

Spoofed AI crawlers: how much traffic claiming to be an AI or search bot is real?

EasyxLab · EasyByte Hub S. Coop. Mad. · Study S2

Working draft Published 2026-10-02 · Updated 2026-10-03

EasyxLab, study S2 — working draft, 2026-10-02. Not peer-reviewed. See §8 for what was done by AI agents.

Abstract

Website owners increasingly count "AI crawler" requests in their logs as a signal of whether language models read, index or cite them. The User-Agent header that carries the crawler's name is free text. We verified every request claiming to be one of 23 AI or search crawlers across three small production websites (one 68-day log, two 7–9-day logs, plus a 30-day Cloudflare edge view), using only each operator's own published verification method: IP-range JSON files and forward-confirmed reverse DNS. Of 16,548 origin-log requests carrying a known crawler name, 49.5% were verified, **39.3% were spoofed**, 4.9% were unverifiable (no method published) and 6.3% indeterminate. Among requests whose claim could be tested, 44.3% were false. The problem is concentrated in exactly the categories site owners read as evidence of "being used by AI": 67.8% of *user-initiated fetch* requests (ChatGPT-User, Claude-User, Perplexity-User, MistralAI-User, DuckAssistBot) were spoofed, and for four of those five names the genuine share was between 0% and 4.2%. Spoofing came from at most 54 IP addresses, overwhelmingly in cloud and small hosting networks (one cloud provider's customer space alone produced 66% of spoofed requests), and was mostly not crawling: 97% of spoofed requests came from sources that also probed for credentials (`.env` , `config.json` , `service-account.json`), 92% from sources that rotated through two or more operators' identities, and 88% got a 4xx. On the main site, at the CDN edge, the spoofed volume was 3.1 times what the origin logged, and the CDN's free-plan defaults stopped 0.7% of it. We release `ai-bot-verify` , a dependency-free verifier that any webmaster can run on an access log and that publishes only aggregates.

1. Introduction

Two things changed in 2024–2026. First, AI operators split their crawlers by purpose — training (GPTBot, ClaudeBot, CCBot), search indexing (OAI-SearchBot, Claude-SearchBot, PerplexityBot) and fetches made at the moment a user asks something (ChatGPT-User, Claude-User, Perplexity-User). The last group is read by many site owners as the closest thing to an "AI citation" signal in their own logs. Second, most of these operators now publish machine-readable IP lists, which makes verification cheap.

Published figures on how much of this traffic is fake disagree widely. HUMAN Security reported that its Satori Threat Intelligence team "analyzed traffic associated with 16 well-known AI crawlers and found that 1 in every 18 requests using an AI crawler user agent is fake" (≈5.6%; HUMAN blog, 2026-02-28). A single-site account reports the opposite extreme: "Over the last two weeks, my logs claimed 33 AI assistants visited, a little better than two a day. That number is a lie. The real number? Six." and, for Googlebot, "Of 799 requests carrying the Googlebot name, only 107 came from a verified Google address. The other 692,

roughly 87%, were not Google." (D. Forrester, Search Engine Journal, 2026-06-25; the article's title gives the first ratio as 81.8%). The gap is large enough to change decisions. The operator of the sites studied here had itself estimated user-initiated AI fetches partly from user-agent counts.

This study asks, for small real websites: what share of requests claiming to be an AI or search crawler really comes from that operator; who sends the rest, by network rather than by address; what they ask for; and when.

2. Data

Three production sites run by the authors, all static content behind the same stack (Cloudflare → reverse proxy → nginx), all with a permissive `robots.txt` (`User-agent: *` / `Allow: /`). They are identified here only by type and age:

SITE	DESCRIPTION	ORIGIN-LOG WINDOW	DECLARED-BOT REQUESTS
Site A	a calculator site, online since mid-2026	2026-07-27 → 10-02 (68 calendar days)	14,637
Site B	a niche tool site launched in September 2026	2026-09-24 → 10-02 (9 calendar days)	1,481
Site C	a niche tool site launched in September 2026	2026-09-26 → 10-02 (7 calendar days)	430

nginx logs the real client address from `CF-Connecting-IP` ; the reverse proxy only trusts forwarding headers from Cloudflare's published ranges. Cloudflare's GraphQL analytics for the same zones add a 30-day edge view (2026-09-02 → 10-02) that includes requests answered from cache and requests the origin does not log (see §4.6).

