

94% of Press Releases Are Open to AI Crawlers

A pre-registered census of AI-crawler rules on 1,503 press-release websites



By **Émile Bolduc** · PPN World
Published October 8, 2026

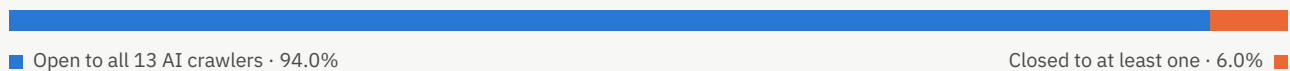
We read the robots.txt file behind every one of 500,129 press releases. 94.0% of them are open to every AI crawler we tested. Where a publisher does refuse, it refuses to train AI — not to be quoted by it.

Edition 1 · Files read October 7, 2026 · Published October 8, 2026 · Pre-registered

<https://www.ppnworld.com/research/press-releases-ai-crawlers> · Text, figures and data: CC BY 4.0

94.0%

of the 500,129 press releases are open to every AI crawler we tested



500,129

RELEASES CHECKED

1,503

WEBSITES

13

AI CRAWLERS TESTED



PRE-REGISTERED

ABSTRACT

AI assistants and answer engines now repeat press releases, which raises a practical question for the people who publish them: are those releases open to AI crawlers? Before reading a single file, we registered three hypotheses. We then read the robots.txt file of every website behind the 500,129 press releases in PPN World's archive from July 1, 2026 to October 6, 2026 — 1,503 websites run by companies, distributors, governments, cities, universities and leagues — and applied each file, as RFC 9309 specifies, to the address of every release, for 13 AI crawlers and for Google's and Bing's. 94.0% of the releases sit on websites that let every one of those AI crawlers in, and 99.0% are open to all three crawlers that feed AI search answers. Refusals are rare and aimed at training: 3.3% of websites turn away OpenAI's training crawler against 0.6% for Googlebot (H1), and 28 websites refuse that training crawler while admitting OpenAI's search crawler, and none does the reverse (H2). Public bodies and commercial publishers do not differ (3.8% against 4.2%; H3 not supported), and archived copies show the same picture twelve months earlier. By the end of 2023, nearly half of the most-read news websites were blocking OpenAI's crawlers (Fletcher, 2024); press releases have stayed open.

CONTENTS

01	Key findings	02	Why measure this	03	How we measured	04	Almost every release is open
05	Training refused, answering admitted	06	Who refuses	07	A year earlier	08	The other gate: the firewall
09	What it means for communicators	10	Method notes	11	Data, pre-registration and citation		

01 Key findings

01

94.0%

94% of press releases are open to every AI crawler

73 of the 1,443 websites that answered (5.1%) refuse at least one of the 13 AI crawlers we tested; weighted by the releases each carries, 6.0%. The five largest websites that answered — carrying 42.1% of those releases — refuse none.

02

99.0%

AI search is open even wider

Only 34 websites (2.4%) turn away even one of the crawlers that feed AI search answers — OpenAI's, Anthropic's or Perplexity's. 99.0% of releases are open to all three.

03

3.3% vs 0.6%

AI crawlers are refused more often than Google (H1)

48 websites (3.3%) refuse GPTBot; 9 (0.6%) refuse Googlebot. 39 refuse GPTBot and admit Googlebot; none does the opposite ($p < 0.001$). Expected, and now replicated at full scale.

04

28 vs 0

Publishers refuse training, not answering (H2)

28 websites refuse GPTBot, OpenAI's training crawler, while admitting OAI-SearchBot, which feeds ChatGPT search. Not one does the reverse ($p < 0.001$). Every website that turns away any AI crawler turns away a training crawler.

05

3.8% vs 4.2%

Governments and companies behave alike (H3, not supported)

3.8% of public-body websites and 4.2% of commercial ones refuse GPTBot (difference -0.3 pts, 95% CI -3.4 to $+2.0$ pts). Distributors refuse it more often than companies' own newsrooms (8.3% against 2.0%), but neither is common.

06

4.3% → 4.7%

A year changed nothing

Against archived copies of the same files from autumn 2025 (1,152 websites), blocking aimed at AI went from 4.3% to 4.7% of websites: 21 started, 17 stopped ($p = 0.63$).

