

Who May Read the Web?

Measuring How Top Websites Admit, Refuse, and Invite AI Crawlers and Agents (October 2026)

Max Song October 10, 2026 Version 1.0 79 min read

Self-published preprint · Not peer reviewed · Prepared with AI assistance

Abstract

Websites now say two opposite things to automated readers. In robots.txt many refuse the crawlers of AI companies; in a newer set of files and signals some invite AI systems and agents in. This report measures both on one stratified random sample of 10,000 Tranco domains, collected in one pass on 2026-10-10 by a crawler that declares itself, under a protocol frozen before any request was sent. The research identity obtained a robots.txt with rules from 56.8% of sampled domains and was refused by 7.6%. Among sites with rules, GPTBot is explicitly refused by 15.8% of the top thousand and by 6.9% across the top million; 11.9% refuse at least one AI token by name. Refusal follows purpose: GPTBot is refused while OAI-SearchBot is admitted on 398 sites, the reverse on 4. Intermediary-written blocks account for every AI refusal on only 1.9% of refusing sites, but 16.3% of files with rules carry a platform's default text. A valid llms.txt is served by 8.6% of sampled domains and any valid invitation by 9.4%; counting status 200 alone would overstate llms.txt 2.0× and the API catalog 28.1×. Of invited sites, 6.5% refuse a crawler of the audience they invite. The policy file was answered for a browser string but not for the research identity on 2.4% of sites, against 0.8% the other way. All six pre-registered tests fall below the Bonferroni threshold. Code, sample and per-site data are published.

中文摘要

网站如今对自动化读者说着两种相反的话：robots.txt 里拒绝 AI 公司的爬虫，一批新的文件与信号又邀请 AI 系统和智能体进来。本研究在同一份分层随机样本上同时测量两者：10,000 个 Tranco 域名，2026-10-10 一趟采完，爬虫亮明研究身份，方案在发出任何请求之前冻结。研究身份从 56.8% 的抽样域名取得了带规则的 robots.txt，被 7.6% 拒绝。在有规则的站点中，前一千名里 15.8% 点名完全禁止 GPTBot，百万名的分层加权估计为 6.9%；11.9% 点名拒绝至少一个 AI 令牌。拒绝按用途区分：禁 GPTBot 却放行 OAI-SearchBot 的站点有 398 个，反向只有 4 个。中介代写的块只在 1.9% 的拒绝站点里包揽了全部 AI 拒绝，但有规则的文件中 16.3% 含有某平台的默认文本。8.6% 的抽样域名提供了结构有效的 llms.txt，9.4% 发布了至少一种有效邀请；只看 200 状态码会把 llms.txt 高估 2.0×、把 API catalog 高估 28.1×。

发出邀请的站点中，6.5% 同时拒绝了它所邀请的那一类爬虫。策略文件对浏览器字符串有应答、对研究身份没有的站点占 2.4%，反向为 0.8%。六个预注册检验全部低于 Bonferroni 门槛。代码、样本与逐站数据一并公开。

1. Introduction

A website speaks to machines in two voices. The older voice is robots.txt, a plain-text file in which a site tells crawlers which paths they may fetch ¹. Since 2023 that file has filled with the names of crawlers run by AI companies, most of them followed by `Disallow: /`. The newer voice is a set of files and signals that do the opposite: an `llms.txt` that hands a language model a curated map of the site ², a `Content-Signal` line that grants or withholds search, AI input and AI training in one record ³, an API catalog at a registered well-known path ⁴, an Agent Card that tells an agent what the site can do for it ⁵. One voice refuses; the other invites.

Both voices are consequential in October 2026. In the United States a federal court held in December 2025 that robots.txt is a request rather than a technological measure that controls access ⁶; in the European Union the voluntary Code of Practice for general-purpose AI names the file, as specified in RFC 9309, as the minimum its signatories will honour ⁷; a large intermediary writes robots.txt text on behalf of its customers and changed its defaults in mid-September 2026, three and a half weeks before this measurement ^{8,9}; and the opt-out for the largest search engine's AI features lives in a console that no outside crawl can see ¹⁰. What the files say is therefore a social fact as much as a technical one, and it can be measured.

The study asks five questions on one sample at one moment.

Question	
RQ1 Refusal	How do popular websites treat AI crawlers in robots.txt, token by token? Do they treat the purposes that operators document differently: collecting training data, building an AI search index, and fetching a page because a user asked?
RQ2 Authorship	How much of that robots.txt text did the site write, and how much was written by an intermediary or copied from a template?
RQ3 Invitation	How many sites publish files or signals that invite AI systems and agents, when a file is counted only if it is what it claims to be?
RQ4 Consistency	How often does one site both invite and refuse the same purpose?
RQ5 Reachability	Can a crawler that says what it is read the policy file at all, compared with a client that sends an ordinary browser string?

The design is a cross-sectional measurement: 10,000 registrable domains drawn by a public hash from the Tranco list ¹¹, stratified by rank; at most twenty-six `GET` requests per site, in one pass, from one vantage point, on 2026-10-10; validators that reject HTML pages, empty bodies and catch-all responses before any file is counted; and a protocol, frozen on 2026-10-05 before any request was sent, that fixed the definitions, the six primary outcomes and the six tests reported here.

The study is descriptive. It claims no priority for any single quantity: every rate reported here has been published by someone for 2025 or 2026, among them Liu et al. for robots.txt rules on Tranco sites to October 2024¹², Hoffmann et al. for robots.txt and llms.txt on about four million domains, with data that the record of their published artefact dates to October 2025¹³, Bouchaud and Ramaciotti for disallow rates by rank bucket in 2025¹⁴, Cloudflare Radar for weekly aggregates^{15,16}, and AI Visibility for a vendor-defined file set in October 2026¹⁷. What was not found in the searches described in Appendix B is one study that puts a purpose-aware typology of robots.txt, standards-track invitation files, a site-level rate of contradiction and an identity-paired test of the policy file itself on one ranked sample with open per-site data. That combination, and the care taken over what counts as a file, is the contribution.

The study measures what sites state, not whether any crawler complies; it sees one vantage point on one day; and its second identity is a header, not a browser.

2. Background

2.1 robots.txt and RFC 9309

The Robots Exclusion Protocol was standardised in 2022 as RFC 9309¹. A file at `/robots.txt` consists of groups: one or more `User-agent` lines followed by `Allow` and `Disallow` rules. A crawler matches its product token case-insensitively, merges every group that names it, and falls back to the `*` group only when no group names it. The most specific rule, measured in octets, wins; when an `Allow` and a `Disallow` are equivalent the `Allow` should win; a path that no rule matches is allowed, and `/robots.txt` itself is always allowed. Lines the crawler does not understand must not end a group, which is what makes `Sitemap`, `Crawl-delay` and the newer `Content-Signal` and `License` records legal extensions. Two rules about fetching matter for this study. A 4xx answer, 401 and 403 included, means the file is unavailable and the crawler may fetch anything; a 5xx answer or a network failure means the file is unreachable and the crawler must assume everything is disallowed.

Google documents where its own parser departs from or adds to the RFC: typo tolerance for field names, truncation of a token at its first invalid character, a global group for `*` followed by trailing text, and a staged treatment of 5xx answers that eventually falls back to no restrictions¹⁸. Its open-source implementation is the reference against which this study's parser was checked¹⁹. Earlier measurement work found home-grown parsers to be a source of error on the order of a tenth of files¹², hence the check.

The RFC is explicit that its rules are not access authorisation. Compliance has been measured and found selective: on one university's websites, crawlers' compliance varied by directive and by class of bot²⁰; a vendor that sells bot paywalls reports that about fifteen per cent of AI bot fetches on its publisher customers' sites in the first half of 2026 ignored an active disallow²¹. Per-crawler favouritism in robots.txt is not new; two census-style studies from 2007 and 2008 counted it^{22,23}, and what is new since 2023 is the kind of crawler named.

2.2 What AI crawlers say they are for

The study classes crawler tokens by the purpose their operator documents, because the question of RQ1 is whether sites separate purposes, and they can only separate what operators expose as separate tokens. Eight operators now publish such splits, and a ninth documents one token without a split. OpenAI documents `GPTBot` for training, `0AI-SearchBot` for search features and `ChatGPT-User` for actions a user takes ²⁴; Anthropic documents `ClaudeBot`, `Claude-SearchBot` and `Claude-User` ²⁵; Amazon `Amazonbot`, `Amzn-SearchBot` and `Amzn-User` ²⁶; Mistral AI `MistralAI-Training`, `MistralAI-Index` and `MistralAI-User` ²⁷; Perplexity `PerplexityBot` and `Perplexity-User` ²⁸; Meta `meta-externalagent`, `meta-webindexer` and `meta-externalfetcher` ²⁹; Apple `Applebot` and a control-only `Applebot-Extended` ³⁰; Google `Googlebot`, a control-only `Google-Extended`, and user-triggered fetchers such as `Google-Agent` ^{31,32}; Common Crawl documents `CCBot` without stating a use within these three classes ³³.

