

Which websites block AI crawlers: a census of robots.txt rules on 6,094 lower-ranked sites, October 2026

Preprint (web measurement)

Yuma Heymans

AIData.Broker, <https://aidata.broker> | hello@aidata.broker | 18 October 2026

Abstract

Website operators who object to the use of their pages for training AI models usually say so in robots.txt, the plain-text file that tells automated clients what they may fetch. This paper reports a census of those files. A reader takes the Tranco ranking of popular sites, keeps registrable domains with eleven address endings within the top 300,000 (181,046 sites), and reads one slice of 4,000 sites each morning in rank order, so that a full pass takes 46 mornings. The current pass began on 15 October 2026 at slice 42, set by date arithmetic, and by 17 October had read slices 42 to 44: Tranco ranks 270,775 to 291,236, well below the most visited sites. Of 12,000 requests, 6,094 returned a usable robots.txt. Of these, 696 (11.4%) disallow at least one of 15 named AI user-agent tokens from the whole site, and 514 (8.4%) disallow two or more, one of them a token used for training. GPTBot is refused most often (457 sites, 7.5%), followed by CCBot (7.2%) and Bytespider (7.0%). Sites refuse training tokens more often than the search and user-request tokens of the same operator: GPTBot is refused 3.7 times as often as OAI-SearchBot. Refusing sites name 6.4 tokens on average, including names that the operator's current documentation does not list, which suggests maintained or copied block lists. We set out the sampling frame, fetch protocol and parser, and the limits of a stated-refusal measure taken on three slices of the ranking.

Keywords: robots.txt; Robots Exclusion Protocol; web crawling; AI crawlers; web measurement; Tranco; training data; opt-out signals; RFC 9309

1. Introduction

Corpora built from crawled web pages are among the most widely used sources of training text for language models [13]. The signal a website can send to the collectors, and the one their operators' documentation asks sites to use, is the robots.txt file at the root of its address [6, 7, 8, 9]. The Robots Exclusion Protocol, standardised in 2022 as RFC 9309, defines the file as a set of groups: each group names one or more crawlers by a "product token" in user-agent lines, followed by allow and disallow rules for paths [5]. The file is advisory. RFC 9309 states the rules crawlers are asked to follow; it does not give the operator a way to enforce them.

AI developers have adapted the file to their own purposes by publishing separate tokens for separate uses. OpenAI documents GPTBot for collecting content that may be used for training and OAI-SearchBot for search, and states that "each setting is independent of the others", so that a site may admit the search crawler while refusing the training crawler [6]. Google and Apple publish tokens that do not crawl at all: Google-Extended is "used in a control capacity" to decide whether content Google already crawls may be used to train Gemini models [8], and Applebot-Extended, which "does not crawl webpages", lets

publishers opt out of training Apple's foundation models while Applebot still crawls [9]. A refusal in robots.txt is therefore increasingly a statement about use, not about access.

The signal has acquired legal weight in some places. The EU text and data mining exception applies only where use "has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online" [15], and the AI Act requires providers of general-purpose models to have a policy "to identify and comply with" such reservations [16]. In the United States, a federal court held in December 2025 that robots.txt files have "no effect on the bot's ability to access" a site's works unless the bot's creator chooses to heed them, so disregarding them was not circumvention of an access control [17]. Developers are also now asked to document where their training data came from [21]. None of this is legal advice; it explains why the count of refusals matters to those who build datasets and to those who license content to them.

Two earlier measurements frame this one. An audit of the websites behind widely used training corpora found that, between 2023 and 2024, robots.txt restrictions fully restricted about 5% or more of all tokens in the C4 corpus and 28% or more of its most actively maintained, critical sources [13]. Cloudflare reported that in 2025 Anthropic's crawlers fetched up to 500,000 pages for each visitor they referred back, while OpenAI's ratio peaked at 3,700 to one [14]. Both concern the most visible part of the web. This paper describes a census designed to read the whole of a large ranking, slice by slice, and reports its first three slices, which lie in the lower part of that ranking. The contribution is a documented, repeatable method, per-token counts for 15 AI user-agent tokens, and a statement of what the counts can and cannot show.

