



CiteHawk

[Read your own logs](#)

— THE VISITORS YOU NEVER COUNTED

# The crawler question.

A whole population of AI bots is reading your website right now, and almost nobody has looked at their own logs to see which ones. This paper maps the AI crawler ecosystem, explains why they are not one thing, and gives you a blocking policy that protects your content without deleting your brand from the answers.

## They are not one thing: the five jobs

"Block AI bots" treats a diverse population as one animal, and that mistake has consequences. Every AI-related agent in your logs is doing one of five jobs, and the trade-off for blocking each is different:

| JOB                         | WHAT IT DOES                                                                                                                                                 | WHAT BLOCKING COSTS YOU                                                               |
|-----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| <b>AI model<br/>fetcher</b> | <b>Fetches your page in real time because a user asked a question right now (ChatGPT-User, Claude-User, Perplexity-User, DuckAssistBot, MistralAI-User)</b>  | <b>The answer being composed this second loses your page. Highest immediate cost.</b> |
| AI search<br>crawler        | Retrieves and verifies sources to cite in AI search results (OAI-SearchBot, Claude-SearchBot, PerplexityBot, YouBot, Meta-ExternalFetcher)                   | Your pages fall out of the citable pool.                                              |
| Training<br>crawler         | Collects content for future model training (GPTBot, ClaudeBot, Google-Extended, Applebot-Extended, Amazonbot, Bytespider, CCBot, Meta-ExternalAgent, Cohere) | Long-term representation in future models; no effect on today's answers.              |
| Fetcher /<br>preview        | On-demand page fetches for previews and app features (Anthropic-AI, GoogleOther)                                                                             | Broken previews, degraded app surfaces.                                               |
| Classic<br>index            | Traditional search indexing (Googlebot, Bingbot, Applebot, DuckDuckBot)                                                                                      | Your search rankings, and increasingly the retrieval layer AI search leans on.        |

Agent names above are the published user-agents each company documents publicly. CiteHawk's bot analytics identifies and classifies all of them from your traffic.

## A blocking policy that does not delete you

The honest framing: this is a rights question and a visibility question at the same time, and the right answer depends on what your site is. A publisher monetizing content and a brand trying to be discovered should not run the same policy.

### If you are a brand that wants to be recommended

**Never block AI model fetchers or AI search crawlers.** These are buyers' questions being answered in real time. Blocking them is walking out of the room mid-pitch.

**Think harder about training crawlers.** Allowing them invests in your representation inside future models; blocking them is defensible for proprietary content but costs long-term memory. Most brands should allow.

**Verify, don't assume.** An accidental blanket rule (a CDN bot-protection preset, an old robots.txt line) silently blocks the fetchers you want. This is one of the most common findings in our GEO audits: sites that opted out of AI answers without ever deciding to.

### If you are a publisher

The calculus genuinely differs: training crawlers consume the asset you sell. Blocking training agents while allowing AI model and search fetchers (which drive attribution and referrals) is a coherent middle path, and the per-purpose table on the previous page is exactly the distinction robots.txt lets you draw, because these companies publish separate user-agents for separate jobs.

**"The most expensive robots.txt line on the internet right now is the one that blocks a buyer's AI from reading your pricing page."**

## Read your own logs this week

Three questions your server logs can answer in an afternoon, and what each tells you:

**Which AI agents visit, and how often?** AI model-fetcher traffic is the strongest signal that AI is reading you live on behalf of real users. Zero AI model traffic on a category-relevant site is itself a finding.

**Which pages do they read?** The pages AI fetches most are the pages currently speaking for you in answers. If that list is your careers page and a 2023 blog post, you have found your problem.

**Who gets a 403?** Cross-reference blocked requests against the five-job table. Every blocked AI model fetcher is an answer you were edited out of.

Impersonation is real (anyone can claim a user-agent string), so serious analysis verifies agents against published IP ranges rather than trusting the string alone. Good bot analytics does this for you.

### See every AI bot on your site, classified.

CiteHawk's bot analytics identifies the full AI crawler population in your traffic, splits it by job (AI models, search, training, indexing), and shows exactly which pages each one reads. It ships alongside your free AI visibility report.

Start free → [citehawk.com/signup](https://citehawk.com/signup)

© 2026 ScaleUp Teams Pty Ltd · Agent names per each provider's public bot documentation  
· Reproduction with attribution welcome.