Five of these operators say in their own documentation that their user-triggered fetcher may not follow robots.txt: OpenAI for `ChatGPT-User`, Google for its user-triggered fetchers, Meta for `meta-externalfetcher`, Amazon for `Amzn-User` and Perplexity for `Perplexity-User`. A `Disallow` aimed at those tokens is therefore a preference the operator has said it may not honour. Two tokens, `Google-Extended` and `Applebot-Extended`, are controls that never appear in an HTTP request and can be observed only in robots.txt.

One blind spot is structural and limits everything this report says about Google. Googlebot crawls for Search and for the AI features inside Search alike, and Google states that no separate token exists for the latter ³⁴. The control that lets a site withhold its content from AI Overviews and AI Mode is a setting in Search Console, available worldwide since 31 August 2026 ¹⁰; the United Kingdom's competition authority has ordered Google to provide such controls under a conduct requirement that takes effect on 3 December 2026 ³⁵. Nothing fetched from a website shows that choice. Statements about Google in this report are limited to the training-side token `Google-Extended`.

Community lists classify tokens differently from operators. The `ai.robots.txt` project publishes a single group of about 180 tokens, all disallowed ³⁶; the Known Agents directory, formerly Dark Visitors, documents a generated robots.txt with one group per agent ³⁷ and an index of blocking rates ³⁸; Cloudflare sorts the crawlers it manages into Search, Agent and Training ³⁹. The study carries two purpose schemes side by side, one from operators' documentation and one derived from Cloudflare's published categories and lists, and names the scheme next to every purpose figure.

2.3 The invitation layer

The files and signals that invite machines in are young, and most are not standards. The study groups them by status.

Registered paths. The IANA registry of well-known URIs holds permanent registrations for `api-catalog`, the two OAuth metadata documents, `openid-configuration`, `security.txt` and `agent-card.json`, and a provisional one for `tdmrep.json` ⁴⁰. The API catalog of RFC 9727 is a linkset of a site's APIs ⁴. The A2A Agent Card describes an agent's name, skills and interfaces; its path was renamed from `agent.json` in 2025, so both paths are probed ⁵.

Proposals, drafts and vendor conventions. `llms.txt` is a Markdown file whose required part is a level-one heading, proposed in 2024 and revised to a second version in August 2026²; `llms-full.txt` is a convention of documentation platforms that the proposal itself does not define. Markdown content negotiation answers a request with `Accept: text/markdown` by a Markdown rendering of the page, as Cloudflare's Markdown for Agents does⁴¹. The Model Context Protocol's specification of 2026-07-28 defines no well-known discovery file; a proposal under review on 2026-10-04 routes domain-level discovery through an AI Catalog at `/.well-known/ai-catalog.json`⁴², and the Agentic Resource Discovery proposal of August 2026 renames that catalog to `/.well-known/ard.json` and adds an `Agentmap` record for `robots.txt`⁴³; an older `/.well-known/mcp/server-card.json` survives in Cloudflare's scanner¹⁶. The Agent Skills index is a Cloudflare draft⁴⁴; `auth.md` is a convention that scanner checks; `agents.json`⁴⁵, `ai-plugin.json`, the manifest of a retired plugin programme⁴⁶, and `ai.txt`, a 2023 proposal whose generator still answers⁴⁷, have no structural rule that could be verified against a specification and are counted on the lenient tier only.

Signals inside robots.txt. Cloudflare's Content Signals Policy adds a `Content-Signal` record with three keys, `search`, `ai-input` and `ai-train`, each `yes` or `no`; an absent key grants nothing and withholds nothing^{3,48}. Cloudflare's individual Internet-Draft that proposed the three labels to the IETF, dated 1 October 2025, expired on 4 April 2026 without a successor⁴⁹. The IETF's AIPREF working group drafts a different vocabulary, `train-ai`, `ai-use` and `search` with values `y` and `n`, carried in a `Content-Usage` header or `robots.txt` rule; both documents were active working-group drafts, not RFCs, on 2026-10-04^{50,51}. The two vocabularies are never translated into each other in this report. Really Simple Licensing 1.0, an industry specification published in December 2025, adds a `License:` line that points at an XML licence document⁵². The W3C TDM Reservation Protocol, a Community Group report of May 2024, reserves text-and-data-mining rights through a JSON file, a header or a meta element⁵³. An expired Microsoft draft once defined `DisallowAITraining` and `AllowAITraining` lines⁵⁴; the study records them as a legacy signal.

Infrastructure that agents need but that invites no one. OAuth protected-resource metadata⁵⁵ and authorization-server metadata⁵⁶, with OpenID configuration, signal that a site has an API with delegated authorisation. `security.txt`⁵⁷ is a baseline file with a known adoption history. The key directory of Web Bot Auth is published by parties that send signed bot traffic, not by sites that receive it, and was not probed⁵⁸. NLWeb exposes conversational endpoints rather than a static file and cannot be measured by a `GET` of a fixed path⁵⁹.

Every one of these is counted three ways in this report, and only the strictest count is reported as adoption (Section 3.3).

2.4 Earlier measurements

Academic work has measured the refusal layer at several moments. Liu et al. followed `robots.txt` rules for AI user agents on sites that stayed in the Tranco top 100,000 across monthly lists to October 2024, counting only explicit rules, documented explicit allow rules for GPTBot on dozens of sites, and found that 14 per cent of the top 10,000 blocked pages for one AI company's crawler strings but not for a browser string¹². Hoffmann et al. measured `robots.txt`, `ai.txt`, `llms.txt` and `llms-full.txt` on roughly four million Tranco domains, with a crawl that the record of their published artefact dates to October 2025,

and report cross-file conflicts; only the abstract and code were read for this report, so none of its numbers is used here ¹³. Bouchaud and Ramaciotti fetched robots.txt for the Chrome UX Report's top million in September 2025 and report GPTBot fully blocked on 10.6 per cent of them, 25.2 per cent of the top thousand, with retrieval agents blocked less than training agents ¹⁴. Steinacker-Olsztyn, Gosain and Dao compared reputable and misinformation news sites in May 2025 with an identified crawler and a control request ⁶⁰; Longpre et al. weighted restrictions by training-corpus tokens to April 2024 ⁶¹; Fletcher counted blocking among 150 top news sites at the end of 2023, letting a wildcard block count ⁶²; and Dinzinger, Heß and Granitzer counted 45 hosts serving `tdmrep.json` in a Common Crawl snapshot from the end of 2023 ⁶³.

The HTTP Archive's Web Almanac counts agents *named* in robots.txt rather than disallowed, and reports a valid `llms.txt` on about 2.1 per cent of Chrome UX Report origins in July 2025, with four in ten of those files carrying one SEO plugin's default comment ^{64,65}. Cloudflare Radar publishes, weekly, the share of its top 10,000 domains whose parsed robots.txt fully disallows each AI agent ¹⁵, and since April 2026 the adoption of agent-facing standards on about 160,000 popular domains; its launch post, from a scan of about 200,000 domains, reported robots.txt on 78 per cent of scanned sites, Content Signals on 4 per cent, Markdown negotiation on 3.9 per cent and MCP server cards and API catalogs together on fewer than 15 sites ¹⁶. AI Visibility, on 1 October 2026, scanned a vendor-defined set of ten AI files and the AI rules of robots.txt on about 1,900 Tranco and UK domains, found any such file on 10.2 per cent of the 1,896 reached, defined the state of inviting while blocking as a contradiction without publishing its size, and documented one `llms.txt` that was served only to certain user-agent strings ¹⁷. Three `llms.txt` scans published in September 2026 give 8.3 per cent of the Tranco top 10,000 with 42 per cent unknown ⁶⁶, 8.6 per cent of the Radar top 100,000 with more HTML answers at status 200 than real files ⁶⁷, and 3.5 per cent of an Open PageRank top million ⁶⁸. On an API-provider catalogue, most hosts that answered 200 at the Agent Card path did not serve a card ⁶⁹.