2. Data and method

2.1 Sampling frame

The frame is the Tranco list, a research ranking of popular sites that aggregates several existing top-site lists and is designed to resist manipulation [3, 4]. The census uses the copy of the list downloaded on 15 October 2026 (it keeps a copy for up to 30 days). From the list it keeps an entry when three conditions hold: its rank is 300,000 or better; the listed name is itself a registrable domain, so subdomains and infrastructure hosts are dropped; and the address ends in one of eleven endings: .com, .net, .org, .io, .co, .ai, .app, .dev, .info, .us or .uk [1, 2]. The endings correspond to the regions in which the compiler works; a generic ending says nothing about where a site's operator is.

Applying those rules to the cached copy of the list gives 181,046 sites in reach. The reader divides them, in rank order, into slices of 4,000 and reads one slice each morning; $\text{ceil}(181,046 / 4,000) = 46$ slices make one pass, which matches the cycle the reader recorded with each sweep [2].

The order of reading matters for what the first sweeps represent. When the reader first ran, it had no record of a previous slice, so it chose the slice by date arithmetic: the number of days since 1 January 1970, modulo 46. On 15 October 2026 that number is 20,741, and 20,741 modulo 46 is 41, which is the 42nd slice. Each later morning continued from the slice last read. The current pass had therefore read slices 42, 43 and 44 by 17 October 2026, covering the 164,001st to the 176,000th sites in reach. Re-applying the reader's rules to the cached list shows that these are Tranco ranks 270,775 to 291,236. Table 1 gives the composition of the 12,000 requested sites by address ending.

Table 1. The 12,000 sites requested in the current pass, by address ending

Ending	Sites requested	Share
.com	8,751	72.9%
.net	1,128	9.4%
.org	1,012	8.4%
.uk	364	3.0%
.io	288	2.4%
.info	126	1.1%
.ai	96	0.8%
.co	75	0.6%
.app	73	0.6%
.us	47	0.4%
.dev	40	0.3%
All	12,000	100.0%

Source: computed by re-applying the census's selection rules [1, 2] to the Tranco list of 15 October 2026 [4], slices 42 to 44.

2.2 Fetch protocol

For each site the reader sends one HTTPS GET request for /robots.txt at the registrable domain, with a timeout of 8 seconds, no retry, and at most 40 requests in flight. Requests carry a User-Agent string that names the compiler and links to its site. A response counts as "answered" only if all of the following hold: the request succeeded with a success status; the body is not empty; it does not begin with "<" (an HTML page served in place of the file); it is shorter than 500,000 characters; and, after any redirects, the final address has the same registrable domain as the site requested. The last rule reflects a choice: a robots.txt served from another domain (for example a parent company's) is treated as speaking for that domain, not for the site requested. RFC 9309 instead asks crawlers to follow up to five redirects, even across hosts, and apply the result to the original site [5]. Sites that time out, refuse the request, answer with an error status or serve HTML are not counted as answered and are not in the denominator of any share below.

2.3 Parser

The parser strips comments, reads each line as a field and a value, and builds groups in the RFC 9309 sense: one or more consecutive user-agent lines followed by rules. A token is counted as "shut out of the whole site" when it appears, by case-insensitive exact match of the user-agent value, in a group that contains the rule "Disallow: /". Three consequences follow, and Section 4 returns to each. First, a group addressed to the wildcard "*" does not count, even with "Disallow: /", because the census measures refusals that single out AI crawlers. Under RFC 9309 a crawler named by no group obeys the wildcard group [5], so a site that refuses all crawlers in this way refuses AI crawlers too, but is not counted here. Second, a disallow rule for a narrower path (for example "Disallow: /articles/") does not count. Third, an allow rule in the same group is not evaluated: a group that disallows "/" but allows one path is counted as a whole-site refusal, although under the RFC's most-specific-match rule the allowed path stays open [5].