07

75 websites

The firewall is the other gate

75 websites (5.0%) refused our automated client at the release page itself, even when it presented itself as a browser — about as many as refuse an AI crawler in robots.txt. Whether a real AI crawler gets through those firewalls cannot be seen from outside.

02 Why measure this

A press release is written to be repeated. In 2026, more and more of the repeating is done by AI: assistants that answer questions, search engines that summarize, and the models behind both. Whether a release can reach them depends first on a plain text file, robots.txt, through which a website tells each crawler where it may go. The file distinguishes crawlers by name, so a publisher can refuse a crawler that collects training data and admit one that fetches pages to answer questions.

News publishers answered the question early and loudly: by the end of 2023, 48% of the most widely used news websites across ten countries were blocking OpenAI's crawlers (Fletcher, 2024), and restrictions on AI crawlers spread quickly across the web's most-used sources (Longpre et al., 2024). Press releases have been argued about rather than measured — communicators hear both that AI answers now quote releases and that publishers are shutting AI out. This paper measures it, for a complete population of press-release websites, with the analysis registered before any file was read.

03 How we measured

The population is every release in PPN World's archive with a publication date from July 1, 2026 to October 6, 2026, from the sources the archive serves: 500,129 releases with a usable link, on 1,503 websites. A website here is an origin — scheme, host and port — because that is what a robots.txt file governs. Each website takes the publisher type of the sources that link to it: 781 belong to governments and public bodies (national, regional and local) and 346 to companies and commercial distributors.

On October 7, 2026 we requested each website's robots.txt as an identified research client, and once more as a browser if refused. Each file was applied exactly as RFC 9309 specifies (Koster et al., 2022) — the group that names the crawler, otherwise the * group; the longest matching rule wins — to the address of every release on that website, not just its home page, because a site can close one directory and leave the rest open. A website counts as refusing a crawler when more than half of its releases are closed to it. The 60 websites whose file could not be reached at all are left out of the rates and counted as refusals in a robustness check.

We tested 13 AI crawlers in three families fixed in advance from each operator's documentation (Anthropic, n.d.; Apple, n.d.; Common Crawl, n.d.; Google, n.d.; Meta, n.d.; OpenAI, n.d.; Perplexity, n.d.): crawlers that collect training data (GPTBot, ClaudeBot, CCBot, Meta-ExternalAgent, Bytespider, and the training controls Google-Extended and Applebot-Extended); crawlers that index pages for AI search answers (OAI-SearchBot, Claude-SearchBot, PerplexityBot); and fetchers that open a page when a user asks (ChatGPT-User, Claude-User, Perplexity-User). Googlebot and Bingbot are the reference. Three hypotheses were registered, with a Holm correction across them: AI crawlers are

refused more often than Googlebot (H1); OpenAI's training crawler more often than its search crawler (H2); and public bodies differ from commercial publishers (H3).

