57% of sites that block GPTBot leave ChatGPT-User unrestricted

Richard Fentner
Founder and Managing Director — Fentner Group
United Kingdom
research@fentner.com

***Abstract*—On 2 September 2026 we fetched robots.txt for the Tranco top 20,000 domains and parsed every file against RFC 9309. Of the 2,207 sites that block GPTBot, OpenAI's training crawler, 1,266 — 57.4% — leave ChatGPT-User unrestricted, the agent OpenAI sends when somebody asks ChatGPT about a page. The block stops the crawler that collects pages for training and leaves untouched the one that reads the page while the user watches. The pattern is not confined to OpenAI and it is not marginal: 2,559 sites block at least one vendor's bulk crawler, and 2,322 of them — 90.7% — leave at least one live sibling of that same vendor unrestricted. Most of these files never mention the live agents at all, which is not how a deliberate policy is written. This is a measurement of files, not of crawler behaviour.**

WHAT "READ THEM LIVE" MEANS Every large vendor now runs several crawlers under separate robots.txt product tokens, and they do different jobs. OpenAI documents GPTBot for model training, OAI-SearchBot for the ChatGPT search index, and ChatGPT-User for a fetch triggered by a person's question [1]. Anthropic documents ClaudeBot, Claude-SearchBot and Claude-User [2]. Perplexity documents PerplexityBot and Perplexity-User [3]. Google splits its fleet into common crawlers, special-case crawlers and user-triggered fetchers [4]. Meta and Amazon have each done the same [5, 6].

Directives are addressed to one token and reach no other, and the vendors confirm this twice over. First, groups in robots.txt are matched per product token: RFC 9309 has a crawler use the group naming it, or the * group if none does. Second, several vendors state that the user-triggered fetch is outside robots.txt altogether. Perplexity writes that "Since a user requested the fetch, this fetcher generally ignores robots.txt rules" [3]. Google writes that "Because the fetch was requested by a user, these fetchers generally ignore robots.txt rules" [4]. Meta writes that its user fetcher "may bypass robots.txt because it performs fetches that were requested by the user" [5]. OpenAI is softer: "Because these actions are initiated by a user, robots.txt rules may not apply" [1].

Anthropic is the exception and it makes the same point from the other side. It states that its bots "respect 'do not crawl' signals by honoring industry standard directives in robots.txt" [2], and it lists Claude-User as a separate token, writing that disabling it "prevents our system from retrieving your content in response to a user query" [2]. Both robots.txt examples on that page address ClaudeBot. Claude-User obeys a rule addressed to Claude-User. It does not obey a rule addressed to ClaudeBot.

So there are two mechanisms and one result. Either the live agent is documented as ignoring the file, or it obeys only the rule the site never wrote. In neither case does a block on the training crawler stop the assistant reading the page.

What we measured

Sample: the Tranco daily list generated on 1 September 2026, list `K9QPW`, ranks 1 to 20,000. Each domain was fetched twice, once with a research user-agent naming this study and once with a browser user-agent, at 20

Date of publication 02 09 2026; date of current version 02 09 2026

concurrent requests. The second fetch matters: 389 domains (1.9%) refused the research agent and served the browser agent, and counting those as "no AI rules" would have been wrong. Anything that failed or was refused was retried through a rotating proxy, which recovered a further 291 domains that a single address could not reach; a refusal from one address is not a fact about a website.

12,139 domains (60.7%) returned a usable robots.txt. 6,619 returned nothing to either agent. 1,049 (5.2%) returned HTTP 200 with an HTML page at `/robots.txt` — a single-page-application catch-all or a soft 404 — which we count as having no robots.txt at all, whatever the status code says. A further 49 served a bot challenge instead, and are excluded rather than counted either way.

Parsing follows RFC 9309: product tokens compared case-insensitively, consecutive user-agent lines forming one group, groups naming the same token merged, unnamed crawlers falling back to `*`, and longest-match evaluation with `allow` winning ties. Two predicates run through everything below. A site *blocks* a crawler when the file names it and disallows it something. A crawler is *restricted* when it is either named and disallowed, or unnamed while `*` disallows the whole site — so a site with a blanket `Disallow: /` is never counted as leaving anything open.

Every token list was read from the vendor's own documentation the same day rather than from a bot directory. Amazon documents three tokens today, Mistral three and Meta five.

The gap, vendor by vendor

vendor	blocks the bulk crawler	leaves a live sibling open	leaves the user-triggered agent open
OpenAI (GPTBot)	2,207	1,537 (69.6%)	1,266 (57.4%)
Anthropic (ClaudeBot)	1,932	1,519 (78.6%)	1,472 (76.2%)
Meta (meta-externalagent)	1,626	1,428 (87.8%)	1,200 (73.8%)
Amazon (Amazonbot)	1,604	1,478 (92.1%)	1,475 (92.0%)
Perplexity (PerplexityBot)	1,018	512 (50.3%)	512 (50.3%)

Amazon is the worst case: 1,475 of 1,604 Amazonbot blockers leave Amzn-User unrestricted, the agent Amazon documents for answering an Alexa query that needs current information. Perplexity is the least bad, and only because it has one sibling to miss rather than two.