3. Method

Full detail in `METHOD.md` . In short:

- Scope.** A request is in scope if its User-Agent contains one of 23 crawler tokens from 13 operators (case-insensitive, most specific first).
- Verification, by the operator's own rules.** Each operator's documentation was downloaded on 2026-10-02 and its method applied literally: OpenAI, Anthropic, Perplexity, Google, Bing, Apple, Common Crawl, DuckDuckGo and Mistral publish IP-range JSON files; Google, Bing, Apple, Common Crawl and Yandex also document forward-confirmed reverse DNS. Two tokens are, per their owners, never sent as a User-Agent: Google says "Google-Extended doesn't have a separate HTTP request user agent string", and Apple says "Applebot-Extended does not crawl webpages". Any request bearing them is spoofed by definition. ByteDance publishes no method (Bytespider: *unverifiable*); Amazon publishes lists only through a JavaScript page (*indeterminate*). Anthropic's support page links a single list for all its bots (`claude.com/crawling/bots.json`) with the wording "If a crawler has a source IP address on this list, it indicates that the crawler is coming from Anthropic."

3. **Behaviour.** For each source (in memory) we record whether it requested a credential or config path, whether it claimed more than one operator, whether it read `robots.txt` first, the status codes it received and the hour. ASNs of non-verified sources come from Team Cymru.
4. **Privacy.** Only aggregates are published. The tool refuses to write output containing any client IP seen in the input; an independent check over the 7,781 client IPs in the raw logs found none in the published files. Site paths are published only for scanner probes; the holder of one AS is a natural person, so its number and name are withheld.

4. Results

4.1 Overall

	REQUESTS	SHARE
verified	8,186	49.5%
spoofed	6,498	39.3%
unverifiable (no published method, or client-side tool)	818	4.9%
indeterminate (Amazon)	1,046	6.3%
total claiming a known crawler	16,548	

Spoofed as a share of *testable* claims (verified + spoofed): 44.3%.

The three sites differ sharply. On Site A, the oldest, 44.1% of all claims were spoofed. On the two sites launched in September, spoofing was 0.2% (Site B, 3 requests) and 9.8% (Site C, 42 requests from two addresses within six days of launch). With three sites and windows of 68, 9 and 7 days, we cannot separate site age from window length or content. In addition, since 2026-09-27 the sites' nginx no longer logs a list of common scanner paths (it closes them with status 444 and `access_log off`), which covers almost the entire window of Sites B and C and lowers their origin-log spoofing counts.

4.2 By crawler

Share of testable requests that were genuine (all sites pooled):

CRAWLER	OPERATOR	PURPOSE	REQUESTS	VERIFIED	SPOOFED	% GENUINE
Google-special (AdsBot, Mediapartners...)	Google	other	110	104	6	94.5
Bingbot	Microsoft	search	1,902	1,790	112	94.1
DuckDuckBot	DuckDuckGo	search	96	79	17	82.3
YandexBot	Yandex	search	642	513	129	79.9
Googlebot	Google	search	2,548	1,720	828	67.5
ClaudeBot	Anthropic	training	1,808	1,139	669	63.0
OAI-SearchBot	OpenAI	search index	1,569	891	678	56.8
ChatGPT-User	OpenAI	user fetch	1,713	835	878	48.7
Applebot	Apple	search	519	240	279	46.2
GPTBot	OpenAI	training	979	448	531	45.8
PerplexityBot	Perplexity	search index	839	325	514	38.7
CCBot	Common Crawl	training	452	84	368	18.6
DuckAssistBot	DuckDuckGo	user fetch	166	7	159	4.2
Claude-User	Anthropic	user fetch	318	11	307	3.5
Perplexity-User	Perplexity	user fetch	309	0	309	0.0
Claude- SearchBot	Anthropic	search index	196	0	196	0.0
MistralAI-User	Mistral	user fetch	146	0	146	0.0

CRAWLER	OPERATOR	PURPOSE	REQUESTS	VERIFIED	SPOOFED	% GENUINE
Google-Extended	Google	(robots token)	371	0	371	0.0
Google-fetcher (user-triggered)	Google	user fetch	1	0	1	0.0

Not testable: Amazonbot 1,046 (indeterminate), meta-externalagent 437 and Bytespider 376 (unverifiable), and 5 requests from a client-side coding agent that sends "Claude-User" from the user's own machine.