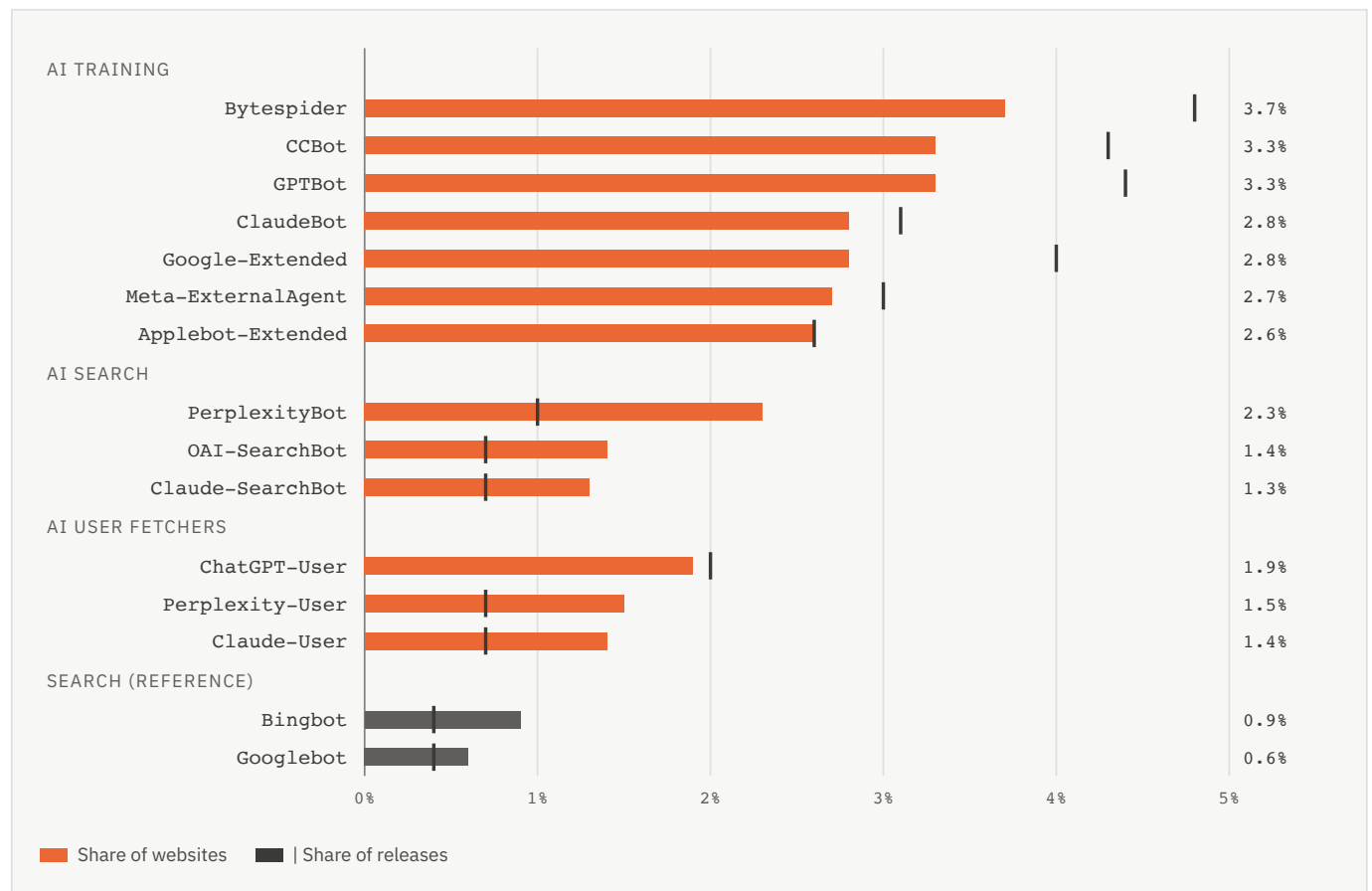
04 Almost every release is open

Almost none of the files close the door. 73 of the 1,443 websites that answered (5.1%, 95% CI 4.0% to 6.3%) refuse at least one AI crawler, and 34 refuse an AI search crawler. Weighted by the releases each website carries, 94.0% of releases are open to every AI crawler and 99.0% to all three AI search crawlers. The five largest websites that answered — carrying 42.1% of those releases — refuse none.

FIGURE 1

The crawlers press-release websites refuse

Share of websites whose robots.txt closes most of their release addresses to each crawler (bars), and share of releases closed to it (ticks). Websites that answered for their robots.txt.



The crawlers refused most often are those that collect training data: Bytespider, GPTBot and CCBot head the list. The least refused are the AI search crawlers and the fetchers that open a page for a user. Googlebot is refused by 9 websites, 8 of them because they close their releases to every crawler on the list. H1 is supported: 39 websites refuse GPTBot while admitting Googlebot, and none does the opposite (exact McNemar test, Holm-adjusted $p < 0.001$).