The error runs one way. Only 80 sites block an OpenAI live agent without blocking GPTBot, against 1,537 that block GPTBot and leave a live agent open, so this is not noise in the parse. It is a population of files written once, against the crawler that was in the news, and not revisited.

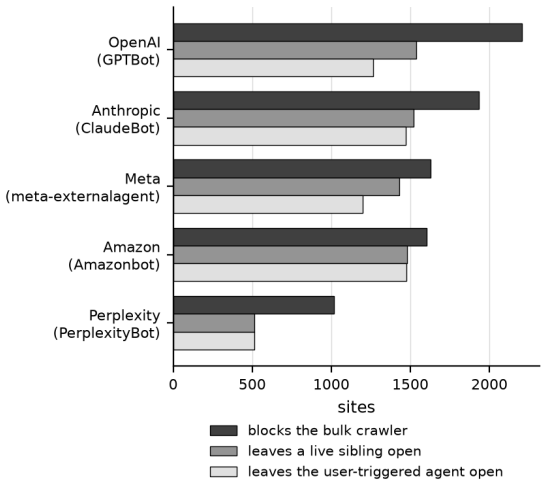

Figure 1. For each vendor, the number of sites that block the bulk crawler, the number of those that leave a live sibling open, and the number that leave the user-triggered agent open.

The gradient by rank says the same thing. Among GPTBot blockers, the share leaving a live sibling open runs from 42.5% in the top thousand to 74.8% in the 10,001–20,000 band. Care is a function of how much attention a site can afford, and it decays quickly.

Not a policy, a stale snippet

A site could block training and permit live citation on purpose. That is a defensible position, and for some of these files it is surely the intent. It cannot be the explanation for most of them, because most of them do not mention the live agents at all: 1,167 of the 2,207 GPTBot blockers (52.9%) name neither OAI-SearchBot nor ChatGPT-User anywhere in the file. A policy names the thing it is permitting.

The shape of the files points the same way. 1,813 of the 2,207 GPTBot blockers also name CCBot, the two names that travel together in the blocks people copy, and 1,106 name no live token at all — search or user-triggered — from any of the six vendors that publish one alongside a bulk crawler.

Named files, each re-fetched while this was being written and saved with the dataset:

- **medium.com**, **latimes.com**, **webmd.com**, **healthline.com**, **patreon.com** and **cbsnews.com** each disallow GPTBot and never name OAI-SearchBot or ChatGPT-User.
- **forbes.com** goes further and writes the permission out: GPTBot and ClaudeBot are disallowed, and `OAI-SearchBot`, `Claude-SearchBot` and `Claude-User` are given `Allow: /`.
- **canva.com** disallows GPTBot, ClaudeBot, `anthropic-ai` and `Claude-Web`, then allows PerplexityBot and Perplexity-User.
- **apnews.com** blocks Claude-User but not ChatGPT-User, which is the same mistake with the vendors swapped.

Sites that closed the gap exist and are easy to find: **bbc.com**, **cnn.com**, **amazon.com**, **tiktok.com** and **bloomberg.com** each disallow GPTBot, OAI-SearchBot and ChatGPT-User by name. Nothing about the format prevents it. None of the five is complete across every vendor: bbc.com blocks ClaudeBot and never names Claude-User, and bloomberg.com blocks PerplexityBot and never names Perplexity-User.

Two Anthropic crawlers that are not in the documentation

1,122 sites (9.2% of those with a robots.txt) write rules for `anthropic-ai` or `Claude-Web`. Neither token appears anywhere in Anthropic's current crawler documentation. Anthropic publishes no statement retiring them, so the honest description is that they are undocumented rather than dead — but a rule addressed to a name the vendor no longer publishes is a rule with nothing on the other end.

127 sites name one of the two and never name ClaudeBot, Claude-User or Claude-SearchBot at all. Two of them are worth stating plainly. **cloudflare.com**, which sells AI-bot blocking as a product, writes `Allow: /` groups for `Anthropic-AI` and `Claude-Web` and names no current Anthropic crawler anywhere in the file — a welcome addressed to nobody. **chatgpt.com** takes the opposite position with the same two names, disallowing `anthropic-ai` and `Claude-Web` and never mentioning ClaudeBot. **espn.com** and **science.org** disallow `anthropic-ai` and name no current Anthropic crawler either.

Rules that cannot match anything

27 files contain a user-agent token with a character that is not ASCII, which under RFC 9309 cannot match the product token it is aiming at. 13 of those involve a character that looks exactly like a hyphen and is not.