Several interested parties publish the most-cited figures, and the report names them as such wherever their numbers appear. Cloudflare proposed Content Signals, publishes the largest adoption series, and writes robots.txt text for its customers. AI Visibility measures a file set it proposes and sells a checker for. Several `llms.txt` scans come from SEO vendors. Trackers run by Originality.ai ⁷⁰, Known Agents ³⁸, the News Homepages project ⁷¹ and TollBit ²¹ use lists and counting rules that are only partly documented. None of this makes the figures wrong; it makes them comparators rather than ground truth (Section 4.9).

2.5 The social setting, by jurisdiction

The report states what courts and regulators have said and draws no conclusion about what is lawful.

In the United States, a district court held in December 2025, in the consolidated OpenAI copyright litigation, that robots.txt directives are requests and not a technological measure that effectively controls access, and that disregarding them is not circumvention under the anti-circumvention provision of the copyright statute ⁶. In July 2026 another judge of the same court let anti-circumvention claims proceed where a scraper allegedly defeated a JavaScript and CAPTCHA challenge, treating the enforced gate differently from the declared preference ⁷². In August 2026 the Ninth Circuit vacated a preliminary injunction against an AI browsing agent, holding on that record that it was the user, not the agent's maker,

who accessed the site ⁷³. On training itself, two district courts found fair use on records about books in June 2025 ^{74,75}, while the Third Circuit in September 2026 affirmed a finding against fair use for a non-generative system ⁷⁶. The Copyright Office published a pre-publication report on generative AI training in May 2025 ⁷⁷.

In the European Union, the 2019 copyright directive lets rightholders reserve text-and-data-mining rights by machine-readable means without naming a protocol ⁷⁸. The Code of Practice for general-purpose AI, published in July 2025, commits its signatories to crawlers that read and follow robots.txt as specified in RFC 9309 ⁷. The EU Intellectual Property Office's study of May 2025 reports stakeholders' view that the protocol is not an optimal reservation mechanism ⁷⁹; the Commission opened a consultation in December 2025 on which protocols should count, and no list had been published on the pages read for this report ⁸⁰. In the United Kingdom, the government's report of March 2026 says a copyright exception with opt-out is no longer its preferred way forward ⁸¹, and the Competition and Markets Authority imposed in June 2026 a conduct requirement that obliges Google to let publishers withhold Googlebot-collected content from AI training and grounding without a search penalty, in force from 3 December 2026 ³⁵.

The economic context is contested and the report keeps the sides apart. Pew Research Center's browsing data for March 2025 show users clicking a result link on 8 per cent of visits to Google pages with an AI summary against 15 per cent without ⁸²; the Reuters Institute's 2026 survey across 48 markets finds weekly use of AI chatbots for news up from 7 to 10 per cent in a year ⁸³; a publishers' trade body reports a median 10 per cent decline in Google search referrals among its members in mid-2025 ⁸⁴; Google states that total organic click volume was relatively stable year over year ⁸⁵. Cloudflare reported in August 2025 crawl-to-referral ratios that differ by orders of magnitude between operators, and in July 2026 that 52 per cent of crawler requests on its network were for AI training by the crawlers' stated purpose ^{86,87}. Licensing runs alongside: the Associated Press licensed part of its archive in July 2023 ⁸⁸, Reddit announced an expanded partnership with Google in February 2024 ⁸⁹, and Microsoft announced a publisher marketplace in February 2026 ⁹⁰. The costs of being read fall unevenly: the Wikimedia Foundation reported in April 2025 that bots were at least 65 per cent of its most expensive traffic ⁹¹, a documentation host reported 73 terabytes taken by one crawler in a month ⁹², and a proof-of-work gate written in January 2025 for a small git server has spread across open-source infrastructure ⁹³. Cloudflare said in August 2025 that it had observed undeclared crawling by an answer engine that it named; the company's response was not read for this report, and the matter is recorded here as an allegation ⁹⁴.

The intermediary's role most shapes an October 2026 robots.txt measurement. Cloudflare, which says more than a fifth of the web and 36 per cent of the most-visited sites sit behind it ⁸⁷, began blocking AI crawlers by default for new domains in July 2025, said in September 2025 that its managed robots.txt was switched on for over 3.8 million domains and that it would serve a comment-only robots.txt for free-plan sites without one ³, announced in August 2026 a feature that writes dashboard choices into robots.txt between marker comments ⁸, and on 15 September 2026 retired its legacy block-all toggle in favour of settings that distinguish search, agent and training uses ⁹. The counter-movement that invites machines in has its own economy: agent payment protocols announced by Google and by Stripe with OpenAI in September 2025 ^{95,96}, and an acquisition of about 1.9 billion dollars in the market for visibility in search and generative engines ⁹⁷, all assume that sites want to be read by models, while Google's search documentation says no new machine-readable files are needed to appear in its AI features ³⁴.

3. Method

The protocol was written, frozen and dated 2026-10-05 before any request was sent; this section summarises it and the protocol is published beside the data. Where the two differ in wording, the protocol governs.

3.1 Frame, strata and sample

The population is the registrable domains of Tranco list 647LX, generated on 2026-10-04 and downloaded the day after¹¹. Tranco averages several popularity sources over thirty days, which makes it more stable than any one list, but three of its sources include infrastructure names, so a substantial share of entries serve no web content^{98,99}. The study therefore reports every rate against more than one denominator and never replaces a sampled domain that does not answer.

The list was cut into four strata at the rank bounds used by the Chrome UX Report, the Web Almanac and Bouchaud and Ramaciotti: ranks 1 to 1,000 (S1), 1,001 to 10,000 (S2), 10,001 to 100,000 (S3) and 100,001 to 1,000,000 (S4). S1 is a census of 1,000 domains. From S2, S3 and S4 the study drew 3,000, 3,000 and 3,000 domains out of 9,000, 90,000 and 900,000: within a stratum the domains were ordered by the SHA-256 hash of a public salt and the domain, and the leading ones taken. Anyone with the list can recompute the sample, and a script in the materials does so. The sample is 10,000 domains.

Results are reported in three forms, and the form is named each time. A **stratum** figure describes one stratum. A **weighted** figure for the top 10,000, 100,000 or million combines strata with weights equal to their population sizes, 10,000, 100,000 and 1,000,000 domains. A **sample** figure is the unweighted share of the 10,000 sampled domains; it describes the sample, in which the top thousand is a tenth, and estimates no population.

3.2 Instrument and identity

The collector is a Node.js program published with the data. For each site it finds an origin by trying `https://` and `https://www.` and then the same two over `http://`, and sends at most twenty-six `GET` requests to it: the policy file twice, a home page as HTML and once with `Accept: text/markdown`, two control requests for paths that should not exist, and the discovery files and signals listed in the protocol and on the study page. It waits at least one second between requests to a host, follows at most five redirects within the site, retries at most once, cuts bodies at one megabyte, sends no cookie, runs no JavaScript and attempts no challenge.

Every request carries the string `WhoMayReadTheWeb-Research/1.0`, with a link to the author's contact page, as `User-Agent`. One request per site is the disclosed exception: `/robots.txt` is fetched a second time with the string of a common desktop browser, from the same client, with identical headers in identical order; which of the two goes first is fixed per domain by one bit of a hash. No request ever carries the name of another organisation's crawler. The collector reads robots.txt before anything else and requests nothing the file disallows for its own token or, failing that, for `*`, taking the stricter of the RFC reading and the

reading of Google's parser so that a typo cannot undo an operator's opt-out. It leaves a site alone when robots.txt answers 5xx or cannot be reached, when both the policy file and the home page are refused, and when the site keeps answering 429.

Collection ran in four sessions on one day, 2026-10-10 (UTC), from one residential fibre connection in California, United States, over IPv4 only. Every one of the 10,000 domains has a record: 10,000 of 10,000 were collected, and none was excluded at an operator's request (0 of 10,000), stopped by a collector crash, the record state `collector-error` (0 of 10,000), or left without a record (0 of 10,000).

The robots.txt parser was written against RFC 9309 and checked against Google's open-source parser on 649 hostile files, 1,037,102 file-token-path cells, the fixture of the parser's golden test published with the code: a profile that imitates Google's implementation agrees with it on every cell, and the primary RFC profile differs on 51,787 cells, each attributed to one of twelve named differences between the RFC and Google's implementation. The primary reading reported here is the RFC profile; Google's reading is a sensitivity analysis.

A **codebook** frozen with the protocol lists fifty-four crawler tokens with two purpose labels each, the twenty-six probes with their validation rules, twenty-three fingerprints of robots.txt text written by platforms, generators and intermediaries, and sixteen signatures of block and challenge pages, each with an evidence level. A fingerprint counts only once verified, which under the protocol's rule means a cited source for the text and a hand check of twenty matching files from this collection with at least nineteen confirmed. The verification was carried out after the collection and before the final analysis; its outcome is in Section 3.4.

