

What Cloudflare's September 15 AI Default Actually Changed: A 10,600-Site Before-and-After

Eric Lancheres · On-Page.ai Research | October 7, 2026 | Part 2 of [the AI Access Census](#)

In August we published a number: a third of the top million websites, about 330,000, sat in the pool Cloudflare's September 15 AI-crawling default could touch. We called it an upper bound at the time and promised to re-crawl the same sites afterward to see what actually happened. This is the follow-up, and the short answer is that the flip landed well inside that bound.

We crawled the same frozen panel of 10,600 domains five times: August 17 (baseline), September 8 (one week before the flip), September 16 (the day after), September 22, and October 1. Every wave used the same domains and the same requests, run by the same code, so the differences between waves are differences in the web, not in the instrument.

ABSTRACT

On September 15, 2026, Cloudflare began blocking AI training and agent crawlers by default for new domains and unchanged Free-tier accounts, while allowing crawlers it classifies as Search. We measured the change on a frozen, rank-stratified panel of 10,600 domains from the Tranco top million, crawled on August 17, September 8, September 16, September 22, and October 1 (10,194 to 10,216 reachable per wave), recording each site's robots.txt policy for 24 AI user agents, llms.txt, Cloudflare presence, and the edge response to requests identifying as GPTBot, ClaudeBot, and PerplexityBot against a browser control. Weighted to the top million, the edge-aware exposed share (on Cloudflare, no training block in robots.txt, GPTBot not blocked at the edge) moved from 24.96% to 25.50%; every metric changed in a single step at the flip and was flat for the following 16 days. The largest changes were Cloudflare's: its legacy managed robots.txt text disappeared from 5.96% of the top million to 0.07% (282 of 282 pairable carriers), taking Content-Signal adoption from 6.40% to 0.56% with it, while the edge enforcement behind it remained. PerplexityBot, classified as Search, went from edge-

blocked on 5.47% of sites to 1.24%; the share blocking GPTBot while serving PerplexityBot rose from 0.98% to 6.03%, roughly 50,000 sites. GPTBot edge blocks rose by less than a point (72 new blocks in the flip window against 23 in a 22-day noise window). Cloudflare's replacement managed robots.txt, including a template that blocks agent crawlers, is growing from near zero (0.07% to 0.27%). llms.txt adoption rose from 10.54% to 11.01% independently of the flip.

Key findings

- Counting what Cloudflare's edge actually blocks, the exposed share of the top million moved from 24.96% to 25.50% over the whole study. The August estimate of a third was an upper bound, and the flip landed well inside it: one step on September 15, then no further movement in 16 days.
- Cloudflare removed its legacy managed robots.txt text from 5% of the web overnight. On September 8, 5.96% of the top million carried it; on September 16, 0.07% did, while the edge enforcement behind it stayed in place. Since then, a Cloudflare site's robots.txt is an unreliable guide to what its edge actually enforces.
- The flip opened the web to search crawlers more than it closed it to training crawlers. PerplexityBot went from edge-blocked on 5.47% of the top million to 1.24%. The share of sites blocking GPTBot at the edge while letting PerplexityBot through jumped from 0.98% to 6.03%, roughly 50,000 sites.

1. What we said in August, and what happened

Our August 17 baseline counted 33.0% of the top million as exposed: on Cloudflare, with no AI training crawler blocked in robots.txt. We called that an upper bound, because the new default applied only to Free-tier accounts with untouched settings and to newly added domains, and plan tier is invisible from outside.

A site has gone dark to AI when Cloudflare's edge refuses a request identifying as an AI crawler while serving the same page to a browser, whatever its robots.txt says. We measured exposure that way for this report: on Cloudflare, no training block in robots.txt, and GPTBot not blocked at the edge. By that definition the exposed pool was 24.96% at baseline and 25.50% on October 1.

Top-1M weighted	Aug 17	Sept 8	Sept 16	Sept 22	Oct 1
On Cloudflare	44.42%	44.49%	44.66%	44.72%	44.73%
Exposed, edge-aware (on CF, no robots training block, GPTBot not edge-blocked)	24.96%	25.09%	25.60%	25.58%	25.50%
Protected either way (robots training block or GPTBot edge block)	19.46%	19.41%	19.07%	19.14%	19.23%
"At risk" as defined in August (robots.txt only)	32.99%	32.75%	38.45%	38.39%	38.22%

Exposure under both definitions across all five waves, top-1M weighted. The August definition jumps at the flip because Cloudflare deleted its managed robots.txt text, not because sites opened up (section 3).