By purpose (spoofed share of all requests): user fetch **67.8%**, training 43.8%, search indexing 38.0%, classic search engines 23.9%. The less a name is actually used, the more of its traffic is fake. That is consistent with impostors rotating through a fixed list of names regardless of how often the real crawler visits.

4.3 Who spoofs

At most 54 distinct IP addresses produced the 6,498 spoofed requests (51 on Site A; the sum over sites is an upper bound), against more than 1,600 verified crawler addresses. Spoofed requests by network (pooled, Team Cymru):

NETWORK	SPOOFED REQUESTS
AS396982 Google Cloud Platform (customer address space, not Google's crawlers)	4,288 (66%)
a small AS registered to an individual (number and name withheld)	772
AS400810 BreezeHost	609
AS1004 Ambyre	276
AS14061 DigitalOcean	125
AS400529 Infraly	121
AS399804 Hostodo	115
AS46475 Limestone Networks	69
AS16509 Amazon (AWS)	44
other 7 ASNs (incl. two Cloudflare ASNs and one Spanish residential ISP)	79

By heuristic network type (from the registered AS name): cloud/hosting 5,272 requests (81.1%); "unknown" 1,202 (18.5%), of which 772 come from the single AS registered to an individual and 430 from three small providers whose names the heuristic could not classify plus one ISP; residential ISP 24 (0.4%). The

residential requests are unexplained; they could include test requests from the authors' own tooling on connections with dynamic addresses, which could not be excluded by IP. We found no spoofed request from an address space the impersonated operator itself uses: none of the requests claiming to be OpenAI came from Microsoft/Azure, and none claiming Google came from Google's own AS15169. That argues against the alternative explanation that "spoofed" requests are real crawlers on addresses newer than the published lists.

For the untestable names, the registered owner of the source network is informative but not proof: 82% of "Amazonbot" requests came from non-Amazon networks (mostly Google Cloud); 0% of "Bytespider" requests came from a ByteDance-registered AS; on Site A, 73% of meta-externalagent requests came from outside Meta's AS (on Site B, all came from it).

4.4 What spoofers ask for

	VERIFIED	SPOOFED
requests to known credential/config/admin paths	<0.1%	58.3%
<code>/robots.txt</code>	13.7%	0.2%
HTTP 4xx (incl. 444 connection closed)	1.1%	87.8%
requests from sources that asked for ≥ 1 probe path	4 requests	97.0%
requests from sources that claimed ≥ 2 different operators	0	92.2%

The most requested spoofed paths were `/api/config`, `/env.js`, `/settings.json`, `/env.json`, `/config.json`, `/secrets.json`, `/openapi.json`, `/api/env`, `/app/.env`, `/graphql`, `/application.properties`, `/@fs/.env`, `/service-account.json`, `/terraform.tfstate` and `/@fs/proc/self/environ`. Only one spoofing source (claiming Bingbot) read `robots.txt` before its first request; verified crawlers frequently did. Because `robots.txt` allows everything, "respecting `robots.txt`" could only be observed as "reading it", and spoofers essentially do not. The remaining 37% of spoofed requests went to ordinary site pages, so not all of it is probing; but the dominant pattern is a credential scanner wearing a rotating list of AI and search crawler names.

4.5 When

Spoofing is bursty. On Site A, the five busiest days held 48% of all spoofed requests (max 806 in one day) against 17% for verified crawlers; spoofed traffic appeared on 48 of 68 days. Weekly spoofed volume ranged from 144 to 1,460 requests with no trend. No human-day pattern; bursts dominate: 27% of spoofed requests fall at 04:00 and 16:00 UTC (980 and 755 requests).

4.6 Edge vs origin

For the same 30 days on Site A, Cloudflare's edge saw 10,454 spoofed requests where the origin logged 3,412 ($\times 3.1$), and 7,950 verified ones where the origin logged 4,455 ($\times 1.8$, the difference being mostly cached `robots.txt`, sitemaps and assets). The origin undercounts spoofing because nginx closes common scanner paths with status 444 *without logging them*, and undercounts verified crawling because of the

cache. At the edge, the spoofed share of testable claims was 56.8%, and 63.5% of spoofed requests went to probe paths. Cloudflare's free-plan defaults blocked 72 of the 10,454 spoofed requests (0.7%).