2.4 Tokens tested

The census tests the 15 tokens in Table 2. For the counts it uses, it also classes nine of them as tokens that gather training data; the other six are search, user-request, undocumented or general-purpose names. The operators' own documentation, read on 18 October 2026, describes several of them as follows: OpenAI separates GPTBot (training) from OAI-SearchBot (search) and ChatGPT-User (fetches made when a user asks), and says of the last that "because these actions are initiated by a user, robots.txt rules may not apply" [6]. Anthropic currently documents ClaudeBot (training), Claude-User and Claude-SearchBot [7]; the census tests ClaudeBot and two other Anthropic names, anthropic-ai and Claude-Web, that the current documentation does not list, and does not test Claude-User or Claude-SearchBot. PerplexityBot "is not used to crawl content for AI foundation models" [10]. Amazon states that Amazonbot "may be used to train Amazon AI models" [11], although the census does not class it as a training token. Common Crawl describes CCBot as the crawler of its open repository of web crawl data [12].

Table 2. The 15 tokens tested, their operators and the census's classification

Token	Operator	Use, as documented or described	Census class
GPTBot	OpenAI	Training [6]	Training
OAI-SearchBot	OpenAI	Search [6]	Other
ChatGPT-User	OpenAI	Fetches on a user's request [6]	Other
ClaudeBot	Anthropic	Training [7]	Training
anthropic-ai	Anthropic	Not in current documentation [7]	Training
Claude-Web	Anthropic	Not in current documentation [7]	Other
CCBot	Common Crawl	Open web archive [12]	Training
Google-Extended	Google	Control token for Gemini training [8]	Training
Applebot-Extended	Apple	Control token for model training [9]	Training
meta-externalagent	Meta	Meta's crawler [1]	Training
Bytespider	ByteDance	ByteDance's crawler [1]	Training
cohere-ai	Cohere	Not described in the sources read	Training
PerplexityBot	Perplexity	Search results; not for foundation models [10]	Other
Amazonbot	Amazon	Products and services; may train models [11]	Other
Diffbot	Diffbot	Not described in the sources read	Other

Source: the census code [2] for the tokens and classes; operators' documentation as cited.

2.5 Measures

Each sweep records five counts for its slice: robots.txt files requested; files answered (Section 2.2); sites that shut at least one of the 15 tokens out of the whole site; sites that shut out two or more tokens, at least one of them in the training class; and, for each token, the number of sites that shut it out [1, 2]. Only counts are kept in the public series. The census does not publish the list of sites, and this paper names none. Unless stated otherwise, shares are of the sites that answered. Because the slices of one pass do not overlap, the counts of the three sweeps add up to one reading of 12,000 distinct sites; the per-token counts are summed in the same way. The full series is reproduced in Appendix A, and each sweep has its own public record [2].

3. Findings

3.1 Response

Of the 12,000 requests, 6,094 (50.8%) returned a usable robots.txt: 2,107 of 4,000 on 15 October, 2,089 on 16 October and 1,898 on 17 October (52.7%, 52.2% and 47.4%). The census does not record why the other requests failed. The group includes sites that publish no robots.txt (which, under RFC 9309, means no rules apply [5]), sites that refused or timed out on a single request from an unfamiliar client, sites that served an HTML page instead of the file, and sites whose file came from another domain. The distinction matters for interpretation, and Section 4 discusses it.

3.2 How many sites refuse AI crawlers

Of the 6,094 sites that answered, 696 (11.4%) shut at least one of the 15 tokens out of the whole site. Expressed against all 12,000 sites requested, the same count is 5.8%. Of the 696, 514 (8.4% of answering sites, 73.9% of refusing sites) shut out two or more tokens, at least one of them in the training class. The remaining 182 sites (3.0%) shut out a single token, or two or more tokens none of which is in the training class.

The three slices give similar shares. The share refusing at least one token was 11.1%, 11.7% and 11.5% on the three days, and the share refusing two or more was 7.9%, 9.1% and 8.3%. The slices are adjacent in rank, so this agreement says that the measure is stable across neighbouring slices, not that it holds across the ranking.

