

# We read the robots.txt of 600 startup websites: almost none block AI search crawlers

Elementera, AI Crawler Access Report, October 2026. Web version: <https://www.elementera.com/resources/ai-crawler-access-report>

By [Mojtaba Ebadi](#), Co-Founder and Chief Operating Officer. Reviewed by [Soroush Mehri](#), Co-Founder and Chief AI Officer. Last updated October 13, 2026.

**Key point.** Almost no startups block the crawlers AI search engines use: 3 of the 569 websites we could read block one. Training crawlers are blocked more often, and nearly always on purpose. For most companies, crawler access is not where AI visibility is lost.

## 3 of 569

startup websites block at least one AI search crawler (OAI-SearchBot, Claude-SearchBot or PerplexityBot): 0.5%, 95% interval 0.2% to 1.5%

Elementera AI Crawler Access Report, robots.txt of 600 sampled Y Combinator-backed B2B software, fintech and developer-tool sites, October 7, 2026

We read the robots.txt file of 600 randomly sampled websites of Y Combinator-backed B2B software, fintech and developer-tool companies and checked, crawler by crawler, what each file allows. This is the AI Crawler Access Report: a measured answer to a question every marketing team now asks, "are we blocking ChatGPT?", for the companies most likely to care about the answer.

The short version: they are not. Of the 569 sites whose answer we could read, 467 of 569 (82%; 95% interval 79% to 85%) have no robots.txt rule that names an AI crawler at all, so every AI crawler is allowed by default. The blocks we did find are mostly on training crawlers, and the file usually names them, so someone chose them.

| Interactive on the web version: <https://www.elementera.com/resources/ai-crawler-access-report>

Want the same answer for your own site? Paste your robots.txt into the free checker below. It reads the file with the same rules this study used and shows, engine by engine, which crawlers may read your pages. Nothing you paste leaves the page.

## Key takeaways

- **AI search crawlers are almost never blocked.** 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%) block at least one of OAI-SearchBot, Claude-SearchBot and PerplexityBot.
- **Training crawlers are blocked more often, and on purpose.** 18 of 569 sites (3.2%; 95% interval 2.0% to 4.9%) block at least one AI training crawler, and in 15 of those 18 the file names the crawler.

- **Nobody blocks search while allowing training.** 0 of 569 sites (0%; 95% interval 0% to 0.7%) do the reverse of the split most site owners say they want.
- **Most files say nothing about AI at all.** 467 of 569 sites (82%; 95% interval 79% to 85%) name no AI crawler, so every AI crawler is allowed by default.
- **A newer kind of line is still rare.** 26 of 569 sites (4.6%; 95% interval 3.1% to 6.6%) carry a Content-Signal line in robots.txt.
- **Access is rarely the problem.** For most companies, AI visibility is won or lost after the crawler gets in: in whether the engine reads, trusts, describes and recommends the page.

Interactive on the web version: <https://www.elementera.com/resources/ai-crawler-access-report>

## How to cite this report

Journalists, researchers and anyone writing about AI crawlers are welcome to quote these figures. Please cite the report and link to this page, so readers can check the method and the sample size behind each number.

Suggested citation:

Elementera (2026). The AI Crawler Access Report: robots.txt rules for AI crawlers on 600 startup websites. Ebadi, M. (author), Mehri, S. (reviewer). Data collected October 7, 2026.  
<https://www.elementera.com/resources/ai-crawler-access-report>

The headline figure, ready to quote: "3 of 569 startup websites (0.5%; 95% interval 0.2% to 1.5%) block at least one AI search crawler (Elementera AI Crawler Access Report, October 2026)." The full per-site data is on [the dataset page](#), and the CSV can be [downloaded directly](#).