5. Discussion

User-agent counts are not evidence of AI usage. Across the three sites, counting "Claude-User" by name gives 318 user-initiated fetches; 11 were real. On Site A alone it is 313, of which 6 were real. "Perplexity-User": 309, none real. Only ChatGPT-User survives verification in volume (835 genuine across the three sites, 811 on Site A), and even there half of the claims were false. Any report built from UA counts of user-fetch agents is dominated by scanners.

The spoof rate may be a property of the site's exposure rather than of the crawler. One explanation consistent with, but not tested by, our data: a small number of scanner sources produce a roughly fixed volume per target, so on low-traffic sites the same scanners are a large share of the total. That would reconcile HUMAN's $\approx 5.6\%$ over a large base with the single-site figures above. The two new sites show what may be the early phase: almost nothing at launch, a first spoofing source within days.

Rare names are mostly fake. Names like Claude-SearchBot, MistralAI-User, DuckAssistBot and Google-Extended appear in impostor rotations whether or not the real crawler ever visits. For Google-Extended the answer is categorical: the token is never a User-Agent.

Verification is cheap and mostly automatable. Nine of thirteen operators publish machine-readable ranges. Two gaps remain: Amazon publishes lists only via a JavaScript page, and ByteDance publishes nothing. One more caveat for practitioners: Googlebot does not only crawl from 66.249.0.0/16. 48 of the 170 IPv4 prefixes in Google's current list are outside it, and 223 of the 1,720 verified Googlebot requests here (13%) came from them; a rule restricted to that block — common in tutorials, and the rule the authors' own earlier monitoring used — would call them fake.

Implications. Allow-listing by User-Agent invites exactly the traffic seen here; allow-listing by published range does not. Operators that publish one list for all their bots (Anthropic) cannot be verified per purpose: an IP on the list could be any of its three bots.

6. Limitations

- **Small, self-selected sample.** Three static sites run by the same team on one stack. Site A provides 88% of the data. Rates on large or dynamic sites will differ.
- **Short windows.** 68, 9 and 7 calendar days of origin logs; 30 days at the edge.
- **Retroactive ranges.** Lists downloaded on 2026-10-02 were applied to requests up to 68 days older. A crawler address added or removed in that period could be misclassified. The behaviour of the "spoofed" group (credential probing, multi-identity, no source in the operator's own networks) makes material misclassification unlikely, but it is not excluded.
- **Unlogged probes.** Since 2026-09-27 the sites' nginx does not log a list of scanner paths, so origin-log spoofing is a floor; the edge view partly corrects this but is sampled.
- **Edge sampling.** Cloudflare adaptive sampling applied to 54% of Site A's bot rows; edge counts are raw sampled counts (lower bound). The edge window ends at the time of the run.
- **Header trust.** The client IP comes from `CF-Connecting-IP`. A client reaching the origin directly could forge it; 11 of $\sim 120,000$ lines lacked a Cloudflare ray id.

- **Heuristic network typing.** Network type is inferred from the registered AS name; VPN and residential-proxy traffic cannot be reliably identified this way.
- **Methods not applied.** Amazon (indeterminate) and Meta/ByteDance (unverifiable) are excluded from the spoof rate; their real share is unknown.
- **Undeclared crawlers are out of scope.** This study only checks requests that *claim* a crawler name. Cloudflare reported the reverse problem — "Perplexity uses not only their declared user-agent, but also a generic browser intended to impersonate Google Chrome on macOS when their declared crawler was blocked" (Cloudflare blog, 2025-08-04) — which this method cannot see.
- A third-party summary of DataDome's 2026 report could not be checked against the source (HTTP 1. and is not cited.

7. Tool and data

`scripts/ai_bot_verify.py` (CLI wrapper `ai-bot-verify`): Python ≥ 3.10 , standard library only, 15 offline tests. Input: one or more access logs (`.gz` accepted). Output: aggregated CSV/JSON by bot, verdict, ASN, network type, path category, hour and day. The tool never writes an IP address.