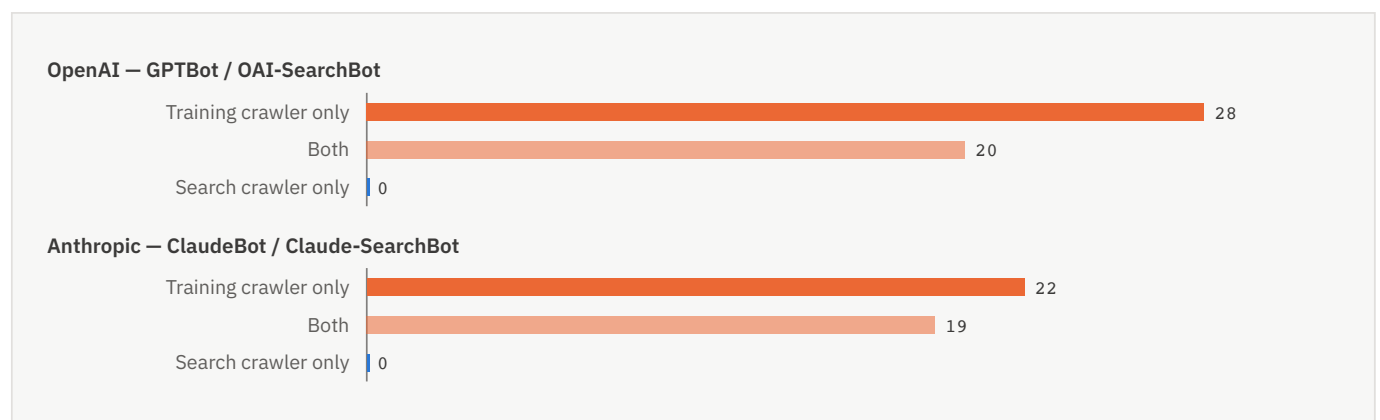
05 Training refused, answering admitted

H2 asked whether publishers tell training apart from answering. They do, in one direction only. Of the websites that answered, 20 refuse both of OpenAI's crawlers; 28 refuse GPTBot but admit OAI-SearchBot, the crawler OpenAI says surfaces websites in ChatGPT search (OpenAI, n.d.); and none does the reverse (exact McNemar test, Holm-adjusted $p < 0.001$; H2 supported). Anthropic's pair splits the same way: 22 websites refuse ClaudeBot alone, none Claude-SearchBot alone.

FIGURE 2

Training refused, answering admitted

Websites refusing each company's training crawler, its search crawler, or both. No website refuses the search crawler alone.



Part of the split comes from how the files are written. 35 of the 48 websites that refuse GPTBot name it in their file; the newer search crawlers are rarely named, so they fall under the * group, which usually admits them. Some publishers choose the split outright — Ireland's gov.ie names GPTBot, ClaudeBot, Google-Extended and Meta-ExternalAgent and leaves the search crawlers alone. Either way, a release closed to training is, in most cases, still open to being quoted.