The last row is the definition we published in August, and it moved up 5.5 points at the flip, which would read as 55,000 sites opening up to AI if you took it at face value. The jump comes from Cloudflare deleting its own text from those robots.txt files (section 3), so a robots-only definition of exposure stopped working on September 15. We count that as one of the findings.

We also followed the August cohort directly. Of the 3,062 panel domains in the at-risk pool on August 17, we re-observed 3,049 on October 1. 39 of them added a training-crawler block to robots.txt, 7 picked up a new Cloudflare-managed robots.txt, and 41 left Cloudflare. Among the 2,049 with a clean browser response in both waves, GPTBot edge blocks went from 223 to 224. PerplexityBot edge blocks in the same set fell from 180 to 49.

2. One step on September 15, then flat

Every weighted metric moved less than half a point between September 22 and October 1, and the domain-level changes in the last window sit at or below the pre-flip noise floor. In the data the flip looks like a single step rather than the start of a trend.

Paired domain flips (on / off)	Noise: Aug 17 to Sept 8 (22 days)	Flip: Sept 8 to 16 (8 days)	Sept 16 to 22	Sept 22 to Oct 1
GPTBot blocked at Cloudflare's edge	23 / 50	72 / 20	8 / 21	9 / 19
PerplexityBot blocked at Cloudflare's edge	20 / 25	7 / 317	6 / 3	4 / 6
GPTBot blocked, PerplexityBot allowed (split)	4 / 29	367 / 7	7 / 23	9 / 18
All three bots blocked at the edge	20 / 23	7 / 316	4 / 2	4 / 5
Cloudflare-managed robots.txt present	21 / 12	3 / 282	5 / 0	7 / 4

Domain-level changes between consecutive waves, counted only on domains observable in both waves. Read the flip window against the noise window on its left.

Against the noise window, the 72 new GPTBot edge blocks in the flip window are a real change, though a small one. The large movements went the other way: 317 PerplexityBot edge blocks disappeared (25 did in the noise window) and 282 managed robots.txt files were removed (12 in the noise window).

3. Cloudflare removed its managed robots.txt from 5% of the web

For two years, Cloudflare's “Block AI bots” toggle did two things: it blocked AI crawlers at the edge, and it prepended a managed block to the site's robots.txt, delimited by a “BEGIN Cloudflare Managed content” marker, listing the blocked user agents and carrying Content-Signal lines such as ai-train=no. In August that block sat on 5.75% of the top million and on 12.9% of Cloudflare sites.

On September 16 it was on 0.07% of the top million. Among the sites we could pair with a valid robots.txt on both September 8 and September 16, 282 lost the block and 3 gained one; the noise window before had 12 and 21. The rest of each file was unchanged byte for byte. The legacy template, present on 459 panel domains on September 8, had not returned anywhere by October 1. Sites that

lost it include patreon.com, kick.com, vaticannews.va, aps.org, gsu.edu, airasia.com, documentcloud.org, standardnotes.com, nationalreview.com, and x.ai, whose per-bot Content-Signal lines (which we noted in August) turned out to be Cloudflare's text, not theirs.

Content-Signal adoption went with it: 6.40% of the top million on September 8, 0.56% on September 16, 0.87% on October 1. Nearly all of the Content-Signal directives we counted in August lived inside the managed block.

Cloudflare announced part of this. On August 21 it published “Say it once: Introducing Bot Preference Sync,” a replacement that writes robots.txt from the dashboard's Search, Agent, and Training settings so that, in Cloudflare's words, “what you say to the world and what you enforce at the edge are kept in sync.” The post says existing users of “the legacy managed robots.txt feature” would be prompted to review their preferences and transition. It does not say the legacy text would be removed from their files on September 15, and nothing we found in Cloudflare's changelog or docs says so either. The removal of the legacy text is the part nobody announced, and it's what we measured on September 16.

The enforcement stayed. Among Cloudflare sites with a clean browser response, the share blocking a request identifying as GPTBot at the edge was 14.16% on September 8 and 16.00% on September 16, so the sites that lost the text kept the block. Since September 15, a Cloudflare site's robots.txt says less about its AI policy than it did before, and section 7 covers what that means if you audit access.

4. The flip opened the web to search crawlers

Cloudflare's new taxonomy sorts crawlers into Search, Training, and Agent, with Search allowed by default. PerplexityBot is classified as Search. On September 15 that classification was applied at the edge, and the result is the largest single movement in the study.