3.3 Definitions

Outcome of a robots.txt request. Each request falls into one of seven groups. Three are *answered*, because the server gave a definitive answer about its policy: `rules` (status 200, not HTML, at least one line with a robots.txt field), `no-rules` (a file with no such line: empty, comments only, or an intermediary's comment-only policy text) and `absent` (404 or 410, a body equal to the site's answer for a random path, an HTML page, or a redirect to another path). Four are not: `refused` (401 or 403 from the origin, a verified block or challenge page, or 429), `error` (5xx, other 4xx, a redirect loop), `unreachable` (DNS, TLS, connection or timeout failure) and `not-collected`. "Sites with robots.txt rules" always means that the request under the research identity ended in `rules`. The RFC's own reading of each outcome, which turns 4xx into permission and failure into prohibition, is reported beside the groups and never substituted for them.

Reading a file for a token. A token is *fully disallowed* when `/` is disallowed for it, *partially disallowed* when `/` is allowed and at least one non-empty `Disallow` survives, and unrestricted otherwise. Under the **explicit** rule only groups that name the token count; a token that no group names has no explicit level. This is the rule of Liu et al. and of Cloudflare Radar. Under the **effective** rule the crawler's own group counts if one names it, else the `*` group, else nothing: what a compliant crawler ends up with, the rule of AI Visibility and of Fletcher's factsheet. Every refusal figure is produced under both rules and over two denominators, sites with rules and all sampled domains, so that it can stand beside studies that chose differently.

Purposes. Fifty-two of the fifty-four tokens are AI tokens; `Googlebot` and `bingbot` are the classic-search baseline, since neither operator documents a token that a site could refuse without leaving the search engine. Under the **first-party scheme**, the primary one, a token's purpose is what its operator documents: `training` (seven tokens), `search-index` (eleven), `user-fetch` (ten); the other twenty-four, among them CCBot, Google-Extended, Bytespider and meta-externalagent, have a mixed or undeclared purpose and form `other`. Under the **community scheme**, derived from Cloudflare's published categories and its managed block, the classes hold eighteen, four and six tokens. The schemes disagree on twenty-seven tokens.

Policy types. A site with rules is typed by whether it fully disallows at least one token of each of the three purposes under the effective rule: `open` (none, and no `other` token either), `other-only`, `training-only`, `training-user`, `training-search`, `all-purposes` (all three, but not every token), `blanket` (every token of every class) and `not-training` (search or user-triggered tokens refused while training is not). Two further types cover the rest of the sample, `no-robots-rules` and `unknown`.

Authorship. An intermediary block is the text between `# BEGIN Cloudflare ...` and `# END Cloudflare ...` marker lines. Every refusal figure, policy type and signal figure is produced twice, for the file as served and for the site's own text with such blocks removed. Each file receives the first authorship class that applies: `intermediary-comment-only`, `intermediary-only`, `intermediary-plus-own`, `platform-template` (contains the default text of a content-management system or site builder, by a verified fingerprint), `generator-template` (contains a published block list or a generator's output, by a verified fingerprint), `unverified-template` (matches only unverified fingerprints), `no-known-template` and `no-directives`. `platform-template` means that the file carries a platform's default text, whether the platform wrote the file or a person kept the default lines while editing, not that the platform wrote it; the residual `no-known-template` does not show that a person at the site wrote the file.

Invitations. Twenty-one files and behaviours are probed and reported in three groups that are never added together without saying so: *registered* (the API catalog of RFC 9727 and the A2A Agent Card), *proposal* (llms.txt and its companions, the agent-discovery drafts and conventions of Section 2.3, the TDMRep file and the RSL licence document) and *infrastructure* (OAuth and OpenID metadata and security.txt, which agents need and which invite no one). Each is counted three ways over all sampled domains: **status 200**, whatever the body; **present**, when the body passes generic rejections (not HTML, not empty, not reached through a path-changing redirect, not a recognised challenge page, not equal to what the site returned for a random path at the root or under `/.well-known/`) and the file's lenient rule; and **valid**, when it also passes a strict rule taken from the file's specification. Only the third is reported as adoption, and the ratio of the first to the third is reported per file. Signals inside robots.txt, `Content-Signal`, `Content-Usage`, `License`, `Agentmap` and the records of an expired draft, are counted when a file carries one; a key's site-wide value is what a record in the `*` group or before any group states for `/`, and a key that no such record states is unstated, which is not "no". A site **has a valid invitation** when it has at least one valid inviting file or a site-wide signal that grants training or AI input.

Conflict. Each inviting artifact addresses one audience: training (`ai-train=yes`, `train-ai=y`), inference at answer time (llms.txt, llms-full.txt, Markdown negotiation, `ai-input=yes`, `ai-use=y`), or agents acting for a user (the API catalog, Agent Cards, the MCP and Agentic Resource Discovery files, the Agent Skills

index, auth.md). A **same-purpose conflict** is a valid invitation whose site refuses at least one listed crawler of that audience: for a file the test is exact, whether the file's own path is disallowed for such a crawler; for a signal it is a full disallow of `/`. A **void** invitation is one where every crawler of the audience is refused. **Co-occurrence** is the broad count: a valid invitation on a site that fully disallows any AI token under the effective rule, whatever the purpose; a site that serves llms.txt for answer-time use and refuses training crawlers is counted here and is not in conflict with itself.

Reachability. The paired comparison is made among collected domains on the outcome *answered*: concordant when both requests were answered or neither was, discordant otherwise. Both requests come from the same HTTP/1.1 client with the same TLS fingerprint and no JavaScript, so the pair isolates the `User-Agent` string.

Categories. Five labels, education, government, news, retail and community, come from open sources only: Wikidata items whose official website is the sampled domain, and domain-name rules for government and education namespaces. News, government and education are the categories reported; retail and community are too thin outside the top strata and are shown for description. The EU proxy is a top-level domain, one of the twenty-seven member states' country codes or `.eu`; it stands for where a site is addressed, not for where its operator is established. A category is compared in a stratum only when it has at least thirty sampled domains there.

3.4 Pre-registration and deviations

The protocol fixed the sample, the probes, the two identities, the request limits, the definitions above, six primary outcomes, eleven sensitivity analyses and the six tests before any request was sent. Ten entries were added to its deviations log; none changed a definition, a request or a test. Before collection, the link in the research `User-Agent` was changed from the study page, which could not be deployed in time, to the author's contact page, which was live throughout. After 1,200 sites, the code that lets a resumed session learn which sites already have a record was found to split lines at U+2028 as well as at a line feed; the reader was changed to split at line feeds only, no site was requested twice, and the run manifest records the new source hash. A reproducible tool for drawing the twenty files of a fingerprint hand check was added. While collection was still running, with 4,200 records written, the analysis pipeline was rehearsed on them and the author saw a rehearsal's headline refusal table by stratum; this is logged so that a reader can weigh it against the rule that no fingerprint is added or verified after results have been seen by stratum. After collection and before the final analysis the verification rule was applied to all twenty-eight unverified codebook entries: five robots.txt fingerprints and seven edge signatures were raised to verified, and sixteen stayed unverified, for want of a vendor source, for a hand check below the threshold, or for want of any match; only evidence levels, source links and description text changed. One entry records a defect found by that check and not corrected, which Section 4.1 and Section 6 describe. After the manuscript was written, a closing review added four entries: the caption and the cell escaping of one appendix table were corrected; two hand-check strings in the codebook were corrected to the counts on which both reviewers agreed, and on that count the raise of one edge signature, CloudFront's, was withdrawn the same day it had been made; and keys were added for the counts that Section 4.1 and Section 4.4 cite. None changed a collection record or a request; the withdrawal moved CloudFront's policy-file refusals from the count of refusals by a verified signature to the unidentified count.

3.5 Statistics

Within a stratum an estimate is a count over a denominator, printed with both, and the 95 per cent interval is Wilson's score interval, or the exact Clopper–Pearson interval when the count or its complement is below ten ¹⁰⁰. S1 is a census: for the list itself it has no sampling error, and its intervals are given under a super-population reading. Weighted figures use the combined ratio estimator for a stratified sample with its linearised variance and a finite-population correction; the census stratum contributes none, and an interval is printed only when at least ten sampled events and ten non-events stand behind it. Unweighted sample figures carry no interval.