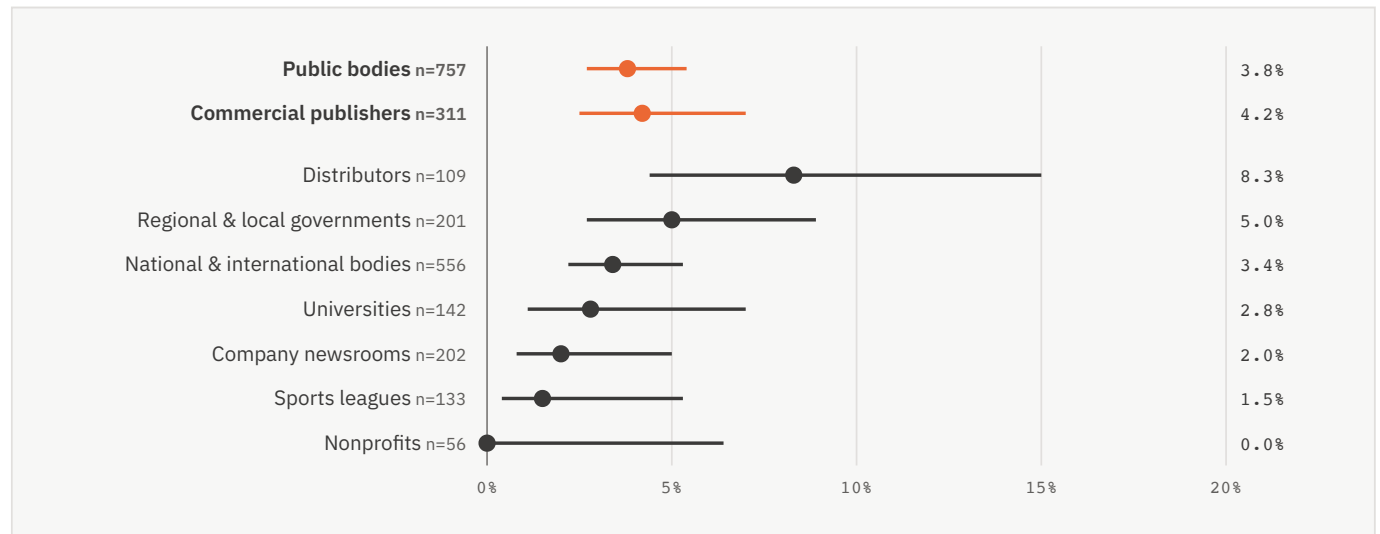
06 Who refuses

H3 found no difference between public bodies and commercial publishers. 3.8% of public-body websites (29 of 757) and 4.2% of commercial ones (13 of 311) refuse GPTBot; the difference, -0.3 pts (95% CI -3.4 to $+2.0$ pts), is indistinguishable from zero (Holm-adjusted p 0.86; H3 not supported). Within the commercial group, distribution websites refuse GPTBot more often than companies' own newsrooms: 9 of 109 (8.3%) against 4 of 202 (2.0%).

FIGURE 3

Who refuses GPTBot

Share of websites refusing GPTBot, by publisher type, with 95% intervals. The two pooled rows are H3's comparison. Websites that answered for their robots.txt.



Public bodies are public record, so the refusers among them can be named. Those carrying the most releases are Moscow's city government, the Generalitat Valenciana, Barcelona's city council, Ireland's gov.ie, Spain's competition authority (CNMC), the Texas governor's office, Austria's interior ministry and the World Trade Organization. Tuscany's regional news service goes further and closes its release pages to every crawler, Google included.

Among countries with at least 20 websites, refusing an AI crawler is more common in parts of Europe — the Netherlands (16.7%), Spain (16.1%), Germany (13.0%) — than in the United States (4.2%), Canada (4.4%) or Japan (1.9%). The groups are small and these figures are descriptive.

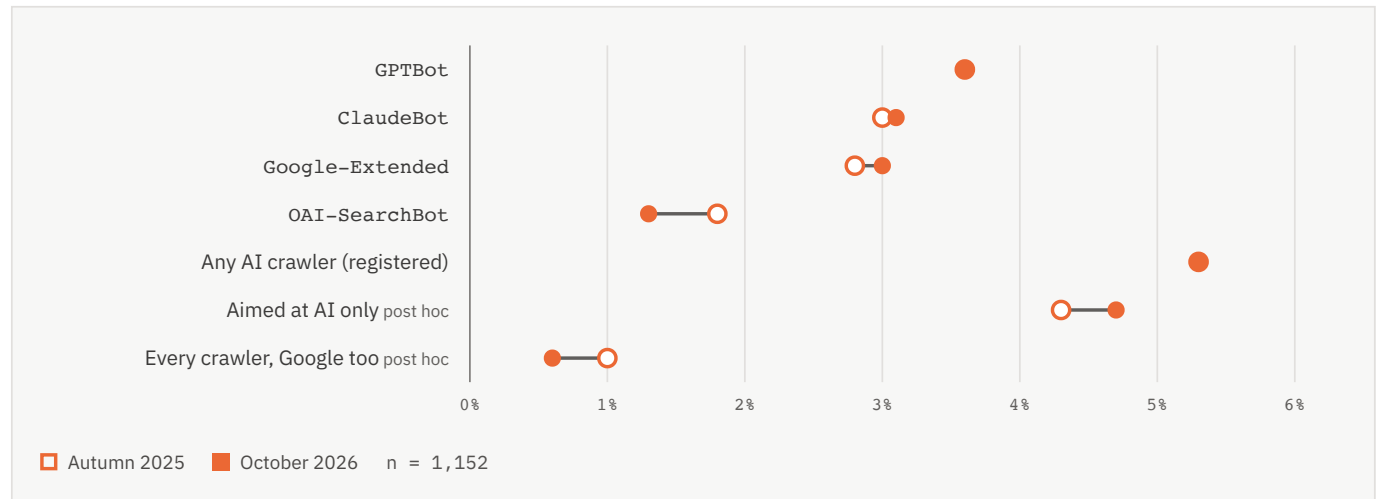
07 A year earlier

To see whether this is changing, we read the archived copy of each website's robots.txt closest to October 7, 2025 from the Internet Archive's Wayback Machine (Internet Archive, n.d.), within two months either side, and applied it to the same release addresses. 1,152 websites (80% of those that answered) had a usable copy; the median copy was 6 days from the target date.