13 sites carry the identical line `User-agent: Perplexity-User` written with the character U+2011 (non-breaking hyphen) in place of the hyphen-minus: **chatgpt.com**, **launchpad.net**, **academia.edu**, **seekingalpha.com**, **foreignpolicy.com**, **f95zone.to**, **forth.gr**, **weatherbug.com**, **tempo.co**, **unil.ch**, **thisoldhouse.com** and **haufe.de** among them. The same invisible defect in thirteen files, on twelve unrelated operators, is a fingerprint: it was typed once somewhere that autocorrects hyphens, and pasted from there. Three of the thirteen attach `Allow: /` to the broken token and one writes the token correctly elsewhere in the file as well; the other nine lose a disallow they meant to have.

ptsecurity.com shows the other way to write a file that does nothing. Every AI user-agent line carries an explanatory phrase appended without a `#`, so the token becomes `GPTBot OpenAI` followed by an em dash and a Russian-language gloss on ChatGPT model training, and thirteen consecutive groups — GPTBot, OAI-SearchBot, ChatGPT-User, Google-Extended, ClaudeBot, Claude-Web, `anthropic-ai`, PerplexityBot, Perplexity-User, Amazonbot, Applebot-Extended, YouBot and YandexAdditional — address nothing at all. **lemonde.fr** has `User-agent: InMédia Technologies`, with an accent, in an otherwise careful file.

Sites with no robots.txt that believe they have one

1,049 of the 20,000 domains (5.2%) answer `/robots.txt` with HTTP 200 and an HTML page. **office.com**, **azure.com**, **cloud.microsoft**, **github.io** and **pages.dev** are among them. There are no rules on any of those hosts; there is a web page where the rules would be, and any crawler reading it finds no directives and proceeds.

Google-Extended does not do what most of its users think

1,754 sites disallow Google-Extended. Google-Extended is not a crawler; it is an opt-out control covering Gemini model training and grounding in

Gemini apps and Vertex AI, and Google states that it "does not impact a site's inclusion in Google Search nor is it used as a ranking signal in Google Search" [7]. A page must be indexed for Search to appear in AI Overviews or AI Mode, and Google's own page on AI features says that "robots.txt directives for Googlebot is the control for site owners to manage access" [8].

1,583 of those 1,754 sites (90.3%) leave Googlebot unrestricted. They have opted out of Gemini training and remain fully available to the Google AI surface most of them are likely thinking about. That is a coherent choice for a publisher who wants search traffic. It is unlikely to be the choice most of them think they made.

What this does not show

This is a measurement of files, not of behaviour. We did not observe a single crawler and we did not ask any AI product to read a blocked page, so nothing here shows that any bot did or did not honour a rule. Compliance is a separate study and a harder one.

Nor does it show intent. A site that blocks training and permits live citation may mean exactly that, and we cannot read minds; the strongest thing the data supports is that the majority of these files never mention the live agents, which is not how a deliberate policy is written.

A site counts as blocking GPTBot if it disallows the crawler anything, not only the whole site; 1,841 of the 2,207 write `Disallow: /`, and 1,202 of those (65.3%) still leave ChatGPT-User unrestricted. The mirror of that rule makes the open counts conservative: a site that disallows a live agent one directory is counted as restricting it, not as leaving it open.

The sample is the Tranco top 20,000 domains, which includes CDN, DNS and tracking hosts that serve no site at all, and that is part of why only 12,139 of them returned a robots.txt; it describes prominent domains and says nothing about the long tail. robots.txt is also only one lever: a site may block the same traffic at the firewall, at the CDN, or through a licensing deal, and none of that is visible in the file. Fetches came from many addresses and from two user-agents, and a site that varies its robots.txt by region or by client would be recorded as whichever version was served here. Finally, a robots.txt is a snapshot. These files change, and thirteen of them will probably have had a hyphen fixed by the time you read this.

Data availability

The full dataset, the parser, the analysis script and the saved robots.txt of every site named above are published under CC BY 4.0 at https://fentner.com/data/gptbot-block-leaves-chatgpt-open/:

- https://fentner.com/data/gptbot-block-leaves-chatgpt-open/robots_20000.csv
- https://fentner.com/data/gptbot-block-leaves-chatgpt-open/numbers.json
- https://fentner.com/data/gptbot-block-leaves-chatgpt-open/lookalike_sites.csv
- https://fentner.com/data/gptbot-block-leaves-chatgpt-open/METHOD.md
- https://fentner.com/data/gptbot-block-leaves-chatgpt-open/README.md

Every figure above is computed by `analyse.py` from the saved fetch and appears in `numbers.json`.

References

[1] OpenAI, *Bots*. https://developers.openai.com/api/docs/bots

[2] Anthropic, *Does Anthropic crawl data from the web, and how can site owners block the crawler?* https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler

[3] Perplexity, *Bots*. https://docs.perplexity.ai/guides/bots

[4] Google, *Google user-triggered fetchers*. https://developers.google.com/search/docs/crawling-indexing/google-user-triggered-fetchers

[5] Meta, *Web crawlers*. https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/

[6] Amazon, *Amazonbot*. https://developer.amazon.com/amazonbot

[7] Google, *Google common crawlers*. https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers

[8] Google, *AI features and your website*. https://developers.google.com/search/docs/appearance/ai-features