Six comparisons are tests, and nothing else in this report is: C1, GPTBot against OAI-SearchBot, and C2, GPTBot against ChatGPT-User, both fully disallowed under the effective rule, paired within site, by the exact McNemar test; C3, explicit full disallow of at least one training token over the four strata, by the linear-by-linear association statistic; C4, the same outcome for news sites against the rest, and C5 for the EU proxy against the rest, both by the Cochran–Mantel–Haenszel test stratified by rank with the Mantel–Haenszel common odds ratio and the Robins–Breslow–Greenland interval; and C6, whether the policy file was answered for the browser string against the research identity, paired within site, by the exact McNemar test. All are two-sided; the family-wise rate of 6 tests is held at 0.05 by Bonferroni, so each is judged at 0.0083; each is reported with its raw p-value and its counts whether or not it falls below the threshold. The paired tests and the trend test are computed on the unweighted sample and ask whether the difference exists among the sampled sites; the per-stratum values show whether its direction holds in each. Every other comparison in this report is descriptive, and comparisons made after the data were seen are labelled exploratory.

3.6 Ethics in brief

The study had no institutional review and was designed against the Menlo Report ¹⁰¹; Section 7 answers the four questions Partridge and Allman ask of a measurement paper ¹⁰² and reports what sites did with the crawler's name.

4. Results

4.1 Collection and data quality

Every sampled domain has a record: 10,000 of 10,000 were collected, none was excluded at an operator's request (0 of 10,000) and none was stopped by a collector crash (0 of 10,000); the three sites that a signature defect, described below, caused the crawler to leave alone early are collected records with fewer probes. Some request received an HTTP answer from 8,584 of 10,000 domains. The share rises down the list, from 79.1% in the top thousand to 88.9% in S4, as expected of a list whose top ranks hold many names that serve no website (Section 3.1): no origin could be reached for 20.5% of S1 against 10.4% of S4. The crawler left 3.8% of sites alone because robots.txt answered 5xx or could not be fetched, 7.0% because both the policy file and the home page were refused, and 0.5% because the site kept answering 429. No robots.txt in the sample names this study's crawler (0 of 10,000); 4.5% of sites disallowed at least one probe for it through a group it fell under, 7.0% in the top thousand.

Table 1. The sample and what was collected, by rank stratum. Counts of sampled domains, with their share of the stratum.

	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)	All
Domains in the list	1,000	9,000	90,000	900,000	1,000,000
Sampled	1,000	3,000	3,000	3,000	10,000
Collected	1,000 (100.0%)	3,000 (100.0%)	3,000 (100.0%)	3,000 (100.0%)	10,000 (100.0%)
Excluded at the operator's request	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Not collected: no usable host name, collector error, or no record	0	0	0	0	0
Some request answered over HTTP	791 (79.1%)	2,533 (84.4%)	2,592 (86.4%)	2,668 (88.9%)	8,584 (85.8%)
robots.txt with rules obtained	519 (51.9%)	1,755 (58.5%)	1,699 (56.6%)	1,709 (57.0%)	5,682 (56.8%)
Left alone: no origin reachable	205 (20.5%)	450 (15.0%)	382 (12.7%)	313 (10.4%)	1,350 (13.5%)
Left alone: robots.txt unreachable (5xx or network failure)	34 (3.4%)	107 (3.6%)	101 (3.4%)	137 (4.6%)	379 (3.8%)
Left alone: robots.txt and home page refused	59 (5.9%)	250 (8.3%)	217 (7.2%)	171 (5.7%)	697 (7.0%)
Left alone: rate limited	7 (0.7%)	19 (0.6%)	12 (0.4%)	10 (0.3%)	48 (0.5%)
A robots.txt group names this study's crawler	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
robots.txt disallowed at least one probe	70 (7.0%)	105 (3.5%)	118 (3.9%)	154 (5.1%)	447 (4.5%)

Under the research identity, a robots.txt with at least one directive was obtained from 5,682 of 10,000 sampled domains: 51.9% of S1 and between 56.6% and 58.5% of the other strata, 56.9% weighted to the top million. The rest splits into a file that states no rule (2.3%; the comment-only policy text that one intermediary serves for sites without a file accounts for 133 of 10,000 sampled domains, most of this group), a definitive absence (15.3%: 9.6% answered 404 or 410, 2.7% returned the same body as for a random path, 2.9% redirected to another path), a refusal (7.6%), an HTTP answer that settles nothing (1.8%) and no HTTP answer at all (16.2%; a DNS failure alone accounts for 11.1% of sampled domains, the larger part of this group). Refusal is highest in S2 (8.9%) and lowest in S4 (6.1%); weighted, 8.7% of the top 10,000 and 6.3% of the top million turned the research identity's request for the policy file away.

Table 2. Outcome of the robots.txt request made under the declared research identity, by rank stratum. Share of all sampled domains; the seven groups are exhaustive and mutually exclusive.

Outcome	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)	All
rules: a robots.txt with at least one directive was obtained	519 (51.9%)	1,755 (58.5%)	1,699 (56.6%)	1,709 (57.0%)	5,682 (56.8%)
no-rules: a file was served that states no rule	9 (0.9%)	36 (1.2%)	71 (2.4%)	111 (3.7%)	227 (2.3%)
absent: the server answered and there is no robots.txt	152 (15.2%)	366 (12.2%)	479 (16.0%)	532 (17.7%)	1,529 (15.3%)
refused: the request was turned away: 401/403, an edge block or challenge page, or 429	69 (6.9%)	268 (8.9%)	241 (8.0%)	182 (6.1%)	760 (7.6%)
error: an HTTP answer that settles nothing	18 (1.8%)	45 (1.5%)	45 (1.5%)	73 (2.4%)	181 (1.8%)
unreachable: no HTTP answer	233 (23.3%)	530 (17.7%)	465 (15.5%)	393 (13.1%)	1,621 (16.2%)
not-collected: the domain has no collected record	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Three readings of these outcomes are kept apart. The RFC reads the 4xx answers among these refusals, 741 of 10,000 sampled domains and nearly all of the refusals, as permission to fetch anything, and the 5xx and network failures as a prohibition: by its rule 65.0% of sampled domains have rules to follow, 17.8% allow everything and 17.3% disallow everything. Of the refusals, 519 of 10,000 sampled domains were recognised as a block or challenge page of a verified signature; 3.2% of domains received a challenge, 2.4% a block page, 2.0% a plain 401 or 403, and <0.1% a 429; the challenge and block-page shares include the pages of unverified signatures, which are reported below as unidentified. By vendor, in descending order of sites, the verified signatures that turned the policy request away are Cloudflare's three (388 of 10,000 sampled domains, most of the recognised refusals), then Akamai's (108 of 10,000), AWS WAF's (10 of 10,000), Vercel's challenge (6 of 10,000), DataDome's (5 of 10,000) and F5's (2 of 10,000); Anubis and Vercel's deny header matched no policy request. A further 39 of 10,000 were turned away by pages of unverified signatures and are reported as unidentified, not under a vendor's name. The per-site table records each site's signature; the verification log gives, for the seven signatures raised after the collection, their matches over all probes of all sampled domains: Akamai 370 responses from 130 domains, DataDome 226 from 29, AWS WAF 146 from 20, Vercel's deny header 34 from 2, Vercel's challenge 13 from 7, F5 5 from 3, Anubis 1 from 1. Two need a gloss. The CloudFront signature, whose 108 matches on 34 domains are all 403 pages generated by CloudFront, stayed unverified: the two reviewers agreed on 16 of its 20 drawn files, below the 19 the rule requires, because 3 of the files are a "Bad request" variant from a misconfigured distribution, refused to both identities alike, and 1 was relayed by another vendor's edge; a raise made on the reading that any CloudFront-generated 403 is a refusal was withdrawn the same day, and its policy-file refusals are among the unidentified ones. Vercel's deny header records a path rule, not a judgement on the identity: its 34 matches come from 2 domains that refuse every `/.well-known/` and control path and serve robots.txt and the home page to both identities. The other two unverified signatures

with any match are described below. Counted among the absences are 33 of 10,000 sampled domains whose bodies are structurally a robots.txt but were reached through a redirect that changes the path; the protocol counts them as absent, and a reader who prefers the RFC's view can add them.

Two defects of the instrument surfaced in the fingerprint hand check and were not corrected, by the author's decision not to alter stored outcomes after the results had been seen; both are disclosed with the verification log's counts. One edge signature, for a vendor's injected script path, matched at any status and content type: it matched 130 responses from 20 domains, among them the real robots.txt of 6 sites, served at status 200 as plain text and recorded as refused with their rules unparsed, and the 404 pages of 2 sites recorded as refused instead of absent; 3 of these sites were then left alone because the home page matched too. If corrected, at most 6 sites would move from refused to obtained and 2 from refused to absent, 0.08 per cent of the sample. In the other direction, one CAPTCHA signature was confirmed on 18 of 20 drawn files; the other 2, one site's responses, could not be judged because their markers lie beyond the stored 2-kilobyte head of the body, and under the rule a file that cannot be judged counts as not confirmed, so the signature was not raised. Over all probes it matched 55 responses from 11 domains; the policy-file refusals among them, recorded in the per-site table as challenge pages of an unverified signature, are reported among the unidentified refusals rather than under a vendor's name, and the verified-edge count above is to that extent a lower bound.