FIGURE 4

Twelve months on, the same

Share of websites refusing each crawler in the archived copy of their robots.txt closest to 7 October 2025 (hollow) and today (filled), on the websites with a usable copy. Rows marked post hoc were split out after the data was seen.



The share refusing GPTBot was 3.6% then and 3.6% now. Under the registered measure — any AI crawler refused — 61 websites refused then and 61 now, with 23 starting and 23 stopping. That churn mixes two things, which we separated after seeing it: blanket blocks that close a site to every crawler, Google included (11 websites then, 7 now), and blocking aimed at AI alone, which went from 4.3% to 4.7% of websites (21 started, 17 stopped; $p = 0.63$). Publishers edit their files in both directions; the total has not moved.

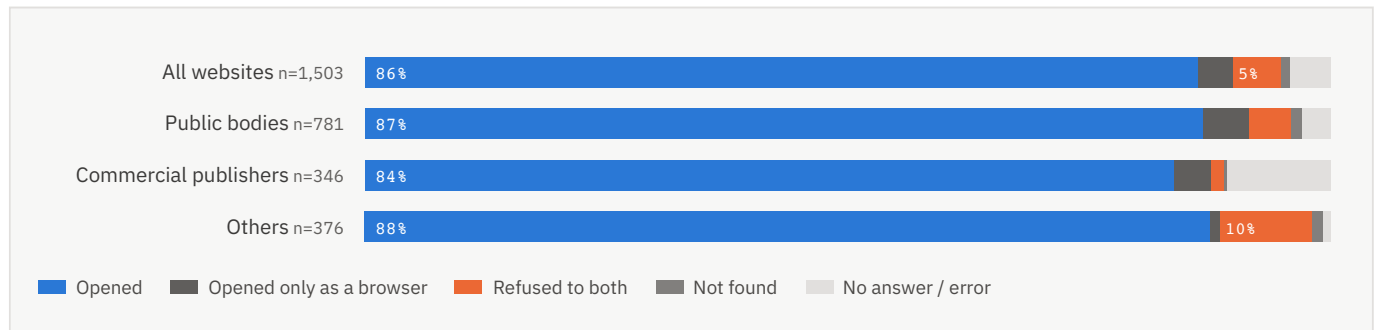
08 The other gate: the firewall

robots.txt is a request, and it is written down. A firewall in front of a website is enforced, and it is not written anywhere a crawler can read. When we opened the most recent release on each website, 75 websites (5.0%) refused our client outright, whether it identified itself honestly or as a browser, and 54 more refused the honest identity but opened for the browser one. 60 websites gave no usable answer for their robots.txt (timeouts, server errors, broken certificates); a second HTTP client, added after the fetch, found that 30 of them — all company websites, most of them investor-relations pages — leave an identified research client without an answer and refuse a browser identity with a 403.

FIGURE 5

What an automated client met at the release page

Outcome of opening the most recent release on each website: opened for an honestly identified client; opened only for a browser identity; refused to both; not found; no answer or server error.



A real AI crawler is not our client. OpenAI and Google publish the network addresses their crawlers use, and a firewall can admit them while refusing everyone else. What can be said is that about as many press-release websites put an automated-client firewall in front of their releases as refuse an AI crawler in robots.txt — and that the second is a line of text the publisher wrote, while the first is often a default the publisher never saw.

Other signals are rarer still. 29 of the 1,351 release pages we opened tell search engines not to index them; 1 carries the unofficial noai tag; 10 robots.txt files carry Content Signals (Cloudflare, 2025), 5 of them refusing AI training. Meanwhile 81 websites (5.4%) publish an llms.txt file (Howard, 2024) — a guide written for AI systems — more than refuse an AI crawler.

09 What it means for communicators

Do press-release websites block AI crawlers?

Rarely. Of the 1,443 websites behind 500,129 releases that answered, 5.1% refuse at least one AI crawler in robots.txt. 94.0% of releases are open to every AI crawler we tested, and 99.0% to the crawlers that feed AI search answers.

Does blocking GPTBot keep a release out of ChatGPT search?