Top-1M weighted, clean browser response	Sept 8	Sept 16	Oct 1
PerplexityBot blocked at Cloudflare's edge	5.47%	1.24%	1.30%
GPTBot blocked at Cloudflare's edge	6.31%	7.16%	7.04%
GPTBot blocked, PerplexityBot allowed (split treatment)	0.98%	6.03%	5.83%
GPTBot, ClaudeBot, and PerplexityBot all blocked	5.31%	1.09%	1.15%

Edge treatment of requests identifying as each crawler, among sites that served a 200 to the browser control in the same wave. Top-1M weighted.

On Cloudflare sites specifically, PerplexityBot edge blocks fell from 12.27% to 2.76%, and the split treatment rose from 2.19% to 13.48%. In domain terms, 367 panel sites went from blocking all three crawlers to blocking GPTBot and ClaudeBot while serving PerplexityBot, and 402 of the 437 we could re-observe were still split on October 1. Weighted to the top million, that is about 5 points, roughly 50,000 sites, that a search-classified AI crawler can now reach and a training-classified one cannot.

GPTBot moved the other way, slightly: 72 new edge blocks in the flip window against 23 in the noise window, and 62 of those 72 were still in place a week later (the ones that lapsed were rotating piracy and adult domains, not policy reversals). Net, GPTBot's edge-block rate rose less than a point and ClaudeBot's tracked it.

Taken together, September 15 moved the web toward a split policy. On the sites where Cloudflare sets the rules, crawlers that send traffic back now get in and crawlers that only collect training data stay out. PerplexityBot's Search classification is Cloudflare's call, and plenty of publishers will disagree with it; what we can report is that the classification took effect at the edge on September 15 and hasn't been rolled back.

5. The new managed robots.txt is growing from zero

Bot Preference Sync writes its own managed block, and we can read it. Two templates account for nearly all of it:

Managed template (by content hash)	Sept 16	Sept 22	Oct 1
Training crawlers only (33 user agents)	3	8	14
Training crawlers plus an "AI Agents" section (48 user agents)	0	5	8
Custom or other	5	5	5

Cloudflare-managed robots.txt templates by content hash, panel domain counts per wave.

Weighted, Cloudflare-managed robots.txt is back to 0.27% of the top million (0.61% of Cloudflare sites), from 0.07% the day after the flip. Content-Signal lines follow it, 0.56% to 0.87%.

The agents template is the one to watch. It blocks ChatGPT-User, Claude-User, Perplexity-User, Google-Agent, and NotebookLM alongside the training crawlers, while leaving search bots allowed. It is on 8 panel domains, including america.gov, and most of the apparent growth is zones becoming

observable (four of the five new ones served a 403 or 404 on robots.txt the week before) rather than confirmed new adoption. Eight domains is too few to call a rate, so we'll count it properly in a later wave.

The training-only template is being adopted by real switches: footballant.com, vitra.com, auto-tests.com, notice-facile.com, and importitall.co.za all went from a valid unmanaged robots.txt on September 22 to the managed one on October 1.

6. What site owners did on their own

Underneath the Cloudflare-driven changes, site owners made few changes of their own, and the ones they made went in both directions. We verified each of these against the raw robots.txt files.

Newly blocking since September 22: airbnb.com, airbnb.fr, and airbnb.be moved their GPTBot rules from path-level disallows to a full Disallow: / and added a ClaudeBot group. depositphotos.com added a GPTBot, ClaudeBot, and anthropic-ai group plus a Content-Signal ai-train=no line. scientificamerican.com, which already blocked GPTBot, now blocks ClaudeBot too. eluniversal.com.mx, meteo.ua, and astrobin.com added training blocks.

Opened up since August and still open: weather.com, sohu.com, justdial.com, and loopnet.com. At the edge, gitlab.com and gitlab.io dropped their GPTBot and ClaudeBot blocks (a 403 in August and September, a 200 for every user agent on October 1), and sportingnews.com dropped its GPTBot block.

llms.txt kept growing on its own schedule, 10.54% to 11.01% over 45 days: 123 panel sites gained a file and 39 lost one. Gains include etsy.com, intel.com, ubuntu.com, unity3d.com, intuit.com, tailscale.com, sephora.com, and zomato.com. Losses include azure.com, indeed.com, kia.com, and searchengineland.com. None of this correlates with the flip; it is the same slow adoption curve we saw in August.

7. What to do now

Four things changed on September 15 that affect how you should manage AI access, whether or not you run on Cloudflare.