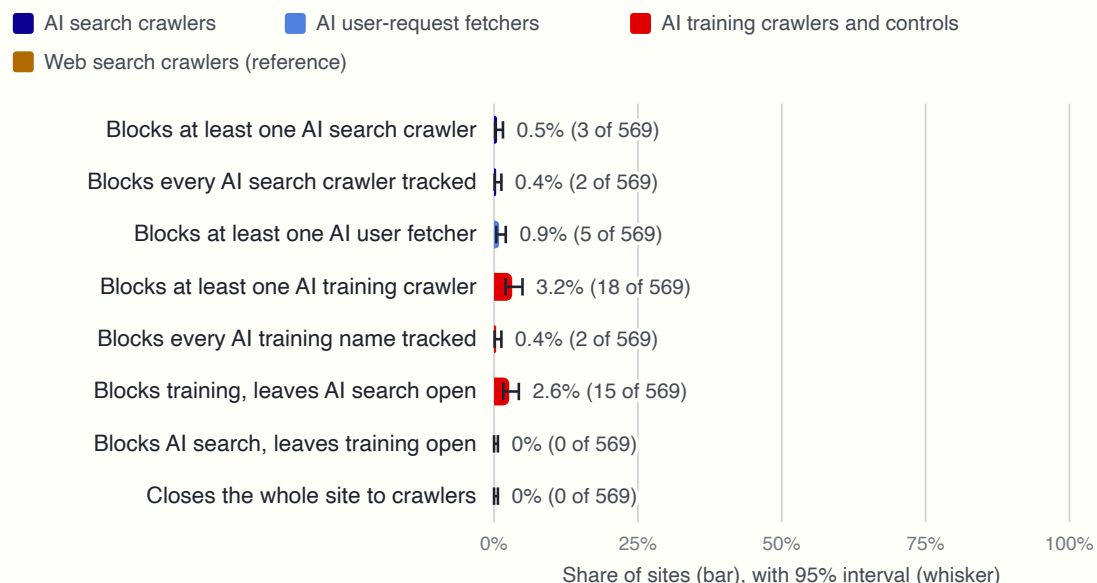
Data licence: [CC BY 4.0](#): free to reuse, including commercially, with credit to Elementera and a link to this page.

## Do startups block the AI crawlers that decide whether ChatGPT can cite them?

Almost none do: of the sites we could read, 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%) block at least one of OAI-SearchBot, Claude-SearchBot and PerplexityBot on at least one of six common paths, and 2 of 569 sites (0.4%; 95% interval 0.1% to 1.3%) block all three. For comparison, 18 of 569 sites (3.2%; 95% interval 2.0% to 4.9%) block at least one AI training crawler. Sites block AI training crawlers more often; the two intervals do not overlap.

## AI search access versus AI training access

n = 569 sites with a readable answer; robots.txt fetched 2026-10-07



Three more site-level figures complete the picture:

- 0 of 569 sites (0%; 95% interval 0% to 0.7%) block an AI search crawler while leaving every AI training crawler open.
- 5 of 569 sites (0.9%; 95% interval 0.4% to 2.0%) block at least one of the fetchers that visit a page when a person asks an assistant to read it (ChatGPT-User, Claude-User, Perplexity-User, Meta-ExternalFetcher).
- 0 of 569 sites (0%; 95% interval 0% to 0.7%) close the whole site to crawlers, Googlebot and Bingbot included.

## What is the difference between an AI search crawler and an AI training crawler?

An AI search crawler fetches pages so an answer engine can find and cite them; an AI training crawler collects pages, or controls whether pages may be used, to train models. Each vendor names them separately in robots.txt, so blocking one does not block the other.

| VENDOR             | AI SEARCH CRAWLER                                 | FETCHES ON A USER'S REQUEST | AI TRAINING CRAWLER OR CONTROL               |
|--------------------|---------------------------------------------------|-----------------------------|----------------------------------------------|
| OpenAI (ChatGPT)   | OAI-SearchBot                                     | ChatGPT-User                | GPTBot                                       |
| Anthropic (Claude) | Claude-SearchBot                                  | Claude-User                 | ClaudeBot                                    |
| Perplexity         | PerplexityBot                                     | Perplexity-User             | None documented                              |
| Google             | Googlebot (Google Search, including AI Overviews) | Not covered here            | Google-Extended, a control name              |
| Apple              | Applebot                                          | Not covered here            | Applebot-Extended, a control name            |
| Meta               | Not covered here                                  | Meta-ExternalFetcher        | Meta-ExternalAgent                           |
| Others we counted  |                                                   |                             | CCBot (Common Crawl), Bytespider (ByteDance) |

The vendors say this in their own words:

