer that doesn't recognize the bot. A 404 means the crawler has a URL you've since removed or renamed. A long chain of 301s means a migration you didn't clean up.

Timeouts mean the bot asked and your server took too long to answer. Sort crawler hits by status code and the broken pipes jump right out.

**Three: which URLs are they requesting?** Map the crawled URLs against the pages you actually care about. The quiet failure here is a crawler that happily fetches your blog index and your homepage but has never once requested your best, most citable page — the deep explainer, the original-data piece, the service page with the clean answer. If the bot never asks for it, your content quality is irrelevant. That is a discovery problem: fix your internal linking, your sitemap, and your llms.txt so the good page is reachable.

**Four: how often, and is it changing?** Crawl frequency is a rough proxy for how much an engine cares about you right now. A page getting re-fetched regularly is a page the engine is keeping current. A sudden drop-off after a site change is a warning you want to catch in weeks, not quarters.

## Turning findings into fixes

The log's whole value is that every problem it surfaces has a concrete fix. If the AI crawlers are getting 403s, the culprit is almost always a web application firewall, a bot-management rule, or a security plugin. Allow-list the official user-agents and, where the provider publishes them, their IP ranges — don't rely on the user-agent string alone for anything security-sensitive, since it's trivially spoofable, but for opening the door to legitimate AI crawlers it's the right lever.

If the bots are hitting dead URLs, reconcile your redirects and your sitemap so the crawler is pointed at live, canonical pages. If they're reaching shallow pages but never your deep ones, that is an internal-linking and sitemap job — the best content on the internet does nothing if the only path to it is a search box a crawler won't use. If they're timing out, that's a performance and hosting conversation, and it's one of the few places where page speed genuinely intersects with AI visibility: a bot on a budget will give up on a slow server.

And if, after all that, the logs are clean — crawlers visiting, 200s across your key pages, your best content getting fetched — then you've learned something equally valuable. The plumbing is fine. Your problem, if you still aren't getting cited, lives upstream in authority, corroboration and positioning, and you can stop blaming the pipes and go work on the thing that's actually broken.

## The limits of what logs can tell you

I want to be honest about the ceiling here, because it's easy to over-read a log. Server logs tell you about *fetching*. They do not tell you about *citation*. A crawler can read your page perfectly and the engine can still choose a competitor, because the decision to name someone in an answer is a

separate, murkier process built on authority and relevance and the whole web's corroboration of your claims. The log proves the door is open. It says nothing about whether you got invited in.

There's also attribution ambiguity. A live fetch from ChatGPT-User on your pricing page strongly suggests your page was being pulled into someone's answer, but the log can't tell you what the prompt was or whether you were cited favorably, mentioned in passing, or used and discarded. Treat these as strong hints, not proof.

So the right workflow is a loop. Read the logs to confirm the engines can reach your important pages. Then go ask the engines the questions your customers ask, and watch whether you show up and how. When you don't show up on a prompt you think you should own, come back to the log and check: did the relevant page even get fetched? That loop — logs for the mechanism, prompt testing for the outcome — is how you stop guessing.

## Make it a habit, not a heroic one-time audit

The reason this matters on a schedule is that the failures are introduced by ordinary, well-meaning changes. Someone tightens the firewall. A developer ships a redirect. A new robots.txt rule goes out with a release. A CDN turns on a stricter bot challenge. Any one of these can silently cut a crawler off from your best pages, and you will not feel it for weeks — right up until you notice you've quietly disappeared from answers you used to win.

A light monthly review is enough for most businesses: filter for the AI user-agents, check status codes on your priority URLs, confirm your best pages are still getting fetched, and note any big change in frequency. Fifteen minutes. The teams that do this catch the broken pipe while it's a line in a log file instead of a mystery about why the recommendations dried up. In a world where being the answer is the whole game, knowing for certain that the engines can actually read you is not a nice-to-have. It's the floor everything else stands on.

### Key takeaways

- Before any AI can cite or recommend you, one of its crawlers has to successfully fetch the page — and your raw server logs are the only record that proves it did.
- Learn the real user-agent strings (GPTBot, OAI-SearchBot, ChatGPT-User, ClaudeBot, Claude-SearchBot, PerplexityBot, Perplexity-User, Google-Extended) so you can tell training crawlers from live-answer fetchers.
- A 200 on your key pages and a lot of 404s, 403s, redirects or timeouts on crawler hits is a fixable visibility problem hiding in plain sight.

- Watch which URLs the bots actually request — if they never touch your best page, the problem is discovery, not content.
- Server logs tell you about fetching, not citation; pair them with manual prompt testing to connect what gets crawled to what gets named.
- Set a simple monthly log review so a robots change, a migration, or a CDN rule can't quietly cut you off from the engines.

## Frequently asked questions

### Which AI crawlers should I look for in my logs?

Start with the ones that matter most for visibility today: OpenAI runs GPTBot (training), OAI-SearchBot (its search index) and ChatGPT-User (fetches triggered by a live chat). Anthropic runs ClaudeBot and Claude-SearchBot. Perplexity runs PerplexityBot and a live Perplexity-User fetcher. Google uses Googlebot for its index and Google-Extended as a separate control for Gemini and AI training. Seeing a mix of training crawlers and live fetchers is the healthy pattern. My read, not a guarantee: the live fetchers (the ones with 'User' or 'Search' in the name) are the ones to watch most closely, because they are the ones pulling pages in the moment someone asks a question.

### I don't have access to raw server logs — what can I do?

A lot of people are on a host or CDN that hides the raw access log. Check first whether your CDN (Cloudflare, Fastly, Akamai) exposes a bot or crawler analytics view, because many now break out AI user-agents directly. If you are on a managed platform, look for a 'logs' or 'traffic' export. Failing that, you can approximate it: add server-side logging of the user-agent header for a few weeks, or use a log-based analytics tool. The goal is just to see which bots hit which URLs with what status code — you do not need anything fancy.

### If a crawler is visiting my pages, does that mean AI will cite me?

No, and this is the trap. A successful crawl is necessary but not sufficient. It means the door is open — the engine can read you. Whether it then chooses to cite or recommend you depends on authority, corroboration, relevance to the prompt, and a dozen things logs can't see. Think of logs as the plumbing check: they tell you water can get through, not that anyone is drinking. Always pair log analysis with actually asking the engines questions and watching whether you show up.

**About the author.** Scott Tischler is the Founder & Chairman of AIrecommend.ai and a practitioner-authority on AI search and Answer Engine Optimization. With 20+ years in marketing technology — including American Express, MetLife, and UBS — and executive and professional study at Wharton, Harvard, and Oxford, he helps businesses become the ones AI recommends.

© 2026 Scott Tischler · scotttischler.com
scott@airecommend.ai · AIrecommend.ai