3.3 Which crawlers are refused

Table 3 reports each token. GPTBot is the most refused, by 457 sites (7.5%), followed by CCBot (437, 7.2%) and Bytespider (427, 7.0%). ClaudeBot is refused by 367 (6.0%), Amazonbot by 333 (5.5%), meta-externalagent by 327 (5.4%) and Google-Extended by 302 (5.0%). At the other end, OAI-SearchBot is refused by 125 sites (2.1%) and PerplexityBot by 186 (3.1%).

Table 3. Sites that shut each token out of the whole site, current pass (three sweeps, 6,094 sites answered)

Token	Sites	Share of answering sites	Share of refusing sites (696)	Range across the three sweeps
GPTBot	457	7.5%	65.7%	6.7% to 8.0%
CCBot	437	7.2%	62.8%	6.8% to 7.5%
Bytespider	427	7.0%	61.4%	6.6% to 7.5%
ClaudeBot	367	6.0%	52.7%	5.6% to 6.7%
Amazonbot	333	5.5%	47.8%	4.7% to 6.0%
meta-externalagent	327	5.4%	47.0%	4.9% to 5.6%
Google-Extended	302	5.0%	43.4%	4.5% to 5.4%
anthropic-ai	292	4.8%	42.0%	4.5% to 5.4%
Applebot-Extended	270	4.4%	38.8%	3.8% to 4.8%
Diffbot	252	4.1%	36.2%	3.9% to 4.4%
cohere-ai	242	4.0%	34.8%	3.7% to 4.3%
Claude-Web	234	3.8%	33.6%	3.4% to 4.3%
ChatGPT-User	221	3.6%	31.8%	3.1% to 4.1%
PerplexityBot	186	3.1%	26.7%	2.8% to 3.3%
OAI-SearchBot	125	2.1%	18.0%	1.7% to 2.3%

Source: computed from the census sweeps of 15, 16 and 17 October 2026 [1, 2]; the per-sweep counts are in Appendix A. The range column gives the lowest and highest share of answering sites among the three sweeps.

The ordering is not fixed at the top. CCBot was the most refused token in the first sweep and GPTBot in the second and third, and the first three tokens are within 30 sites of one another over the pass. At the bottom, by contrast, the order was the same in every sweep: OAI-SearchBot last, PerplexityBot second to last and ChatGPT-User third to last.

3.4 Training tokens against search and user tokens

Where one operator publishes several tokens, sites refuse the training token more often. For OpenAI, GPTBot is refused by 457 sites, ChatGPT-User by 221 and OAI-SearchBot by 125: the training token is refused 2.1 times as often as the user-request token and 3.7 times as often as the search token. OpenAI's documentation describes exactly this configuration as an option, admitting the search crawler while refusing the training crawler [6]. The answer engine Perplexity, whose crawler is described as not used for foundation models [10], is refused by 186 sites, against 457 for GPTBot.

The two tokens that do not crawl are refused at rates close to the crawlers themselves: Google-Extended by 302 sites and Applebot-Extended by 270. Neither token appears in any request a site receives [8, 9], so a site that names them is responding to the operators' published controls rather than to crawler traffic it observed. Amazonbot, which the census does not class as a training token although Amazon says it may be used to train models [11], is refused by more sites (333) than Google-Extended.

3.5 Refusals come in lists

Sites that refuse one AI token usually refuse many. The 696 refusing sites account for 4,472 token refusals in all, an average of 6.4 tokens per refusing site (6.2, 6.8 and 6.3 in the three sweeps). No single token is named by more than 65.7% of refusing sites, so the lists differ from site to site, but they are long.

The Anthropic tokens show where the lists come from. ClaudeBot, the token Anthropic currently documents for training [7], is refused by 367 sites. Two other Anthropic names that its current documentation does not list are refused by 292 sites (anthropic-ai) and 234 (Claude-Web). Conversely, the census does not test Claude-User and Claude-SearchBot, which Anthropic does document. A site that names undocumented tokens and omits documented ones is most plausibly maintaining a list compiled at some earlier date, or copied from a published one. The counts cannot show which, because the census keeps no file text, but they caution against reading a refusal list as a current, deliberate choice about each named operator.