The browser identity obtained a file with rules from 5,831 of 10,000 domains and was refused by 6.2%, against 56.8% and 7.6% for the research identity; Section 4.7 takes the pair apart site by site.

4.2 The refusal layer

Among the 5,682 sites with robots.txt rules, a group that names GPTBot disallows everything for it on 15.8% of the top thousand (95% CI 12.9–19.2%) and 6.6% of S4, with every stratum in Table 3; weighted, 11.7% of the top 10,000 (95% CI 10.6–12.8%) and 6.9% of the top million (95% CI 5.8–8.0%). CCBot is refused by name slightly more often in the top thousand and slightly less often across the top million (17.3% of S1; 6.4% of the top million), ClaudeBot slightly less on both counts (15.6%; 5.0%) and Google-Extended less again (13.1%; 3.5%). Over all sampled domains, the denominator that includes sites with no file, every figure is roughly halved: GPTBot 6.8% of the top 10,000 and 3.9% of the top million. Under the effective rule, which adds sites whose * group disallows everything, GPTBot is refused on 22.2% of S1 and 14.1% of the top million.

Table 3. Full disallow of four widely tracked tokens, under two counting rules and two denominators. "Explicit": a group that names the token disallows "/". "Effective": what a compliant crawler ends up with after falling back to the "*" group. Percentages with 95% intervals; the last two columns are stratum-weighted estimates.

Token	Rule	Denominator	S1	S2	S3	S4	Top 10,000	Top 1,000,000
GPTBot	explicit	robots.txt with rules	15.8% (12.9–19.2%)	11.3% (9.9–12.8%)	9.9% (8.6–11.5%)	6.6% (5.5–7.8%)	11.7% (10.6–12.8%)	6.9% (5.8–8.0%)
GPTBot	explicit	all sampled	8.2% (6.7–10.1%)	6.6% (5.8–7.5%)	5.6% (4.9–6.5%)	3.7% (3.1–4.5%)	6.8% (6.1–7.4%)	3.9% (3.3–4.5%)
GPTBot	effective	robots.txt with rules	22.2% (18.8–25.9%)	14.5% (13.0–16.3%)	14.7% (13.1–16.4%)	14.0% (12.5–15.8%)	15.2% (14.0–16.4%)	14.1% (12.6–15.6%)
GPTBot	effective	all sampled	11.5% (9.7–13.6%)	8.5% (7.6–9.6%)	8.3% (7.4–9.3%)	8.0% (7.1–9.0%)	8.8% (8.1–9.5%)	8.0% (7.2–8.9%)
CCBot	explicit	robots.txt with rules	17.3% (14.3–20.8%)	12.8% (11.3–14.4%)	9.3% (8.0–10.8%)	6.0% (5.0–7.3%)	13.2% (12.0–14.3%)	6.4% (5.4–7.4%)
CCBot	explicit	all sampled	9.0% (7.4–10.9%)	7.5% (6.6–8.5%)	5.3% (4.5–6.1%)	3.4% (2.8–4.1%)	7.6% (6.9–8.3%)	3.6% (3.1–4.2%)
CCBot	effective	robots.txt with rules	25.4% (21.9–29.4%)	16.1% (14.5–17.9%)	13.9% (12.4–15.7%)	9.6% (8.3–11.1%)	17.0% (15.7–18.2%)	10.1% (8.8–11.3%)
CCBot	effective	all sampled	13.2% (11.2–15.4%)	9.4% (8.4–10.5%)	7.9% (7.0–8.9%)	5.5% (4.7–6.3%)	9.8% (9.0–10.6%)	5.7% (5.0–6.5%)
Google - Extended	explicit	robots.txt with rules	13.1% (10.5–16.3%)	8.7% (7.5–10.1%)	6.8% (5.7–8.1%)	3.2% (2.4–4.1%)	9.1% (8.1–10.1%)	3.5% (2.8–4.3%)
Google - Extended	explicit	all sampled	6.8% (5.4–8.5%)	5.1% (4.4–5.9%)	3.9% (3.2–4.6%)	1.8% (1.4–2.3%)	5.3% (4.7–5.8%)	2.0% (1.6–2.5%)
Google - Extended	effective	robots.txt with rules	19.5% (16.3–23.1%)	12.0% (10.5–13.6%)	11.4% (10.0–13.0%)	10.6% (9.3–12.2%)	12.6% (11.5–13.8%)	10.7% (9.4–12.1%)
Google - Extended	effective	all sampled	10.1% (8.4–12.1%)	7.0% (6.1–8.0%)	6.5% (5.6–7.4%)	6.1% (5.3–7.0%)	7.3% (6.6–8.0%)	6.1% (5.3–6.9%)
ClaudeBot	explicit	robots.txt with rules	15.6% (12.7–19.0%)	10.2% (8.9–11.7%)	8.2% (7.0–9.6%)	4.6% (3.7–5.7%)	10.7% (9.6–11.7%)	5.0% (4.1–5.9%)
ClaudeBot	explicit	all sampled	8.1% (6.6–10.0%)	6.0% (5.2–6.9%)	4.7% (4.0–5.5%)	2.6% (2.1–3.3%)	6.2% (5.6–6.8%)	2.9% (2.3–3.4%)

Token	Rule	Denominator	S1	S2	S3	S4	Top 10,000	Top 1,000,000
ClaudeBot	effective	robots.txt with rules	22.4% (19.0–26.1%)	13.6% (12.1–15.3%)	13.0% (11.5–14.7%)	12.2% (10.7–13.8%)	14.4% (13.2–15.6%)	12.3% (10.9–13.7%)
ClaudeBot	effective	all sampled	11.6% (9.8–13.7%)	8.0% (7.1–9.0%)	7.4% (6.5–8.4%)	6.9% (6.1–7.9%)	8.3% (7.6–9.0%)	7.0% (6.2–7.8%)

The gradient by rank is the most regular pattern in the table: every token is refused more often at the top of the list, and for each of the four the largest single step is between the top thousand and the next nine thousand. It is also the pattern of every earlier study that reports rank buckets; C3 in Section 4.8 is its pre-registered test. GPTBot is *named* by 24.9% of S1 files and 15.3% of the sample's; among the sample's sites with rules, 3.1% name GPTBot and disallow only part of the site, and 2.4% name it in a group that restricts nothing, an explicit welcome; ClaudeBot gets the same welcome on 2.1%.

At least one of the fifty-two AI tokens is refused by name on 27.0% of S1 sites with rules (95% CI 23.3–31.0%) and 11.5% of S4; weighted, 20.2% of the top 10,000 and 11.9% of the top million (95% CI 10.6–13.3%); over all sampled domains, 11.7% and 6.8%. The effective rule lifts the top-million figure among sites with rules to 18.9%, because 5.8% of sites with rules disallow everything for `*`, and 3.2% do so without naming any AI token: their refusal of AI crawlers is a by-product of refusing everyone. A `*` group is present on 97.1% of files. The classic-search baseline stays low: Googlebot or bingbot is fully disallowed after the fallback on 3.9% of sites with rules, Googlebot by name on 0.2%, and Googlebot is named in a group that restricts nothing on 4.3% of sites, more than any AI token is welcomed.

