

A Third of the Top Million Websites Could Go Dark to AI on September 15

Eric Lancheres · On-Page.ai Research | August 17, 2026

ABSTRACT

On September 15, 2026, Cloudflare begins blocking AI training and agent crawlers by default for new domains, newly added sites, and Free-tier accounts with untouched settings. To establish what that default will land on, we crawled a stratified panel of 10,600 domains sampled from the Tranco top million on August 17 and got a usable response from 10,216. We recorded each site's robots.txt policy for 24 AI crawler user agents, checked llms.txt, and detected Cloudflare. Weighted to the top million: 44.4% of sites run on Cloudflare, and 33.0% (about 330,000 sites) sit in the exposed pool, on Cloudflare with no AI training crawler blocked today. 15.9% of sites block at least one AI training crawler, double the rate for agent crawlers (7.6%), and 14.0% block AI training while leaving Googlebot untouched. llms.txt adoption reached 10.5%, with about one adopter in six being a Shopify storefront serving a platform-generated file. 5.8% of sites serve a Cloudflare-managed robots.txt and 6.1% carry Content-Signal directives. We re-crawl the same frozen panel on September 8, 16, 22, and October 1, and will publish exactly what the flip changed.

Key findings

- 33.0% of the top million websites (about 330,000 sites) sit in the pool the new default can touch: they run on Cloudflare and do not block a single AI training crawler today.
- llms.txt adoption reached 10.5% of the top million. About one adopter in six is a Shopify storefront serving a platform-generated file, and llms-full.txt (the version that carries actual content) trails at 4.4%.

- The web blocks training, not assistants. 15.9% of sites block at least one AI training crawler, twice the rate for AI agent crawlers (7.6%). And 14.0% block AI training while leaving Googlebot untouched.

1. What changes on September 15

On September 15, Cloudflare changes what happens when an AI crawler visits a website. Bots it classifies as Training or Agent will be blocked by default on ad-bearing pages for new domains, for sites added by existing customers, and for Free-tier accounts that never touched their bot settings. Search crawlers stay allowed. Nobody has to click anything. That is what makes it interesting: the largest single change to AI's access to the web will happen mostly to sites that never made a decision.

There is a second change buried in the documentation that deserves more attention than it has received. Cloudflare classifies mixed-purpose crawlers (those used for both search and training) under training: per its docs, every configuration that blocks AI training also blocks them, and that can include Googlebot on those pages unless the site explicitly opts out. A site that reflexively blocks “AI training” without reading the fine print can end up gating the crawler its search traffic depends on. We could not measure this from outside, so we flag it as a documented behavior, not a finding. It is worth reading Cloudflare's own docs before September 15 if you run ads and touch these settings.

2. The exposed pool: 330,000 sites

44.4% of the top million websites run on Cloudflare (we detected this via nameserver delegation and response headers). Most of them have no AI-crawling opinion on record: 33.0% of the top million, about 330,000 sites, are on Cloudflare and do not block even one AI training crawler in robots.txt.

That 330,000 is an upper bound on the directly affected population, and we want to be precise about why. The default flip only applies to Free-tier accounts with untouched settings and to newly added sites. Plan tier is not visible from the outside, and the exposed pool includes enterprise customers whose settings will not move (openai.com, paypal.com, and zoom.us are all in it). What we can say: this is the pool the default can reach, the flip will land somewhere inside it, and our September re-crawls will measure exactly how much of it moved.