3.6 Placing the counts

These shares are not directly comparable with earlier measurements. The audit of the sources behind C4 and other corpora found the highest rates of restriction among the most actively maintained, critical sources [13]; the census reads, so far, a band of sites ranked 270,775 to 291,236 by popularity. What the census adds is a frame that will, when a pass is complete, cover every one of 181,046 sites in reach once, with the same protocol, so that rates can be compared across bands of the ranking.

What a refusal means also varies. The court in *Ziff Davis v. OpenAI* described robots.txt directives as requests that a bot's creator may or may not choose to heed [17]. Platforms have nonetheless pleaded them as part of their case: Reddit's complaint against SerpApi and other defendants, as the court summarised it when ruling on motions to dismiss, cited its robots.txt prohibiting "unapproved web crawlers from accessing any part of Reddit's website" alongside terms that require a separate agreement for AI training [18]. Publishers have alleged in recent claims that named crawlers took their content without permission: six Canadian news organisations in a claim filed in October 2026 that names ClaudeBot [19], and two reference publishers in a complaint that names PerplexityBot [20]. These are allegations, and this paper offers no view on them (not legal advice). They show why the presence of a named refusal, and its date, can matter after the fact, and why a census that records when a refusal was observed is of use beyond counting.

4. Limitations

Coverage. Three of 46 slices have been read, all in the lower part of the ranking (Tranco ranks 270,775 to 291,236). Shares for the most visited sites, or for the ranking as a whole, may differ in either direction, and nothing in this paper estimates them. The frame also excludes sites outside the top 300,000, subdomains, and every address ending not among the eleven (for example .de, .ca or .au).

Denominator and non-response. Shares are of the 6,094 sites that returned a usable file. The 5,906 that did not are a mixture the census cannot separate: sites with no robots.txt, sites behind bot protection that refused an unfamiliar client, transient failures on a single attempt with no retry, HTML served in place of the file, and files served from another domain. If refusing sites are more likely than others to block unfamiliar clients, the shares understate refusal among sites that publish a file; if sites with no file were counted as answering, the shares would fall towards the 5.8% computed against all requests.

Parser. The rule counts named whole-site refusals only. It misses refusals by wildcard, partial-path refusals, and refusals by tokens outside the 15 tested (for example Claude-User, Claude-SearchBot and other operators' newer names). It overcounts groups that disallow "/" but allow some paths. It requires an exact token match, so a user-agent line that adds a version or other text to a token is not counted. These choices make the measure conservative for "shuts out every AI crawler" and approximate for "shuts out this token".

Classification. The training class is the census's own rule. Moving Amazonbot into it, as Amazon's documentation would support [11], could change the count of sites refusing two or more tokens including a training token, but not the count refusing any token or the per-token counts.

Stated refusal, not behaviour. A robots.txt rule records what an operator asked. It does not show whether any crawler complied, and operators may block crawlers by other means (network filtering, paywalls, terms of service) that the census does not see.

Time. Each slice was read once, on one day. The series cannot yet show change, and the three daily shares differ by slice as well as by date.

Replication. The public series keeps counts, not site lists, so an independent check must re-run the selection and re-fetch the files. The selection can be reproduced exactly from the method above and a copy of the Tranco list of 15 October 2026; the fetched files cannot, because sites change.

Position of the compiler. The census is run by the author's organisation, which acts for organisations that license data to AI developers and uses the census to identify sites that guard their content (Section 7). That interest could bear on how refusal is defined and reported. The method is stated here in full so that it can be checked.