1. If you relied on Cloudflare's "Block AI bots" toggle to publish your policy, your robots.txt no longer says what you think it says. The managed text is gone from every site we could pair that carried it on September 8, all 282 of them. Your edge rules are probably still enforcing, but a crawler reading your file sees no instruction. Either turn on Bot Preference Sync so the file is regenerated from your settings, or write the rules yourself.
2. Decide search versus training explicitly. The default now lets Search-classified crawlers, PerplexityBot included, through sites that block training. If that is not what you want, set Search to disallow rather than inheriting it. If it is what you want, know that roughly 50,000 sites in the top million made that trade on September 15 without choosing it.
3. Check what the edge does, not only what the file says. Since September 15, robots.txt on a Cloudflare site is a partial readout at best. Test what a request identifying as GPTBot or ClaudeBot actually gets back, or read the dashboard. On Cloudflare sites, the file overstated edge blocking before the flip (24.6% declared a GPTBot block in robots.txt, 14.2% enforced one at the edge) and understates it after (12.1% declared, 16.0% enforced).
4. The next default will probably cover agent crawlers. The template that blocks ChatGPT-User, Claude-User, and Perplexity-User already exists and is being switched on, so decide whether you want AI assistants fetching your pages on a user's behalf before that decision gets a default of its own.

We will keep the panel frozen and re-crawl it when Cloudflare's next change lands. The August study, with the baseline numbers, is at api.on-page.ai/research/ai-access-census.

8. Method and limitations

Panel: 10,600 domains sampled 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. The panel did not change between waves. Reachable domains per wave: 10,216 (Aug 17), 9,996 (Sept 8), 10,207 (Sept 16), 10,203 (Sept 22), 10,194 (Oct 1).

Per domain, per wave: robots.txt, llms.txt, llms-full.txt, and the homepage with a browser user agent and with user agents identifying as GPTBot, ClaudeBot, and PerplexityBot. robots.txt was evaluated per RFC 9309 for 24 AI user-agent tokens. "Blocked at the edge" means the browser request got a 200 and the crawler-identifying request got a Cloudflare challenge or a 403 or 503 served by

Cloudflare. “Edge-aware exposed” means on Cloudflare, no training crawler blocked in robots.txt, and GPTBot not blocked at the edge. Cloudflare detection used nameserver delegation and response headers together. Managed robots.txt was identified by Cloudflare's own BEGIN and END markers, and templates were distinguished by hashing the managed block.

The probes are not verified crawler traffic: verification is by IP range and we are not those companies. Every rate built on them is conditioned on the browser request succeeding, and we report probe deltas against that control. Across the five waves the browser-request block rate itself drifted up from 22.2% to 23.8% on the same exit IPs, consistent with slow reputation decay of a fixed datacenter egress pool; the control-conditioned metrics in this report are flat across that drift, and the raw probe rates are not quoted for that reason.

All requests were routed through a fixed pool of datacenter egress IPs: 30 distinct exits for the August 17 and September 8 waves, 27 for the three later waves, with a subset in common. Paired domain comparisons exclude any domain whose browser request failed in either wave. The September 8 wave lost about 220 domains to a proxy outage mid-run; they are absent from that wave's counts and from paired comparisons involving it, and are treated as missing observations, not as changes.

Limitations: robots.txt is declared intent, and after September 15 it is incomplete intent on Cloudflare sites. Cloudflare plan tier is still invisible from outside. Eight domains is too few to call the agents template a trend. The 50,000-site figure is a weighted extrapolation of a 5-point change, not a count.

References

- On-Page.ai Research, “A Third of the Top Million Websites Could Go Dark to AI on September 15” (August 17, 2026), the baseline study
- Cloudflare blog, “Say it once: Introducing Bot Preference Sync” (August 21, 2026, updated September 15, 2026)
- Cloudflare changelog, “New options to manage AI traffic” (July 1, 2026)
- Cloudflare docs, “Block AI bots” (Search, Training, and Agent classifications; September 15 defaults)
- Tranco top sites ranking (tranco-list.eu), list dated August 16, 2026
- OpenAI, Anthropic, and Perplexity crawler documentation (user agent tokens)

The panel stays frozen, and the next wave runs when Cloudflare ships its next change.

Questions about the data: team@on-page.ai.

Cite this study: Lancheres, E. (2026). The AI Access Census, Part 2: What Cloudflare's September 15 AI Default Actually Changed. A 10,600-Site Before-and-After. On-Page.ai Research. <https://api.on-page.ai/research/ai-access-census-part-2>