- **OpenAI:** "Sites that are opted out of OAI-SearchBot will not be shown in ChatGPT search answers, though can still appear as navigational links." And: "Disallowing GPTBot indicates a site's content should not be used in training generative AI foundation models." ([OpenAI crawler documentation](#))
- **Anthropic:** "Disabling Claude-SearchBot on your site prevents our system from indexing your content for search optimization, which may reduce your site's visibility and accuracy in user search results." ([Anthropic crawler documentation](#))
- **Perplexity:** PerplexityBot "is designed to surface and link websites in search results on Perplexity. It is not used to crawl content for AI foundation models." ([Perplexity crawler documentation](#))
- **Google:** "Google-Extended does not impact a site's inclusion in Google Search nor is it used as a ranking signal in Google Search." ([Google crawler documentation](#))

For a software company the practical reading is simple. A buyer who asks ChatGPT, Claude or Perplexity to compare vendors gets an answer built from pages those engines could fetch. A blocked search crawler takes your pages out of that set, whatever you decided about training. Microsoft Copilot and Google's AI features draw on the Bing and Google indexes, so for them Bingbot and Googlebot are the crawlers that matter.

Outside research points the same way for Google. A 2026 study of Google Search, Gemini and AI Overviews found that sites blocking Google's AI crawler were less likely to be retrieved by AI Overviews (Grossman et al., ACM SIGIR 2026, [arXiv 2604.27790](#)). The authors measured a correlation, not a cause, and Google states that Google-Extended does not affect inclusion in Google Search.

## Which AI crawlers are blocked most often?

Bytespider has the highest share of blocks among the crawlers we tracked: 15 of 569 sites (2.6%; 95% interval 1.6% to 4.3%). Its interval overlaps those of 9 other crawlers, so the order among them is not settled at this sample size. The table gives every crawler we tracked. "Blocked" means the file disallows

the crawler on at least one of six common paths (/, /pricing, /blog/, /docs/, /about, /product); a rule that only keeps crawlers out of /admin/ or /api/ does not count.

| CRAWLER              | TYPE                   | BLOCKED ON AT LEAST ONE PATH                      | BLOCKED ON ALL SIX PATHS                          | NAMED IN THE FILE                                 |
|----------------------|------------------------|---------------------------------------------------|---------------------------------------------------|---------------------------------------------------|
| OAI-SearchBot        | AI search              | 2 of 569 sites (0.4%; 95% interval 0.1% to 1.3%)  | 0 of 569 sites (0%; 95% interval 0% to 0.7%)      | 81 of 569 sites (14%; 95% interval 12% to 17%)    |
| Claude-SearchBot     | AI search              | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 0 of 569 sites (0%; 95% interval 0% to 0.7%)      | 42 of 569 sites (7.4%; 95% interval 5.5% to 9.8%) |