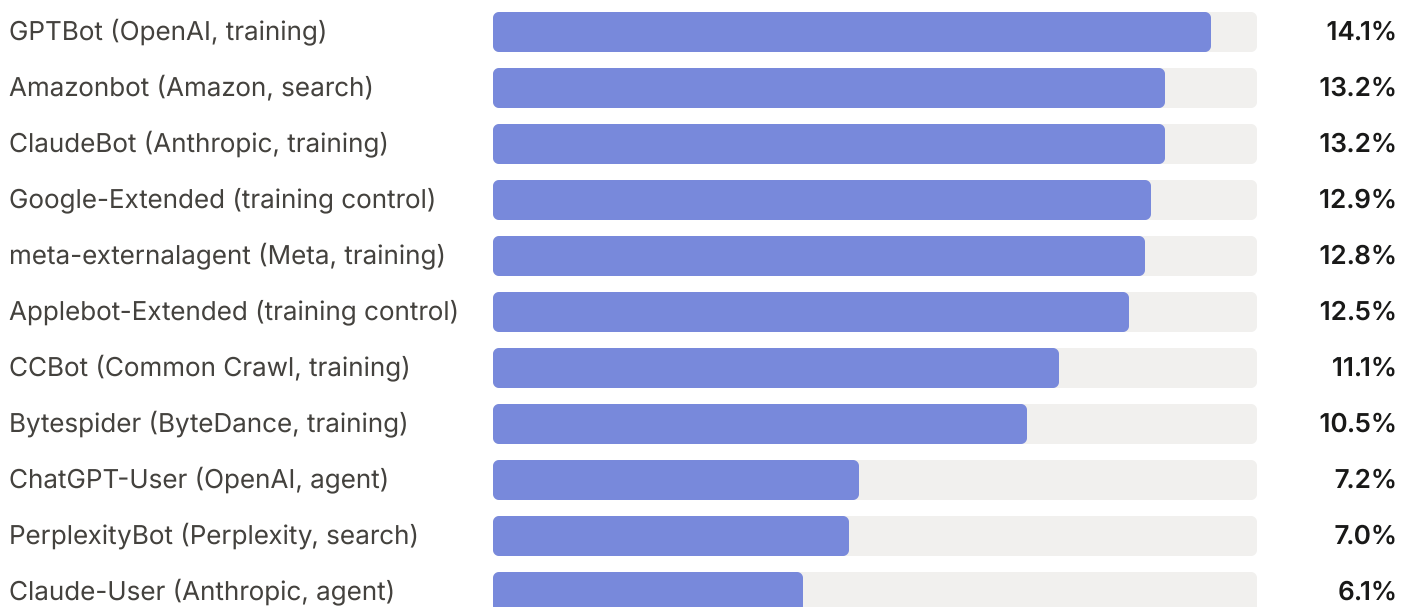
Rank band	On Cloudflare	In the exposed pool
Top 1,000	20.6%	16.5%
1k–10k	33.0%	26.5%
10k–100k	42.7%	33.5%
100k–1M	44.7%	33.0%
Top 1M (weighted)	44.4%	33.0%

Cloudflare footprint and the exposed pool by Tranco rank band. Exposed pool: on Cloudflare, blocking no AI training crawler in robots.txt. Baseline wave, August 17, 2026.

The top of the web is the least exposed: big sites use Cloudflare less, block AI more, and sit on paid plans. The middle and long tail is where the default will actually bite. If AI companies lose access to a meaningful slice of the web next month, it will be the part of the web that never voted.

3. Who blocks AI crawlers today

Across the top million, weighted by rank band, the share of sites whose robots.txt fully blocks each crawler from the site root:



Share of the top million websites blocking each crawler at the site root in robots.txt, weighted by rank band. Baseline wave, August 17, 2026.

Three patterns stand out. First, the web's revealed preference matches the direction Cloudflare is pushing, but only halfway. Sites block training crawlers at roughly double the rate of agent crawlers: 15.9% block at least one training bot, 7.6% block at least one agent bot. Site owners who made a deliberate choice mostly chose to keep assistants and shut out training. Cloudflare's new default blocks both.

Second, blocking AI does not mean blocking search. 14.0% of the top million block at least one AI training crawler while leaving Googlebot fully allowed. Among the top 1,000, the training blockers include nytimes.com, cnn.com, forbes.com, bbc.com, amazon.com, and x.com. A detail we enjoyed: chatgpt.com blocks other companies' training crawlers and allows Googlebot, the same posture as the publishers it trains on.

Third, in the top 1,000, the most-blocked crawlers are not OpenAI's. CCBot (13.1%) and Bytespider (12.7%) outrank GPTBot (11.7%) there. Big publishers started blocking Common Crawl and ByteDance before the current wave of AI-specific tokens existed, and those rules stayed.

4. llms.txt reached 10% of the web, and Shopify did a sixth of it

llms.txt, the proposed standard for telling AI systems what a site is about, was a fringe signal a year ago: Ahrefs, monitoring more than 3 million sites, counted an 8.8x increase in llms.txt files between June 2025 and May 2026. Our panel puts adoption at 10.5% of the top million as of August 17, with 9.2% serving a file that follows the spec (a markdown document opening with an H1 heading). The head of the web tracks close to other recent measurements: Rankability put the top 1,000 at 8.7% in June 2026; we measured 11.4% in August. Adopters in the top 1,000 include github.com, cloudflare.com, adobe.com, samsung.com, dropbox.com, and godaddy.com.

Rank band	llms.txt present	Spec-conformant
Top 1,000	11.4%	10.4%
1k–10k	9.0%	8.0%
10k–100k	10.1%	8.8%
100k–1M	10.6%	9.3%
Top 1M (weighted)	10.5%	9.2%

llms.txt adoption by rank band. Spec-conformant: markdown file opening with an H1 heading. Baseline wave, August 17, 2026.

Two things inflate the raw number, and we measured both. First, 15.7% of all valid llms.txt files we found are near-identical “Agent Instructions” documents on Shopify storefronts (kia.com, dollarshaveclub.com, princesspolly.com, wyzecam.com). One platform decision moved global adoption by more than a point and a half. Second, llms-full.txt, the variant that actually inlines a site's content, sits at 4.4%: adoption of the cheap gesture runs more than double the adoption of the expensive one.