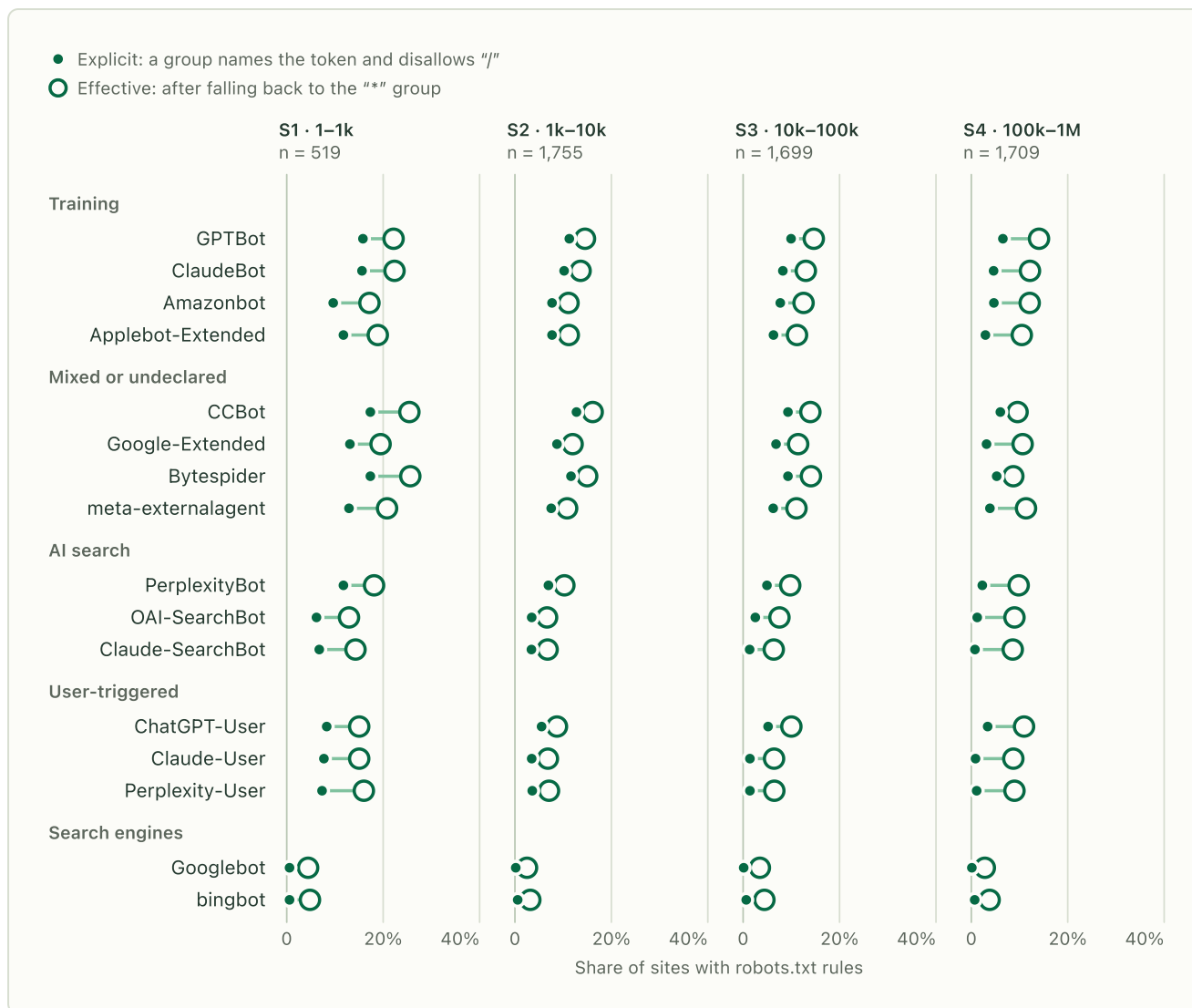


Figure 1. Full disallow of selected crawler tokens, by rank stratum (S1: ranks 1–1,000; S2: 1,001–10,000; S3: 10,001–100,000; S4: 100,001–1,000,000) and counting rule. Filled dot: a group that names the token disallows "/" (explicit rule). Open ring: what a compliant crawler ends up with after falling back to the "*" group (effective rule). Denominator: sites whose robots.txt with rules was obtained under the research identity (n in each panel). Tokens are grouped by the purpose their operator documents; Googlebot and bingbot are a baseline, not AI tokens. Values are in Table A7; four tokens are given with intervals and stratum-weighted estimates in Table 3.

Refusal follows the purpose the operator declares. Under the first-party scheme and the effective rule, at least one training token is refused on 28.9% of S1 sites with rules (95% CI 25.2–32.9%) and 17.9% of the sample's; at least one AI-search token on 27.4% and 16.4%; at least one user-triggered fetcher on 20.4% and 11.7%. The gap is wider at the token level: OAI-SearchBot is refused by name on 6.2% of S1 and 1.4% of the top million against GPTBot's 15.8% and 6.9%, and ChatGPT-User on 8.3% of S1. The * fallback adds about the same five to six points to every class, so the order of the three classes is the same under the explicit rule: at least one training token refused by name on 12.4% of sites with rules, a search token on 11.0%, a user-triggered fetcher on 6.2%. Every one of the seven training tokens is refused on 7.1% of S1 and 5.6% of the sample's sites with rules. The class of mixed or undeclared purpose is refused most of all (31.2% of S1; 19.0% of the sample's), because it contains CCBot, Google-Extended, Bytespider and meta-externalagent, the tokens that block lists have carried longest; moving those tokens into training under the

community scheme raises the training figure to 33.5% of S1 and 21.5% of the sample. Explicit refusal of at least one training token, the outcome of tests C3 to C5, runs from 21.0% in S1 to 8.0% in S4, 8.4% weighted to the top million.

Table 4. Refusal by purpose class: share of sites with robots.txt rules that fully disallow tokens of each class after the "*" fallback. The two schemes assign tokens to purposes from the operators' own documentation (first-party) and from community lists. The last two columns pool the sample, without and with intermediary-written blocks.

Scheme	Purpose class	Refused	S1	S2	S3	S4	Sample, own text only	Sample
first-party	training (7 tokens)	at least one refused	28.9%	18.2%	16.5%	15.4%	17.5%	17.9%
first-party	training (7 tokens)	every one refused	7.1%	3.5%	5.1%	7.8%	5.6%	5.6%
first-party	search-index (11 tokens)	at least one refused	27.4%	17.0%	15.0%	13.9%	16.1%	16.4%
first-party	search-index (11 tokens)	every one refused	6.2%	3.2%	4.9%	3.9%	4.2%	4.2%
first-party	user-fetch (10 tokens)	at least one refused	20.4%	11.2%	10.2%	11.1%	11.6%	11.7%
first-party	user-fetch (10 tokens)	every one refused	6.6%	3.3%	4.8%	7.7%	5.4%	5.4%
first-party	other (24 tokens)	at least one refused	31.2%	19.8%	17.4%	15.9%	18.6%	19.0%
first-party	other (24 tokens)	every one refused	4.8%	2.6%	4.2%	3.4%	3.5%	3.5%
community	training (18 tokens)	at least one refused	33.5%	22.5%	19.8%	18.6%	21.2%	21.5%
community	training (18 tokens)	every one refused	6.0%	3.0%	4.5%	3.6%	3.9%	3.9%
community	search-index (4 tokens)	at least one refused	19.7%	10.9%	10.4%	10.1%	11.3%	11.3%
community	search-index (4 tokens)	every one refused	6.2%	3.4%	5.1%	4.0%	4.3%	4.3%
community	user-fetch (6 tokens)	at least one refused	19.8%	11.7%	10.3%	11.1%	11.7%	11.8%
community	user-fetch (6 tokens)	every one refused	10.6%	4.3%	5.7%	8.1%	6.4%	6.4%
community	other (24 tokens)	at least one refused	25.2%	14.0%	12.2%	11.5%	13.6%	13.7%
community	other (24 tokens)	every one refused	5.2%	2.8%	4.3%	7.5%	4.9%	4.9%

With intermediary-written blocks removed, every figure moves by a few tenths of a point: the training figure above falls from 17.9% to 17.5% of sites with rules, and GPTBot's explicit figure from 9.9% to 9.5%. The policy file is, overwhelmingly, the site's own words; Section 4.4 says how many of those words a platform supplied.

4.3 Policy types

Typing each site with rules by the three purposes it refuses, 344 of 519 S1 sites (66.3%) refuse no AI token at all, against 81.2% of S4 and 81.1% weighted to the top million. Among sites that do refuse, the commonest pattern is to refuse all three purposes without refusing every token, `all-purposes` (13.9% of S1; 7.0% of the sample), followed by `blanket` files that refuse every listed token through `*` or by name (6.2% of S1; 4.1% of the sample), `training-search` (5.0%; 3.2%) and `training-only` (3.7%; 3.1%). A site that refuses a training token and also a user-triggered fetcher but admits AI search is rare (0.5% of the sample), and so is `not-training`, refusing search or user-triggered tokens while admitting training (2.5% of S1; 2.3% of the sample). The `other-only` type, refusal confined to tokens of undeclared purpose, is 2.3% of S1 and 1.6% of the sample. Over all sampled domains, 17.6% state no rule and 25.6% could not be typed because the file was refused, failed or unreachable.

Table 5. Site-level policy types over the three purpose classes (first-party scheme, effective rule), among sites with robots.txt rules: for the robots.txt as served, and for the same file with intermediary-written blocks removed (own text). Within each reading the types are exhaustive and mutually exclusive.