| PerplexityBot        | AI search              | 2 of 569 sites (0.4%; 95% interval 0.1% to 1.3%)  | 0 of 569 sites (0%; 95% interval 0% to 0.7%)      | 93 of 569 sites (16%; 95% interval 14% to 20%)    |
| ChatGPT-User         | AI user fetch          | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 1 of 569 sites (0.2%; 95% interval 0.0% to 1.0%)  | 93 of 569 sites (16%; 95% interval 14% to 20%)    |
| Claude-User          | AI user fetch          | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 0 of 569 sites (0%; 95% interval 0% to 0.7%)      | 42 of 569 sites (7.4%; 95% interval 5.5% to 9.8%) |
| Perplexity-User      | AI user fetch          | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 0 of 569 sites (0%; 95% interval 0% to 0.7%)      | 43 of 569 sites (7.6%; 95% interval 5.7% to 10%)  |
| Meta-ExternalFetcher | AI user fetch          | 4 of 569 sites (0.7%; 95% interval 0.3% to 1.8%)  | 1 of 569 sites (0.2%; 95% interval 0.0% to 1.0%)  | 8 of 569 sites (1.4%; 95% interval 0.7% to 2.8%)  |
| GPTBot               | AI training            | 11 of 569 sites (1.9%; 95% interval 1.1% to 3.4%) | 9 of 569 sites (1.6%; 95% interval 0.8% to 3.0%)  | 98 of 569 sites (17%; 95% interval 14% to 21%)    |
| ClaudeBot            | AI training            | 7 of 569 sites (1.2%; 95% interval 0.6% to 2.5%)  | 5 of 569 sites (0.9%; 95% interval 0.4% to 2.0%)  | 86 of 569 sites (15%; 95% interval 12% to 18%)    |
| Google-Extended      | AI training            | 5 of 569 sites (0.9%; 95% interval 0.4% to 2.0%)  | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 85 of 569 sites (15%; 95% interval 12% to 18%)    |
| CCBot                | AI training            | 13 of 569 sites (2.3%; 95% interval 1.3% to 3.9%) | 10 of 569 sites (1.8%; 95% interval 1.0% to 3.2%) | 61 of 569 sites (11%; 95% interval 8.4% to 14%)   |
| Applebot-Extended    | AI training            | 5 of 569 sites (0.9%; 95% interval 0.4% to 2.0%)  | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 62 of 569 sites (11%; 95% interval 8.6% to 14%)   |
| Meta-ExternalAgent   | AI training            | 7 of 569 sites (1.2%; 95% interval 0.6% to 2.5%)  | 4 of 569 sites (0.7%; 95% interval 0.3% to 1.8%)  | 40 of 569 sites (7.0%; 95% interval 5.2% to 9.4%) |
| Bytespider           | AI training            | 15 of 569 sites (2.6%; 95% interval 1.6% to 4.3%) | 12 of 569 sites (2.1%; 95% interval 1.2% to 3.6%) | 38 of 569 sites (6.7%; 95% interval 4.9% to 9.0%) |
| Googlebot            | Web search (reference) | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 0 of 569 sites (0%; 95% interval 0% to 0.7%)      | 35 of 569 sites (6.2%; 95% interval 4.5% to 8.4%) |
| Bingbot              | Web search (reference) | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 0 of 569 sites (0%; 95% interval 0% to 0.7%)      | 50 of 569 sites (8.8%; 95% interval 6.7% to 11%)  |
| Applebot             | Web search (reference) | 3 of 569 sites (0.5%; 95% interval 0.2% to 1.5%)  | 0 of 569 sites (0%; 95% interval 0% to 0.7%)      | 18 of 569 sites (3.2%; 95% interval 2.0% to 4.9%) |