No. OpenAI documents separate crawlers (OpenAI, n.d.): GPTBot collects training data, OAI-SearchBot surfaces websites in ChatGPT search, and each obeys its own robots.txt rules. 28 press-release websites refuse the first and admit the second; none does the reverse. Google-Extended, likewise, governs Gemini training and grounding, not Google Search (Google, n.d.).

If robots.txt allows AI crawlers, can they read the release?

Not necessarily. A firewall can refuse automated clients whatever robots.txt says: 75 of the 1,503 websites refused our client outright, and 30 company websites, most of them investor-relations pages, gave an identified client no answer at all. Test the newsroom with a non-browser client, not only the robots.txt file.

Which copy of a release should a communicator check?

Every website the release lives on — the newsroom, the distributor, any syndication — since each has its own file and its own firewall. The five largest websites in our population that answered (42.1% of those releases) refuse no AI crawler; the company's own newsroom is the copy it controls.

10 Method notes

Statistics. H1 and H2 use exact McNemar tests (McNemar, 1947) on the paired outcome per website; H3 uses Fisher's exact test, with a Newcombe hybrid-score interval (Newcombe, 1998) for the difference in proportions. Holm's correction (Holm, 1979) covers the three tests. Intervals on shares of websites are Wilson 95% intervals (Wilson, 1927). The population is a census of the archive's websites, so shares of releases are reported without intervals.

Robustness, all published in the data. Counting unreachable files as refusals raises the share of websites refusing any AI crawler to 8.8% (GPTBot 7.2%). Leaving out the five largest websites raises the release-weighted share to 10.3%. Using the identified client's answer alone gives 3.1% for GPTBot. Calling a website a refuser if any one of its releases is closed gives 3.7% for GPTBot and 1.2% for Googlebot. Every file was read a second time at least 6 hours after the first: of the 1,438 websites that answered both times, none changed its GPTBot verdict (5 did not answer the second time). No conclusion changes.

Deviations from the registration

1. Tracking parameters. The registration applied each file to the full release address, query included. Some feeds append campaign-tracking tags (utm_ and similar) to every link, and a site-wide rule against query strings then reads as a refusal of releases that are in fact open — one distributor's whole output flipped this way. Those tags (5,660 release addresses) are removed before matching. With them left in, as registered, GPTBot is refused by 3.5% of websites and any AI crawler by 5.2%; 2 websites change, no conclusion does.
2. Archived copies. The registration asked for the capture with status 200 closest to October 7, 2025. The Wayback Machine's search index answered in about 40 seconds per query, so each capture was located with its closest-capture redirect instead, and kept only if it had been archived with status 200. A website whose nearest capture was an error counts as uncovered.
3. Page-level signals are read only from pages that opened. A refusal page carries its own noindex tag, which says nothing about the release behind it.
4. Added after the data was seen, and labelled as such: the split of the year-on-year change into blanket and AI-specific blocking; the same comparison without websites whose current file is refused to every automated client (48 then, 54 now); and the second-client check of unreachable websites.

Limitations

- robots.txt is a request. This paper reports what each file asks, not whether each crawler complies; OpenAI and Perplexity state that their user-triggered fetchers may not follow robots.txt (OpenAI, n.d.; Perplexity, n.d.).
- Our client is not an AI crawler. Firewalls that admit verified crawlers by address are invisible from outside, so the firewall figures measure exposure to automated clients in general, not to AI.
- The files are those served on one day. A file can change the next morning.
- The population is PPN World's archive — about 1,500 websites, weighted toward English-language publishers and public bodies — not every press-release publisher in the world.
- A release can live on several websites. We test the copy the archive links to; the issuer's own newsroom or a syndicated copy may answer differently.
- A few public bodies' releases link to third-party platforms (Google News, X, Instagram, Mailchimp), whose files are the platform's, not the body's.
- Archived copies cover four in five websites and were captured by the Internet Archive's own crawler, which may not have been served what others were.