5. Conclusion

In the first three slices of a census of 181,046 popular sites, slices that lie between ranks 270,775 and 291,236, about one site in nine that publishes a readable robots.txt shuts at least one named AI crawler out entirely, and about one in twelve shuts out two or more including a training crawler. The refusals discriminate by use: operators' training tokens are refused far more often than their search and user-request tokens. They also come as long lists that include names the operators' current documentation does not list, which suggests that many sites maintain block lists rather than decide operator by operator.

The census will complete its first pass after 46 sweeps, at which point rates can be compared across the whole frame. Three changes would make the measure more informative: recording the reason for each failed request, so that the denominator can be described; testing the tokens operators currently document as they appear; and keeping, for each refusing site, the date on which a refusal was first observed, which is the fact a later dispute is most likely to need.

6. Data availability

Every sweep is published as its own record at <https://aidata.broker/records/census>, and the current pass is summarised at <https://aidata.broker/ai-crawlers> [1, 2]. The complete series used here is reproduced in

Appendix A. The selection rules can be applied to the Tranco list, which is available at <https://tranco-list.eu/> [4]. Site-level results are not published.

7. Conflict of interest

The author is chief executive of AIData.Broker, which compiles and publishes the census and acts as an agent for organisations that license or sell data to AI developers. Beyond the published counts, the compiler uses the census to identify sites that refuse two or more AI crawlers, including a training crawler, as organisations that may hold content of interest to licensees. No site is named in this paper or on the census page.

8. Funding

The census and this analysis were produced by AIData.Broker. No other funding was received.

9. Use of AI tools

Generative AI tools were used to assist with the scripted analysis, with checking the sources and with drafting this manuscript. The author reviewed the text, the tables and every cited source and takes responsibility for the content.

Appendix A. The census series, current pass

Table A1. Counts per sweep

Field	15 Oct 2026 (slice 42)	16 Oct 2026 (slice 43)	17 Oct 2026 (slice 44)	Pass
Requested	4,000	4,000	4,000	12,000
Answered	2,107	2,089	1,898	6,094
Refuse at least one token	233	244	219	696
Refuse two or more, one in the training class	166	190	158	514
GPTBot	141	168	148	457
CCBot	144	156	137	437
Bytespider	140	157	130	427
ClaudeBot	122	139	106	367
Amazonbot	100	119	114	333
meta-externalagent	104	117	106	327
Google-Extended	94	113	95	302
anthropic-ai	95	112	85	292
Applebot-Extended	81	101	88	270
Diffbot	82	87	83	252
cohere-ai	77	89	76	242
Claude-Web	79	90	65	234
ChatGPT-User	76	86	59	221
PerplexityBot	63	69	54	186
OAI-SearchBot	44	49	32	125

Source: the census sweep records [2]; all three sweeps read the Tranco list of 15 October 2026, in a pass of 46 slices.

References

All URLs were accessed on 18 October 2026.

[1] AIData.Broker, "Which sites block AI crawlers: a daily census of robots.txt," web page, measured 17 October 2026. <https://aidata.broker/ai-crawlers>

[2] AIData.Broker, "AI crawler census sweeps," records of 15, 16 and 17 October 2026 (one per sweep), with the census method. <https://aidata.broker/records/census> ; for example <https://aidata.broker/records/census/2026-10-15-part-42>

[3] V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński and W. Joosen, "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation," Network and Distributed System Security Symposium (NDSS), 2019. <https://www.ndss-symposium.org/ndss-paper/tranco-a-research-oriented-top-sites-ranking-hardened-against-manipulation/>

[4] Tranco, "A Research-Oriented Top Sites Ranking Hardened Against Manipulation" (list and download), list of 15 October 2026. <https://tranco-list.eu/>

