Citeme

CITEME RESEARCH · SEPTEMBER 2026

What is actually crawling you

385,057 AI-agent requests across twenty sites. Seven per cent of them are not AI agents at all, the biggest crawler has never read your robots.txt, and no efficiency metric ranks the engines the same way twice.

7.0%
of “AI crawler” traffic hits credential paths

0
robots.txt fetches by Meta in 71,547 requests

3,228
human referrals attributed to an engine

INSIDE

Maxime Mansiet, Mael Bourdin, Alex Teillet

SUMMARY

We logged every AI-agent request to twenty sites for eight months and classified each one by what the agent was doing: bulk-crawling, fetching a page while answering somebody, or sending a person to the site.

The headline is not about any engine's behaviour. It is that **a measurable share of what every GEO tool reports as AI crawler traffic is not an AI crawler.** Seven per cent of our corpus consists of requests for `.env`, `.git`, `terraform.tfstate` and similar, carrying AI user agents. On a fixed set of sites that share rose from 3.3% in July to 18.1% in August. Every per-engine volume in this report, and in every comparable report, is inflated by an unknown amount.

What survives that caveat is still worth knowing: the engine that crawls most returns no measurable humans, the ranking of engines by efficiency changes completely depending on what you divide by, and one large engine has never once read a robots.txt in our logs.

WHAT WE MEASURED

Four kinds of visit

Every request is matched against a user-agent and IP-range signature table, then recorded as one of four types. That split is the whole instrument. A single “AI bot visits” figure mixes a crawler that will never return with a person reading the page right now.

- **Crawl.** Bulk indexing, independent of any question. `GPTBot`, `meta-externalagent`, `bingbot`.
- **Browsing.** A fetch made while answering somebody. `ChatGPT-User`, `Claude-User`, `MistralAI-User`.
- **Search.** Retrieval-time fetch by a search subsystem. `OAI-SearchBot`, `PerplexityBot`.
- **Referral.** A **human** arriving from an AI answer, identified by referring platform rather than bot signature.

256,375
crawls

125,454
answer-time fetches
browsing plus search

3,228
human referrals

READ THIS BEFORE THE NUMBERS

This is not a sample of the web. Twenty sites that bought a generative-engine-optimisation product, skewed toward organisations that already suspected AI visibility mattered to them. No single site exceeds 30% of visits, and the top five are 89% of rows.

The corpus is weighted to the end of the window. Reporting sites grew from one in January to eighteen in August, so about three-quarters of visits fall in July, August and September. Every claim about change over time holds the site population fixed by restricting to a cohort, and says so.