# Which crawlers do sites block?

n = 569 sites with a readable answer; robots.txt fetched 2026-10-07



One parsing detail affects Apple's names. Our checker, like many parsers, lets a group written for Applebot also govern Applebot-Extended when no group names Applebot-Extended directly. In our sample that rule decided the Applebot-Extended result on 5 of 569 sites.

## Are these blocks deliberate or accidental?

Training blocks are almost always deliberate: the file names the crawler. We sorted every block into one of five kinds, using rules we wrote down before we read any of the files:

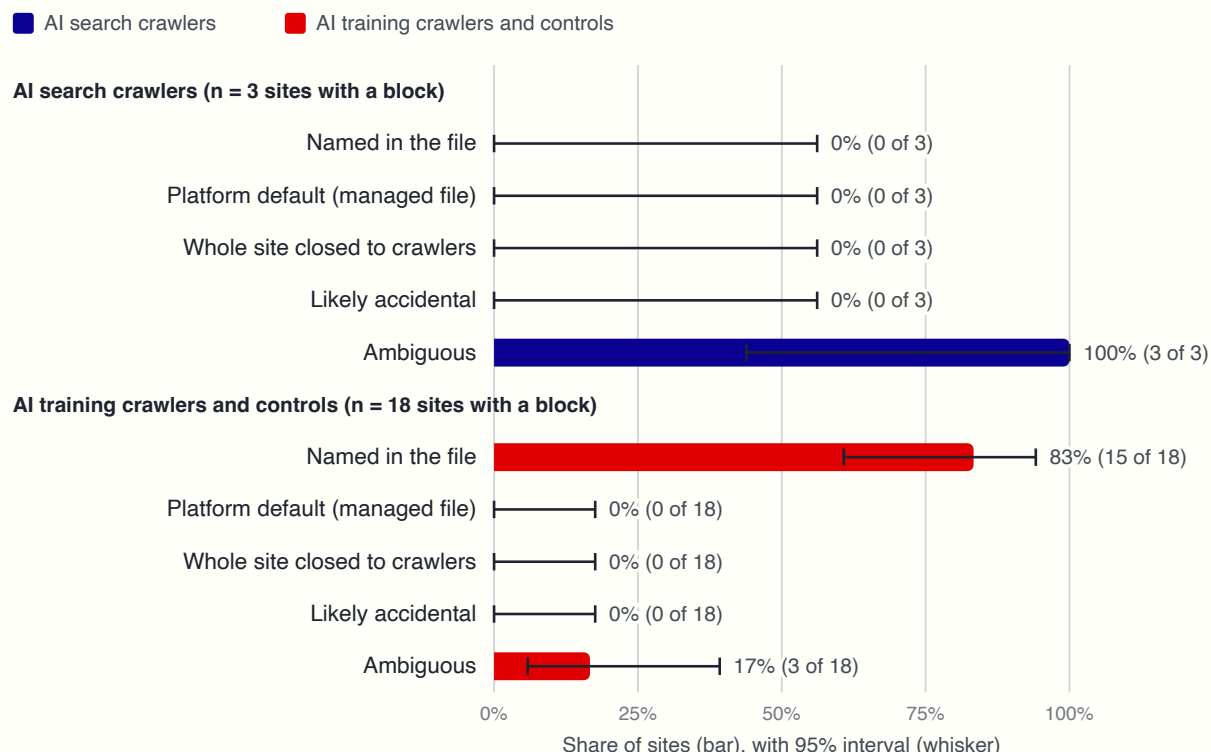
1. **Named in the file:** a group names the crawler. Someone typed its name, or copied a list that contains it.
2. **Platform default:** the rule sits inside a block of rules that a hosting or CDN platform writes for the site, such as Cloudflare's managed robots.txt.
3. **Whole site closed:** a catch-all User-agent: \* rule blocks everything, Googlebot and Bingbot included.
4. **Likely accidental:** the file blocks an AI search crawler while saying it wants to be in search (a Content-Signal line with search=yes), or it names AI training crawlers in their own group and the catch-all rule also catches the search crawlers.
5. **Ambiguous:** everything else, mostly catch-all rules that block AI crawlers while a named group lets Googlebot in. That can be an intended allowlist or an old file written before AI crawlers existed.

Among the 18 sites that block at least one AI training crawler, 15 of 18 sites (83%; 95% interval 61% to 94%) name it, and 0 of 18 sites (0%; 95% interval 0% to 18%) carry the block in a platform-managed section. Among the 3 sites that block at least one AI search crawler, 0 of 3 sites (0%; 95% interval 0% to 56%) show a block that looks accidental, and all 3 are in the ambiguous group.

| KIND OF BLOCK     | AI SEARCH                                     | AI TRAINING                                   |
|-------------------|-----------------------------------------------|-----------------------------------------------|
| Named in the file | 0 of 3 sites (0%; 95% interval 0% to 56%)     | 15 of 18 sites (83%; 95% interval 61% to 94%) |
| Platform default  | 0 of 3 sites (0%; 95% interval 0% to 56%)     | 0 of 18 sites (0%; 95% interval 0% to 18%)    |
| Whole site closed | 0 of 3 sites (0%; 95% interval 0% to 56%)     | 0 of 18 sites (0%; 95% interval 0% to 18%)    |
| Likely accidental | 0 of 3 sites (0%; 95% interval 0% to 56%)     | 0 of 18 sites (0%; 95% interval 0% to 18%)    |
| Ambiguous         | 3 of 3 sites (100%; 95% interval 44% to 100%) | 3 of 18 sites (17%; 95% interval 5.8% to 39%) |

## How the blocks look: deliberate, default or accidental?

Among sites with at least one block in the group; a site can show more than one kind



A site can show more than one kind (for example a named training block and an ambiguous search block), so the shares in each group can add up to more than 100%.

## Are sites starting to use Content-Signal?

Very few sites use Content-Signal yet, and almost none use it to refuse AI input. In our words, as cleared for publication: "A newer kind of line is still rare. 26 of 569 sites (4.6%; 95% interval 3.1% to 6.6%) carry a Content-Signal line in robots.txt, which states preferences such as search, ai-input and ai-train, and 2 of 569 (0.4%; 95% interval 0.1% to 1.3%) say no to AI input. None of the five AI vendor crawler pages we keep dated copies of mentions this line."