`scripts/cf_declared_bots.py` applies the same checks to Cloudflare GraphQL analytics. All aggregates are in `data/`; sites appear only as `site-a`, `site-b`, `site-c`.

8. Automation and review

This study was carried out by AI agents (large-language-model coding agents) working in a terminal under the authors' accounts, with read-only access to the servers:

- **Done by AI agents:** reading the existing in-house scripts and notes; copying the logs; downloading the operators' documentation and range files and extracting the quoted text; writing `ai-bot-verify`, its tests and the Cloudflare and aggregation scripts; running every analysis; assigning network types (by keyword heuristic) and deciding, from a whois record, to withhold one AS as registered to an individual; searching for and quoting the external reports; writing this draft, `METHOD.md` and the other documents. A second AI agent acting as coordinator performed an adversarial review of the draft, and its corrections were applied by the first agent after checking them against the data.
- **Checks:** wherever a document says something was checked (the code and tests, the bot registry and verdict rules, the network-type classification of each AS, the quotations and their sources, the figures in this paper, and the anonymisation), the check was made by an AI agent.

References

- OpenAI, "Overview of OpenAI Crawlers", <https://platform.openai.com/docs/bots> (accessed 2026-10-02).
- 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> (accessed 2026-10-02); list at <https://claude.com/crawling/bots.json>.
- Perplexity, "Perplexity Crawlers", <https://docs.perplexity.ai/guides/bots> (accessed 2026-10-02).

- Google, "Google's common crawlers", <https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers>, and "Verify requests from Google crawlers and fetchers", <https://developers.google.com/search/docs/crawling-indexing/verifying-googlebot> (accessed 2026-10-02).
- Microsoft Bing, "Verify Bingbot", <https://www.bing.com/webmasters/help/how-to-verify-bingbot-3905dc26> (page rendered client-side; no text retrieved), and range list <https://www.bing.com/toolbox/bingbot.json> (accessed 2026-10-02).
- Apple, "About Applebot", <https://support.apple.com/en-us/119829> (accessed 2026-10-02).
- Common Crawl, "CCBot", <https://commoncrawl.org/ccbot> (accessed 2026-10-02).
- DuckDuckGo, "DuckDuckBot", <https://duckduckgo.com/duckduckgo-help-pages/results/duckduckbot> (accessed 2026-10-02).
- Mistral AI, "Robots", <https://docs.mistral.ai/robots/> (accessed 2026-10-02).
- Yandex, "How to check that a robot belongs to Yandex", <https://yandex.com/support/webmaster/robot-workings/check-yandex-robots.html> (accessed 2026-10-02).
- Amazon, "Amazonbot", <https://developer.amazon.com/amazonbot> (accessed 2026-10-02).
- Meta, "Meta Web Crawlers", <https://developers.facebook.com/docs/sharing/webmasters/web-crawlers/> (accessed 2026-10-02; IP list not retrievable as text).
- HUMAN Security, "Controlling AI-driven content scraping with HUMAN", 2026-02-28, <https://www.humansecurity.com/learn/blog/controlling-ai-driven-content-scraping-with-human/>.
- D. Forrester, "81.8% Of My 'AI Assistant' Traffic Was Fake. The Googlebot Number Was Worse", Search Engine Journal, 2026-06-25, <https://www.searchenginejournal.com/my-ai-assistant-traffic-was-fake-the-googlebot-number-was-worse/580052/>.
- G. Corral, V. Singhal, B. Mitchell, R. Tatoris, "Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives", Cloudflare blog, 2025-08-04, <https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/>.
- Team Cymru, IP to ASN mapping service, whois.cymru.com.

Cite this study

CITATION

EasyxLab (2026). Spoofed AI crawlers. Study S2. EasyByte Hub S. Coop. Mad. <https://github.com/easybytehub/easyxlab/tree/main/studies/s2-spoofed-ai-crawlers>

BIBTEX

```
@techreport{easyxlab_s2,  
  title      = {Spoofed AI crawlers},  
  author     = {{EasyxLab}},  
  institution = {EasyByte Hub S. Coop. Mad.},  
  number     = {S2},  
  year       = {2026},  
  url        = {https://github.com/easybytehub/easyxlab/tree/main/studies/s2-  
spoofed-ai-crawlers}  
}
```