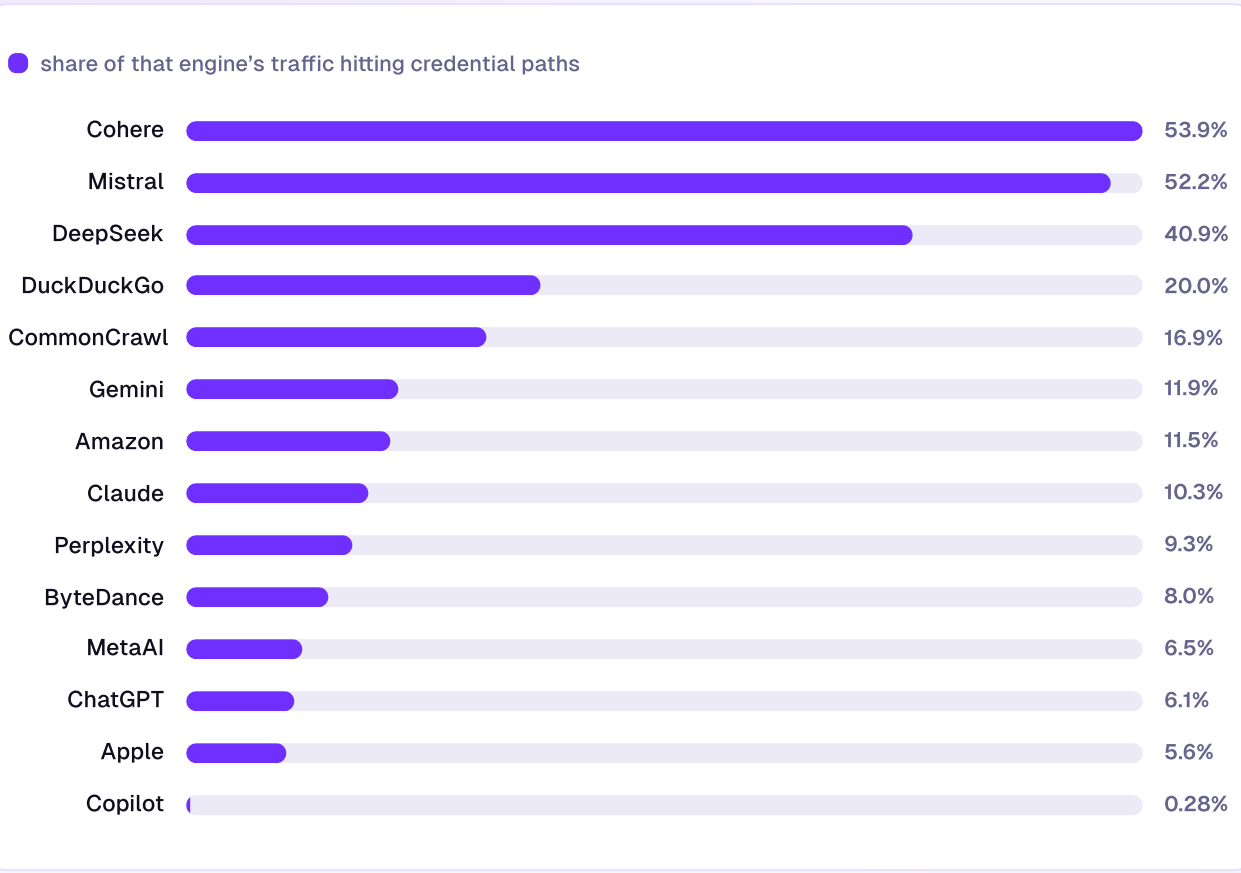
Aggregates only. No site is identified and no raw path appears in this report.

FINDING ONE

Seven per cent of it is not a crawler

Requests recorded as AI-agent traffic include a large and growing share aimed at paths no AI crawler would ever want: `.env`, `.git`, `.aws`, `.ssh`, `terraform.tfstate`, `secrets.yml`, `server.key`, `/actuator` and the Vite path-traversal probe `/@fs`. This is credential harvesting, and it is arriving under the names of named engines.

26,287 of 377,382 requests, 7.0%, across twelve sites, are credential probes. The engine with the lowest probe share is the one bot in the corpus that is cheaply verifiable by reverse DNS.



Probe share by engine. Copilot, which is bingbot and verifiable by reverse DNS, sits at 0.28%. The engines at the top are those whose user-agent strings are easy to forge and whose IP ranges are not published. That ordering is the signature of spoofing.

It is growing, and not because our corpus grew

On a fixed cohort of six sites present throughout, probe share by month was **3.34% in July, 18.09% in August and 15.50% in September**, over 55,073, 78,648 and 35,993 requests respectively. The site population is constant across that series, so this is not an artefact of new sites joining.

FINDING ONE, CORROBORATED

A second signal finds the same thing

If these requests were genuine engine traffic, their timing would look like the rest of that engine's traffic. It does not. Across the fifteen providers with at least 500 requests, an engine's probe share and its weekday share correlate at **Spearman −0.77**. Two independent signals are flagging the same streams.

Under a process indifferent to human calendars, 71.4% of requests would fall on a weekday. Mistral sits at 24.0% weekday. Removing every probe path moves it only to 29.3%, and its browsing traffic is 27.2% Paris-daytime against a 58.3% flat baseline, which is actively nocturnal.

Stripping the probe paths does not restore a normal weekly pattern. Whatever is forging these identities is also fetching ordinary content pages under them, so the probe regex catches only the visible tip and the true contamination is larger than 7%.

WHAT WE CANNOT TELL YOU

We cannot distinguish scanners forging AI user agents from a misattribution bug in our own classifier, because `user_agent` is null on 99.2% of server-log rows. The published consequence is identical either way, that per-engine volumes are overstated, but the mechanism is undetermined.