This is a count we made after the main analysis, from the same files. A Content-Signal line states a preference; it is not a rule any of the vendors above document that their crawlers read.

## Why do other figures about AI crawler blocking look so different?

Other figures look different because they count different things. Figures you may have seen that say a large share of websites block AI crawlers usually measure training crawlers, on large publisher and high-traffic sites, or count any technical barrier a crawler meets. This report measures explicit robots.txt rules for AI search crawlers on startup websites. Both kinds of number can be right at once; they answer different questions.

Four differences explain most of the gap:

1. **Which crawler.** Blocking GPTBot or CCBot keeps content out of model training; it does not take a site out of ChatGPT search, which OAI-SearchBot decides. In our sample, training blocks were far more common than search blocks (18 of 569 against 3 of 569), and the two intervals do not overlap. A figure about "blocking OpenAI" is very often a figure about GPTBot.
2. **Which sites.** News publishers and the most-visited websites protect their content far more than young software companies do. A 2024 audit of about 14,000 web domains that feed AI training datasets found that more than 28% of the most actively maintained, critical sources in one widely used dataset (C4) had become fully restricted (Longpre and colleagues, "Consent in Crisis", arXiv preprint, [arXiv 2407.14933](https://arxiv.org/abs/2407.14933)). That is a different population, a different year and a training question.
3. **Which mechanism.** Some studies count any technical barrier: bot protection, firewall rules, JavaScript-only pages, slow or failing responses. Those can stop a crawler without a word in robots.txt, and they cannot be seen from outside a site. This report reads robots.txt, the one place a company states its rules in public.
4. **Which denominator.** A share of websites, a share of web traffic and a share of training data are three different numbers. A few very large sites can hold a large share of traffic or text while being a tiny share of websites.

So when you see two numbers that seem to disagree, ask what was counted, on which sites, and by which mechanism. For the question a startup marketer actually has, "do companies like mine block the crawlers that decide AI search citations?", this report is the direct measurement.

## What does this mean for your AI visibility?

For most companies, crawler access is the step that already works. If your robots.txt names no AI crawler, every AI search crawler is allowed by default, and the reasons an engine does not cite you sit further down the AI search visibility funnel: whether it reads your page for the buyer's question, trusts it, describes you correctly and recommends you.

That makes a quick check worth more than a rewrite. Run your robots.txt through the checker above, confirm your CDN and firewall let the search crawlers through, and then put your effort where the losses actually happen. Our [llms.txt guide](#) covers the other file AI agents read, and what it does and does not do.

## How can I check my own site?

Use the free checker above, or the full version at </tools/ai-crawler-check>. It reads your robots.txt the same way this study did and tells you, engine by engine, which crawler each rule affects. The check applies to any business with a website, not only software companies.

1. **Check every host you own:** www, the bare domain, and any docs, blog, help or app subdomain.
2. **Find the group each AI search crawler falls into:** OAI-SearchBot, Claude-SearchBot, PerplexityBot, plus Googlebot and Bingbot. If a group names the crawler, only that group counts.

3. **Put training crawlers in their own group**, with a comment that records what you decided and why.
4. **Check your CDN and firewall** for AI bot or bot management settings, and confirm the search crawlers are allowed there too.
5. **Make sure robots.txt answers**. A server error at `/robots.txt` tells standards-following crawlers to stay away from the whole site.

## How we ran this study

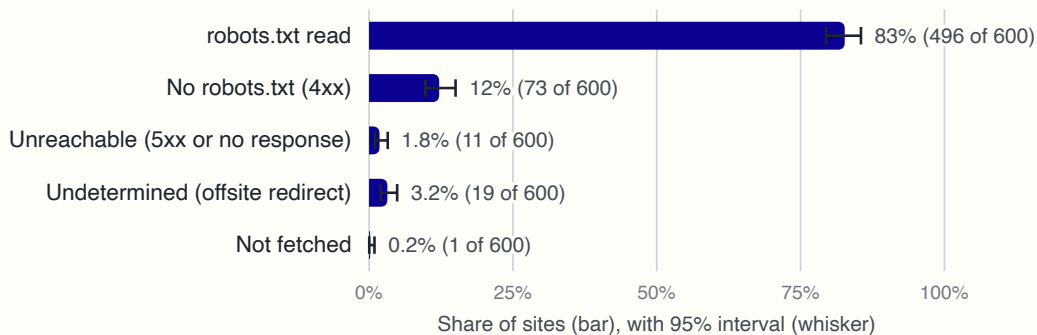
**Who is in the sample.** We started from a public mirror of the Y Combinator company directory (the `yc-oss/api` project on GitHub, at commit `92a11048a71f`). We kept active and public companies in YC's B2B and Fintech industries that list a website, reduced each website to its registrable domain using the Public Suffix List, dropped shared platforms such as LinkedIn pages, and removed duplicates. That left 2,752 sites. We drew 600 of them at random, split into 12 groups by industry (developer tools and infrastructure, fintech, other B2B) and by batch year, with each group's share of the sample equal to its share of the list. The random seed is fixed, so anyone can rebuild the same sample.