References

- Anthropic. (n.d.). Does Anthropic crawl data from the web, and how can site owners block the crawler? Claude Help Center. Retrieved October 8, 2026, from <https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler>
- Apple. (n.d.). About Applebot. Apple Support. Retrieved October 8, 2026, from <https://support.apple.com/en-us/119829>
- Cloudflare. (2025, September 24). Content Signals Policy. <https://contentsignals.org/>
- Common Crawl. (n.d.). CCBot. Retrieved October 8, 2026, from <https://commoncrawl.org/ccbot>
- Fletcher, R. (2024, February 22). How many news websites block AI crawlers? Reuters Institute for the Study of Journalism. <https://doi.org/10.60625/risj-xm9g-ws87>
- Google. (n.d.). Google's common crawlers. Google for Developers. Retrieved October 8, 2026, from <https://developers.google.com/crawling/docs/crawlers-fetchers/google-common-crawlers>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70. <https://www.jstor.org/stable/4615733>
- Howard, J. (2024, September 3). The /llms.txt file. llms-txt. <https://llmstxt.org/>
- Internet Archive. (n.d.). Wayback Machine. Retrieved October 8, 2026, from <https://web.archive.org/>
- Koster, M., Illyes, G., Zeller, H., & Sassman, L. (2022, September). Robots Exclusion Protocol (RFC 9309). Internet Engineering Task Force. <https://doi.org/10.17487/RFC9309>
- Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A. M., Dinh, A., Chitongo, C., Yin, D., . . . Pentland, S. (2024). Consent in crisis: The rapid decline of the AI data commons. *Advances in Neural Information Processing Systems*, 37, 108042–108087. <https://doi.org/10.52202/079017-3431>

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.

<https://doi.org/10.1007/BF02295996>

Meta. (n.d.). Meta web crawlers. Meta for Developers. Retrieved October 8, 2026, from

<https://developers.facebook.com/documentation/sharing/webmasters/web-crawlers>

Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: Comparison of eleven methods. *Statistics in Medicine*, 17(8), 873–890.

[https://doi.org/10.1002/\(SICI\)1097-0258\(19980430\)17:8%3C873::AID-SIM779%3E3.0.CO;2-I](https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8%3C873::AID-SIM779%3E3.0.CO;2-I)

OpenAI. (n.d.). Overview of OpenAI crawlers. OpenAI Developers. Retrieved October 8, 2026, from <https://developers.openai.com/api/docs/bots>

Perplexity. (n.d.). Perplexity crawlers. Perplexity Docs. Retrieved October 8, 2026, from

<https://docs.perplexity.ai/docs/resources/perplexity-crawlers>

Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212.

<https://doi.org/10.1080/01621459.1927.10502953>

11 Data, pre-registration and citation

Every website-level outcome behind the figures is published as CSV under CC BY 4.0, with the date each file was read. Public bodies are listed by address; companies and distributors are listed by type and rank, because PPN World does not name commercial sources it has no agreement with. The pre-registration is published unedited.

↓ Download the data (CSV)

One row per website: type, country, releases, robots.txt status, the crawlers refused now and twelve months earlier.

Read the pre-registration

Registered October 7, 2026, before any robots.txt file was read, and served exactly as committed — so it calls this paper by its working number, No. 03, and mentions an unpublished pilot study.

Cite this paper

Free to quote, chart and republish with attribution. Suggested citations:

APA

Bolduc, É. (2026). 94% of press releases are open to AI crawlers: A pre-registered census of AI-crawler rules on 1,503 press-release websites (PPN World Research Paper No. 1, Edition 1).

PPN World. <https://www.ppnworld.com/research/press-releases-ai-crawlers>

MLA

Bolduc, Émile. 94% of Press Releases Are Open to AI Crawlers: A Pre-Registered Census of AI-Crawler Rules on 1,503 Press-Release Websites. PPN World, 8 Oct. 2026,

www.ppnworld.com/research/press-releases-ai-crawlers.

CHICAGO

Bolduc, Émile. 2026. 94% of Press Releases Are Open to AI Crawlers: A Pre-Registered Census of AI-Crawler Rules on 1,503 Press-Release Websites. PPN World Research Paper 1. PPN World. <https://www.ppnworld.com/research/press-releases-ai-crawlers>.

Licence

Text, figures and data are published under [CC BY 4.0](#)

Corrections

None yet. A correction will be published here with its date; a cited figure is never changed silently.

ABOUT THE AUTHOR



Émile Bolduc

PPN World Research Center

Émile Bolduc wrote this paper for PPN World, which monitors press releases from companies, distributors and public bodies worldwide. Questions about the method or the data, or a cut of the data we have not published: write to the Research Center.

[LinkedIn profile](#)

[Contact the Research Center](#)