We measured presence, not effect. Whether AI systems read these files is a separate question, and the best available evidence says mostly no: in Ahrefs's study, 97% of llms.txt files received zero requests from AI crawlers. What our number does tell you: publishing an llms.txt stopped being a fringe signal and became a mainstream default, and a large share of that happened without site owners doing anything.

5. Cloudflare is already rewriting robots.txt at scale

The September 15 default is edge enforcement, but Cloudflare also operates at the policy layer, and we can see it in the files. 5.8% of the top million serve a robots.txt containing Cloudflare's managed block (delimited by its “BEGIN Cloudflare Managed content” marker), and 6.1% carry Content-Signal directives, the machine-readable policy lines Cloudflare introduced in 2025 (for example: search=yes, ai-train=no). Adopters down the rankings include nvidia.com, roblox.com, and avast.com.

Put differently: about one site in seventeen already has its AI policy written by its CDN rather than its webmaster. Whatever you think of the defaults, the delegation is the story. The entity that terminates the connection increasingly writes the policy too, and on September 15 it starts enforcing a new one.

6. What to check before September 15

If you run a website, three checks take ten minutes:

1. Find out if you are on Cloudflare, and on which plan. Free-tier sites that never touched Security settings, and any domain added after September 15, get the new defaults. If you want AI training or agent crawlers to reach you (or want them blocked everywhere, not only on ad pages), set it explicitly instead of inheriting the default.
2. Read your own robots.txt. 35% of reachable sites in our panel do not serve a valid one at all, and a managed block may be in yours without you having written it. Confirm the file says what you mean.
3. If you block “AI training” in Cloudflare and run ads, read Cloudflare's note on mixed-purpose crawlers before September 15. Blocking training can gate crawlers your search traffic depends on unless you opt out explicitly.

7. Method and limitations

We sampled 10,600 domains from the Tranco top million (list dated August 16, 2026): all of the top 1,000, plus seeded random samples of 2,650 from ranks 1,001 to 10,000, 3,200 from 10,001 to 100,000, and 3,750 from 100,001 to 1,000,000. Top-million figures weight each rank band by its population share. 10,216 domains (96.4%) returned a usable HTTP response and form the denominator for all rates. The panel is frozen: every future wave crawls the same domains.

For each domain we fetched robots.txt, llms.txt, llms-full.txt, and the homepage. robots.txt was evaluated per RFC 9309: a crawler counts as blocked only if the effective rule for the site root is a disallow, either in a group naming that crawler or inherited from the wildcard group. We evaluated 24 AI user-agent tokens spanning OpenAI, Anthropic, Google, Meta, Apple, Amazon, Perplexity, ByteDance, Common Crawl, DeepSeek, Mistral, Cohere, Ai2, DuckDuckGo, and Timpi, with Googlebot and Bingbot as non-AI references. llms.txt hits were validated against soft 404s (HTML, binary, and non-text responses rejected). Cloudflare detection used nameserver delegation and response headers together.

We also sent homepage requests identifying as GPTBot, ClaudeBot, and PerplexityBot. These probes are not verified crawler traffic (verification is by IP range, and we are not those companies), so we treat their absolute block rates as context, not findings, and will use their wave-over-wave deltas,

measured with an identical instrument, in the follow-up study. All requests were routed through a fixed pool of datacenter egress IPs, kept identical across waves.

Limitations: robots.txt is declared intent, not enforcement. Cloudflare plan tier is not observable from outside, so the exposed pool is an upper bound on the population the default flip will move. This is a single-day snapshot; the September waves exist to measure change.

References

- Cloudflare changelog, "New options to manage AI traffic" (July 1, 2026)
- Cloudflare blog, "Your site, your rules: new AI traffic options for all customers"
- Cloudflare docs, "Block AI bots" (September 15 default behavior and the mixed-purpose crawler note)
- Tranco top sites ranking (tranco-list.eu), list dated August 16, 2026
- Ahrefs, "We Analyzed 137K Sites: 97% of llms.txt Files Never Get Read" (May 2026)
- Rankability llms.txt adoption tracker (June 2026)
- OpenAI, Anthropic, and Perplexity crawler documentation (user agent tokens)

We re-crawl the same 10,600-domain panel on September 8, 16, 22, and October 1, and will publish exactly what the flip changed: how much of the exposed pool went dark, whose robots.txt got rewritten, and whether the web's AI-access posture moved by decision or by default. Questions about the data: team@on-page.ai.

Cite this study: Lancheres, E. (2026). The AI Access Census: A Third of the Top Million Websites Could Go Dark to AI on September 15. On-Page.ai Research. <https://api.on-page.ai/research/ai-access-census>