**How we fetched.** On October 7, 2026 our script requested `/robots.txt`, and nothing else, once from each site's main host, identifying itself as `ElementeraResearch/1.0` with a link to this page. It followed redirects only within the same domain.

**What came back.** 496 of 600 sites (83%; 95% interval 79% to 85%) returned a `robots.txt` file, and 73 of 600 sites (12%; 95% interval 9.8% to 15%) returned a "not found" or other 4xx answer, which under the `robots.txt` standard (RFC 9309) means no restrictions. Together they are the 569 sites every share on this page is computed over. 11 of 600 sites (1.8%; 95% interval 1.0% to 3.2%) gave a server error or no answer, and 19 of 600 sites (3.2%; 95% interval 2.0% to 4.9%) redirected their `robots.txt` to a different domain; we report both separately and never count them as blocks. 7 of 600 sites (1.2%; 95% interval 0.6% to 2.4%) answered with 401, 403 or 429, which the standard treats as "no restrictions"; and 11 of 496 files (2.2%; 95% interval 1.2% to 3.9%) were web pages rather than `robots.txt` files, so every crawler counts as allowed there.

### What happened when we asked for robots.txt

n = 600 sampled sites; robots.txt fetched 2026-10-07



**How we read the files.** We used the parser from our crawler checker, which follows RFC 9309: a crawler obeys the most specific group that names it, falls back to `User-agent: *` only when no group names it, and within a group the longest matching rule wins. The crawler list was checked against each vendor's documentation and the community-maintained `ai robots.txt` list on October 6, 2026.

**What robots.txt covers.** `robots.txt` is where a site states its rules for crawlers in public. Firewalls and CDN bot settings act at the network edge and can refuse a crawler that `robots.txt` allows; those settings are not visible from outside a site, so they sit outside this measurement and are worth checking on your own site. A block records the site's request; it does not record whether a given crawler honours it.

**How we count.** Every figure is a count of sites, its share, and a 95% Wilson interval, which shows the range the true share in the full list plausibly falls in. When two intervals overlap, we say the two figures cannot be told apart at this sample size. The figures describe this sample of Y Combinator-backed software companies on the day we fetched it.

By industry group, the share of sites blocking at least one AI search crawler and at least one AI training crawler:

| INDUSTRY GROUP                     | BLOCKS AN AI SEARCH CRAWLER                      | BLOCKS AN AI TRAINING CRAWLER                     |
|------------------------------------|--------------------------------------------------|---------------------------------------------------|
| Other B2B software                 | 2 of 298 sites (0.7%; 95% interval 0.2% to 2.4%) | 10 of 298 sites (3.4%; 95% interval 1.8% to 6.1%) |
| Developer tools and infrastructure | 1 of 168 sites (0.6%; 95% interval 0.1% to 3.3%) | 3 of 168 sites (1.8%; 95% interval 0.6% to 5.1%)  |
| Fintech                            | 0 of 103 sites (0%; 95% interval 0% to 3.6%)     | 5 of 103 sites (4.9%; 95% interval 2.1% to 11%)   |

**When it was run, and what comes next.** Data collected October 7, 2026. When we re-run the study, the new figures and the new date will appear on this page.

## Frequently asked questions

**Does blocking GPTBot remove my site from ChatGPT search?** No. OpenAI documents GPTBot and OAI-SearchBot as separate settings: GPTBot covers training, and OAI-SearchBot decides whether the site can appear in ChatGPT search answers. You can block GPTBot and allow OAI-SearchBot. In our sample, 15 of 569 sites (2.6%; 95% interval 1.6% to 4.3%) block training while leaving every AI search crawler open, and none do the reverse.