- [5] M. Koster, G. Illyes, H. Zeller and L. Sassman, "Robots Exclusion Protocol," RFC 9309, Internet Engineering Task Force, September 2022. <https://www.rfc-editor.org/rfc/rfc9309>
- [6] OpenAI, "Overview of OpenAI Crawlers" (GPTBot, OAI-SearchBot, ChatGPT-User). <https://developers.openai.com/api/docs/bots>
- [7] Anthropic, "Does Anthropic crawl data from the web, and how can site owners block the crawler?" (ClaudeBot, Claude-User, Claude-SearchBot). <https://support.claude.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler>
- [8] Google, "Google's common crawlers" (Google-Extended). <https://developers.google.com/search/docs/crawling-indexing/google-common-crawlers>
- [9] Apple, "About Applebot" (Applebot-Extended). <https://support.apple.com/en-us/119829>
- [10] Perplexity, "Perplexity Crawlers" (PerplexityBot, Perplexity-User). <https://docs.perplexity.ai/guides/bots>
- [11] Amazon, "Amazonbot" (Amazonbot, Amzn-SearchBot, Amzn-User). <https://developer.amazon.com/amazonbot>
- [12] Common Crawl, "CCBot." <https://commoncrawl.org/ccbot>
- [13] S. Longpre, R. Mahari, A. Lee, C. Lund, et al., "Consent in Crisis: The Rapid Decline of the AI Data Commons," arXiv:2407.14933, 20 July 2024. <https://arxiv.org/abs/2407.14933> . Compiled as record consent-in-crisis-c4-restricted-share in AIData.Broker, "Supply of AI training data" (dataset, CC BY 4.0), <https://aidata.broker/datasets/supply.json>
- [14] Cloudflare, "The 2025 Cloudflare Radar Year in Review," 15 December 2025. <https://blog.cloudflare.com/radar-2025-year-in-review/> . Compiled as record cloudflare-radar-2025-crawl-to-refer-anthropic-openai in the supply dataset cited in [13].
- [15] Directive (EU) 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market, Article 4(1) and (3). <https://eur-lex.europa.eu/eli/dir/2019/790/oj>
- [16] Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 53(1)(c). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- [17] In re OpenAI, Inc. Copyright Infringement Litigation, No. 1:25-md-03143-SHS-OTW (S.D.N.Y. 18 December 2025); Ziff Davis, Inc. v. OpenAI, Inc., No. 1:25-cv-04315, Dkt. 302, opinion and order. <https://storage.courtlistener.com/recap/gov.uscourts.nysd.643043/gov.uscourts.nysd.643043.302.0.pdf> . Compiled in AIData.Broker, "Lawsuits and rulings over AI training data" (dataset, CC BY 4.0), <https://aidata.broker/datasets/cases.json>
- [18] Reddit, Inc. v. SerpApi LLC, No. 25 Civ. 8736 (PAE) (S.D.N.Y. 31 July 2026), Dkt. 104, opinion and order, at 6. <https://storage.courtlistener.com/recap/gov.uscourts.nysd.651592/gov.uscourts.nysd.651592.104.0.pdf>
- [19] Broadcast Dialogue, "Canadian news publishers sue Anthropic over AI copyright infringement," 14 October 2026. <https://broadcastdialogue.com/canadian-news-publishers-sue-anthropic-over-ai-copyright->

infringement/ . Compiled in the dataset cited in [17].

[20] Encyclopaedia Britannica, Inc. and Merriam-Webster, Inc. v. Perplexity AI, Inc., No. 1:25-cv-07546 (S.D.N.Y.), complaint, 10 September 2025.

<https://storage.courtlistener.com/recap/gov.uscourts.nysd.649196/gov.uscourts.nysd.649196.1.0.pdf> .

Compiled in the dataset cited in [17].

[21] AIData.Broker, "What Is Training Data Provenance, and Why Buyers Now Ask for It," guide, 17 October 2026 (on Article 53 of the AI Act and California Civil Code section 3111).

<https://aidata.broker/guides/training-data-provenance>

Statements

Use of AI. This paper was prepared with generative AI tools, which compiled and analysed the open datasets and drafted the text. The author takes responsibility for its content.

Licence: CC BY 4.0. Cite as: Heymans, Y. (2026). Which websites block AI crawlers: a census of robots.txt rules on 6,094 lower-ranked sites, October 2026. AIData.Broker, 18 October 2026. <https://aidata.broker/research/paper-crawlers>