Reverse-DNS or published-IP-range verification on a sample of these requests would settle it, and is the first thing we intend to do. Until then, treat every per-engine volume in this report as an upper bound, ours included.

FINDING TWO

Meta has never read your robots.txt

Restricted to the twelve sites that demonstrably record robots.txt fetches, engines differ by three orders of magnitude in how often they check it.

ENGINE	REQUESTS	ROBOTS.TXT	CONTENT PER ROBOTS.TXT
MetaAI	71,547	0	never
Gemini	13,022	21	546
Amazon	14,112	124	99
Copilot	66,677	1,422	40
ChatGPT	90,970	5,705	14
Claude	34,188	10,088	1.3

Meta made 71,547 requests across twelve sites and 152 active days and fetched robots.txt zero times, while fetching `sitemap.xml` 3,248 times. At the other end, Claude-SearchBot fetches robots.txt more often than it fetches content.

THE OBVIOUS OBJECTION, RULED OUT

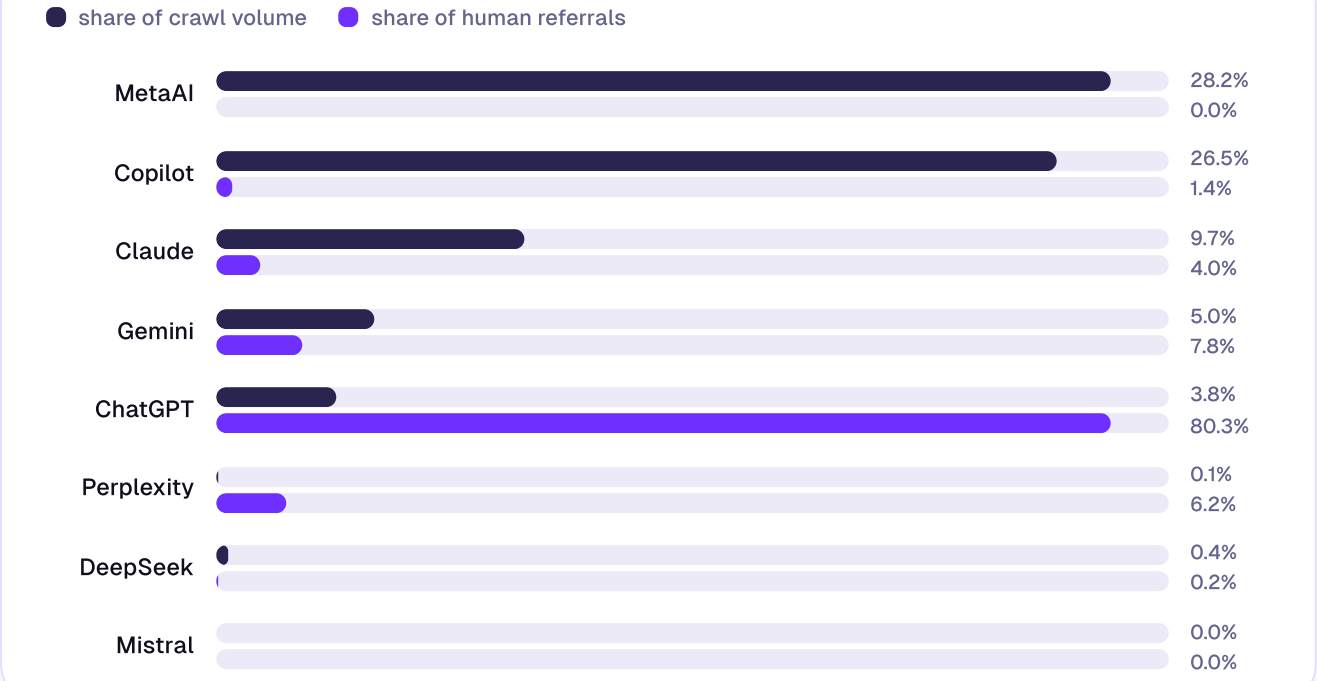
These sites do log robots.txt. Every one of the twelve recorded robots.txt fetches from other engines, between 611 and 7,366 per site, on the same pipeline in the same period. Meta is present on all twelve and fetched it on none.