**What is the difference between ClaudeBot and Claude-SearchBot?** ClaudeBot is Anthropic's training crawler; Claude-SearchBot indexes pages for search in Claude. Anthropic says disabling Claude-SearchBot may reduce a site's visibility and accuracy in user search results. A third name, Claude-User, fetches a page when a person asks Claude to read it. Each takes its own `robots.txt` group.

**Should I allow PerplexityBot?** If you want to be cited in Perplexity's answers, yes. Perplexity describes PerplexityBot as the crawler that surfaces and links websites in its search results, and says it is not used to crawl content for AI foundation models. Blocking it removes the main route by which Perplexity can find your pages.

**Does Cloudflare's AI blocking show up in robots.txt?** Sometimes. Cloudflare can write a managed section into robots.txt, which we count as a platform default, but its AI bot settings also act at the network edge ([Cloudflare](#)), where robots.txt shows nothing. A site can look open in its robots.txt and still refuse AI crawlers at the edge, so check both.

**Why count a 403 on robots.txt as "allowed"?** Because the robots.txt standard says crawlers may treat any 4xx answer as "no restrictions". We follow the standard and report those sites separately, because a 403 often comes from a firewall that may also refuse the crawlers themselves.

**Can I use the data?** Yes. The data is on [our dataset page](#), with a download of the full per-site file. Each site has a random id instead of a name, so the file shows patterns, not companies. Please cite the report when you use it.

## Work with us

There are two ways to work with us. Teams that want the work done for them can use our managed AEO services: [Book a call](#). Teams that want to see where they stand first can start with Lenz, our free analysis of how ChatGPT and Gemini answer their buyers' questions: [Get your free AI visibility report](#).

## How this report was made

Our own scripts collected and classified the data, and the free checker at [/tools/ai-crawler-check](#) reads robots.txt the same way. The sample and the classification rules were fixed before we fetched any site's robots.txt, and the analysis code before the full collection. We used an AI writing assistant (Claude, made by Anthropic) to help draft the text, and Mojtaba Ebadi checked every figure and sentence against the data before publication. Soroush Mehri reviewed the page, and Mojtaba Ebadi approved it for publication.

## Sources

- OpenAI, "Overview of OpenAI Crawlers": [developers.openai.com/api/docs/bots](#)
- Anthropic, "Does Anthropic crawl data from the web, and how can site owners block the crawler?": [support.claude.com](#)
- Perplexity, "Perplexity Crawlers": [docs.perplexity.ai](#)
- Google, "Google's common crawlers": [developers.google.com](#)
- Apple, "About Applebot": [support.apple.com](#)
- Meta, "Meta Web Crawlers": [developers.facebook.com](#)
- Common Crawl, "CCBot": [commoncrawl.org](#)
- ai.robots.txt, community list of AI agents: [github.com/ai-robots-txt/ai.robots.txt](#)
- Cloudflare, "Block AI Bots" (bot solutions documentation): [developers.cloudflare.com](#)
- Grossman, Liu, Chen, Smith, Borcea and Chen. "How Generative AI Disrupts Search: An Empirical Study of Google Search, Gemini, and AI Overviews." ACM SIGIR 2026. DOI 10.1145/3805712.3809667. [arXiv 2604.27790](#)

- Longpre, Mahari and colleagues. "Consent in Crisis: The Rapid Decline of the AI Data Commons." arXiv preprint, 2024. [arXiv 2407.14933](https://arxiv.org/abs/2407.14933)
- Koster, Illyes, Zeller and Sassman, "Robots Exclusion Protocol", RFC 9309, IETF, 2022: [rfc-editor.org](https://rfc-editor.org/rfc/rfc9309/)
- Y Combinator company directory mirror, yc-oss/api: [github.com/yc-oss/api](https://github.com/yc-oss/api)
- Public Suffix List: [github.com/publicsuffix/list](https://github.com/publicsuffix/list)
- Wilson, "Probable inference, the law of succession, and statistical inference", Journal of the American Statistical Association, 1927: [doi.org/10.1080/01621459.1927.10502953](https://doi.org/10.1080/01621459.1927.10502953)