Two things temper it. If Meta fetches robots.txt from infrastructure our classifier does not label as Meta, it would be invisible here, and we cannot rule that out. And OpenAI's split is expected rather than scandalous: `OAI-SearchBot` fetches robots.txt every 1.3 content fetches while `ChatGPT-User` does so once every 3,983, because OpenAI documents user-directed fetches as exempt. Meta's zero is not that.

FINDING THREE

The biggest crawler returned nobody

Meta is 28% of all crawl volume here, and 38% among engines whose referrals we can attribute, with zero referrals recorded. Perplexity is 0.1% of crawl volume and 6% of referrals.



All eight attributable engines, ranked by crawl volume. The bars are broadly inverted at the top of the range and unrelated at the bottom.

Meta's crawling stepped up in July

Raw monthly totals would measure our own growth, so this holds the site population fixed to sites already reporting before April, read as per-site averages.

MONTH	SITES	META CRAWLS PER SITE
March 2026	3	363
April 2026	2	147
May 2026	2	0
June 2026	2	15
July 2026	2	1,943
August 2026	2	2,331
September 2026	2	1,963

That cohort is two sites, and May reads zero where April and June do not, which is more likely a collection gap than a crawler that paused. The `meta-externalagent` signature entered our detector on 6 March 2026, so the July step is not the date we learned to see it. Read the shape, not the magnitude.

FINDING FOUR

Four denominators, four winners

These are the same 3,228 referrals every time, divided four ways.

ENGINE	CRAWL	BROWS.	SEARCH	REFS	/1K CRAWLS	/1K FETCHES	/1K ANS.-TIME
MetaAI	72,207	0	1,748	0	0.00	0.00	0.0
Copilot	67,852	0	43	45	0.66	0.66	1047
Claude	24,832	5,144	5,210	130	5.24	3.69	12.6
Gemini	12,769	331	662	253	19.81	18.38	254.8
ChatGPT	9,605	66,510	15,039	2,591	269.76	28.42	31.8
DeepSeek	1,015	0	0	7	6.90	6.90	n/a
Perplexity	376	2,362	20,312	199	529.26	8.63	8.78
Mistral	17	4,155	0	0	0.00	0.00	0.0

Per thousand crawls, Perplexity looks 800 times better than Copilot. Per thousand total fetches, ChatGPT leads and Perplexity drops to fourth. Per thousand answer-time fetches, Gemini leads by eight times over ChatGPT, and Copilot produces a meaningless 1,047 from 43 fetches.

And a fourth: pages actually covered

Volume and coverage are different leaderboards. On the six-site cohort, counting distinct paths reached rather than requests made:

ENGINE	FETCHES	DISTINCT PATHS	FETCHES PER PATH
MetaAI	19,603	1,009	19.4
ByteDance	10,366	713	14.5
Apple	16,547	1,458	11.3
Copilot	21,536	2,022	10.7
Gemini	7,094	875	8.1
Claude	6,676	1,754	3.8
ChatGPT	4,931	1,469	3.4
CommonCrawl	844	482	1.8

On one site over the same period, Meta reached 451 distinct paths from 13,406 fetches and Claude reached 455 from 1,067. Identical coverage for 12.6 times the requests.

The ranking is a property of the denominator, not of the engines. Any report, ours included, that publishes one of these as an efficiency league table is publishing an artefact. What survives all four: the two largest crawlers, 55% of crawl volume between them, returned 45 people in total.

FINDING FIVE

Answering is not crawling

What an engine fetches while answering a question is structurally different from what it crawls in bulk, and the difference is homepage concentration.

VISIT TYPE	REQUESTS	MEAN PATH DEPTH	HOMEPAGE SHARE
crawl	180,581	0.99	7.3%
search	28,101	0.79	6.2%
browsing	69,247	**0.20**	**37.4%**
referral	3,229	1.08	11.8%

Paired within site, browsing homepage share exceeds crawl homepage share in twelve of twelve sites with at least 100 requests in both arms. Sign test p = 0.00024. The browsing arm is 85% ChatGPT, so this is substantially a statement about `ChatGPT-User`, and much of it is probably people pasting a bare domain into a chat box rather than any editorial preference.

Engines have different appetites, and pace very differently

- **Copilot is the editorial crawler**, 37.9% blog. **Meta is the catalogue crawler**, 45.6% deep paths and 12.5% commercial pages but 0.9% glossary. **Apple is the shallowest**, 26.5% homepage.
- **Perplexity is the reference reader.** Its search fetches are 53.0% glossary, docs and resources, and 3.0% homepage, almost the exact inverse of ChatGPT browsing at 40.1% homepage and 1.3% commercial.
- **Pacing spans 280 times.** Median gap between consecutive requests: Apple 340s, Copilot 170s, Claude 12.4s, Meta 4.5s, ChatGPT 1.2s. ChatGPT sends 71.4% of its requests within five seconds of the previous one, Copilot 14.4%.

Shape mix is partly a property of which sites an engine visited rather than the engine itself. Meta appears on five of six cohort sites and Copilot on six, so those are comparable. ByteDance and CommonCrawl appear on three each and their rows are weak. Perplexity's crawl pacing rests on 373 requests across two sites and is not reported.

FINDING SIX

Citation rates reverse when you control the prompts

We also run tracked prompts against nine providers and record whether a site was cited. Read naively across all 4,051 results, OpenAI cites at 24.4% and ranks second. Restricted to the 59 prompt and site pairs where **all nine providers were actually run**, OpenAI cites at 1.6% and ranks last.

OpenAI swings from 50.2% to 1.6% depending only on whether the prompt was part of a full sweep. That is a thirty-one-fold selection effect, and it moves one engine from second place to last without any engine behaviour changing.

Any “which engine cites you most” figure computed over an unbalanced set of prompts is measuring which prompts were chosen, not how engines behave. This is the same class of error as the denominator problem above, and it is the reason we are not publishing an engine citation ranking.

The fully paired arm is 1,407 results across 59 prompt and site pairs and 8 sites. OpenAI's 1.6% rests on 3 citations from 190 results, so that point estimate is fragile. The contrast with 124 citations from 247 results is not. If the full-sweep prompts happened to be harder or more generic than the rest, prompt difficulty explains the effect entirely, and we could not test that without reading the prompt text. Either way the warning stands.

WHAT IT MEANS

Count the people, and verify the bots

- **Verify before you attribute.** If you publish per-engine AI traffic without reverse-DNS or IP-range verification, some unknown fraction of what you are reporting is a scanner. The engines easiest to forge are the ones that look most active.
- **A single AI-visits number is not informative.** It is dominated by bulk crawl, 67% of requests here, which is the component least related to anything a person does.
- **Do not replace it with a per-fetch ratio either.** Four defensible denominators give four different winners. If you want to know whether an engine sends you people, count the people.
- **Referrals are a floor, not a census.** 3,228 across eight months is a small number, and the mechanism producing it is the referer header, which is fragile.

We are not recommending that anyone block Meta or Copilot, and this data cannot support that recommendation. An engine that returns no measurable clicks may still shape how a model describes you, and that effect is real, commercially significant, and completely invisible to everything measured here.

LIMITATIONS

What this data cannot tell you

We cannot verify bot identity. `user_agent` is null on 99.2% of server-log rows, so findings one and two rest on path patterns and timing rather than on confirmed identity. This is the most important limitation in the report.

Five engines cannot be measured for referrals and are excluded from that comparison: Apple (30,613 crawls), ByteDance (19,058), Amazon (14,391), CommonCrawl (2,775), Cohere (687). They are absent from the referral detector's platform table, so we cannot distinguish an engine that sends no traffic from one we cannot attribute.

Meta's zero referrals is not necessarily an absence of traffic. Meta's assistant is largely rendered inside the WhatsApp, Instagram and Facebook in-app browsers, among the worst environments for referer retention. We can say Meta returned no attributable referrals. We cannot say it returned no people.

robots.txt and probe paths are visible only in server-log ingestion, which covers thirteen of the twenty sites. The seven sites using the JavaScript beacon alone structurally cannot report either, so those findings are restricted to the sites that can.

A crawl row is not a page. Visits are deduplicated per bot name per session, so the crawl column counts crawl sessions, not bytes or pages. No bandwidth claim can be made from it and we make none.

No query-level analysis is possible. An inferred query, derived from the engine's own request rather than anything a user typed, is present on 1,881 of 385,057 visits.