Policy type	File read	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)	All
open: no AI token is fully disallowed	as served	344 (66.3%)	1,357 (77.3%)	1,359 (80.0%)	1,388 (81.2%)	4,448 (78.3%)
other - only: only tokens of mixed or undeclared purpose are fully disallowed	as served	12 (2.3%)	31 (1.8%)	21 (1.2%)	26 (1.5%)	90 (1.6%)
training - only: training refused; AI search and user - triggered fetching admitted	as served	19 (3.7%)	58 (3.3%)	56 (3.3%)	45 (2.6%)	178 (3.1%)
training - user: training and user - triggered fetching refused; AI search admitted	as served	1 (0.2%)	8 (0.5%)	6 (0.4%)	12 (0.7%)	27 (0.5%)
training - search: training and AI search refused; user - triggered fetching admitted	as served	26 (5.0%)	68 (3.9%)	53 (3.1%)	32 (1.9%)	179 (3.2%)
all - purposes: at least one token of each purpose refused, but not every token	as served	72 (13.9%)	129 (7.4%)	86 (5.1%)	112 (6.6%)	399 (7.0%)
blanket: every token of all three purposes refused	as served	32 (6.2%)	57 (3.2%)	80 (4.7%)	63 (3.7%)	232 (4.1%)
not - training: AI search or user - triggered fetching refused while no training token is	as served	13 (2.5%)	47 (2.7%)	38 (2.2%)	31 (1.8%)	129 (2.3%)
open: no AI token is fully disallowed	own text	344 (66.3%)	1,361 (77.5%)	1,366 (80.4%)	1,395 (81.6%)	4,466 (78.6%)
other - only: only tokens of mixed or undeclared purpose are fully disallowed	own text	12 (2.3%)	31 (1.8%)	21 (1.2%)	26 (1.5%)	90 (1.6%)
training - only: training refused; AI search and user - triggered fetching admitted	own text	19 (3.7%)	58 (3.3%)	54 (3.2%)	45 (2.6%)	176 (3.1%)
training - user: training and user - triggered fetching refused; AI search admitted	own text	1 (0.2%)	8 (0.5%)	6 (0.4%)	12 (0.7%)	27 (0.5%)
training - search: training and AI search refused; user - triggered fetching admitted	own text	26 (5.0%)	66 (3.8%)	48 (2.8%)	29 (1.7%)	169 (3.0%)
all - purposes: at least one token of each purpose refused, but not every token	own text	72 (13.9%)	127 (7.2%)	84 (4.9%)	107 (6.3%)	390 (6.9%)
blanket: every token of all three purposes refused	own text	32 (6.2%)	57 (3.2%)	81 (4.8%)	63 (3.7%)	233 (4.1%)
not - training: AI search or user - triggered fetching refused while no training token is	own text	13 (2.5%)	47 (2.7%)	39 (2.3%)	32 (1.9%)	131 (2.3%)
Sites with robots.txt rules		519	1,755	1,699	1,709	5,682
No robots.txt rules (answered)		161	402	550	643	1,756

Policy type	File read	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)	All
Unknown (refused, failed or not collected)		320	843	751	648	2,562

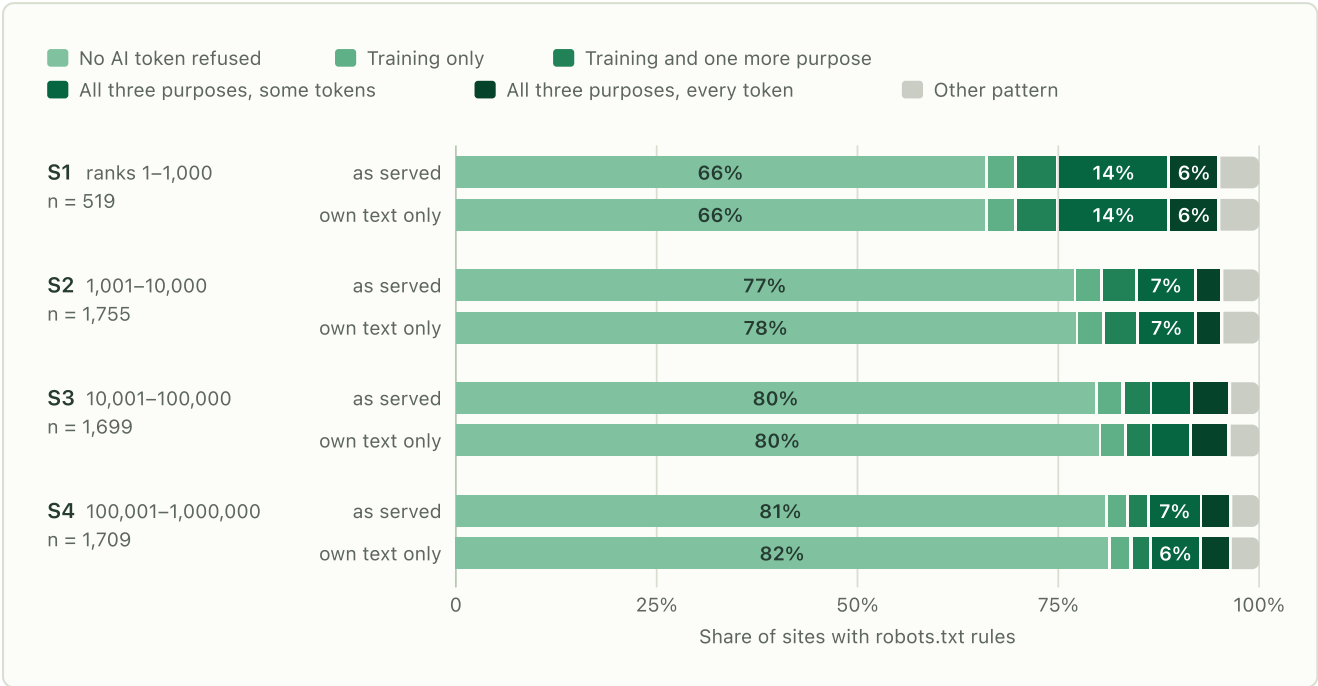


Figure 2. Site-level policy types over three purposes (training, AI search, user-triggered fetching), first-party purpose scheme, effective rule. For each rank stratum (n: sites with robots.txt rules) the upper bar reads the robots.txt as served and the lower bar reads it with intermediary-written blocks removed. Shades run from no refusal to refusal of every listed token; “other pattern” collects sites that refuse only tokens of mixed or undeclared purpose, or refuse search or user-triggered tokens without refusing a training token. Values and the eight underlying types are in Table 5.

Read against RQ1, purpose-selective refusal is a minority of refusal: `training-only` and `training-search` are 2.7% and 2.0% of the top million's sites with rules, against 6.4% for `all-purposes` and 3.8% for blanket files, and selectivity runs one way, training out and search in: `not-training` is 1.9%. Own-text types differ little from as-served types: the largest movement in any cell is in the `open` type, from 78.3% of sites with rules as served to 78.6% in own text (Table 5).

4.4 Who wrote the file

Of the files with rules obtained under the research identity, 83.4% in S1 and 71.3% in S4 match none of the twenty-three fingerprints in the codebook; the class means "not recognised", not "hand-written". Files that carry a platform's default text, by a verified fingerprint, are 6.7% of S1 and 23.5% of S4, 22.8% weighted to the top million (95% CI 21.0–24.6%): the share of files that carry a platform's default text rises steeply down the list. Files that match only unverified fingerprints are 9.1% of S1 and 5.0% of the sample, most of them matches of the per-agent layout fingerprint discussed below rather than of a builder's template; files that contain a published block list or a generator's output, 28 of 5,682 (0.5%). Intermediary-written text of any kind, the comment-only policy file or a managed block, appears in 2.6% of served files, 0.6% in S1 and 4.4% weighted to the top million; the comment-only file alone is 4.2% of S4's served files. Among the sites

that explicitly refuse at least one AI token, every such refusal sits inside an intermediary block on 18 of 942 (1.9%; 0.0% in S1, 3.4% weighted to the top million), none sits inside one on 97.8%, and on 0.3% the refusals sit on both sides of the markers.

Table 6. Who wrote the robots.txt text that the research identity received. A file gets the first class that applies, in the order shown. Counts of sites, with their share of the group named in the second column; fingerprints of platform and generator text are listed in the codebook with their evidence level.

Class	Share of	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)	All
intermediary - comment - only: the policy text an intermediary serves for sites without a robots.txt; no directive	sites that served a file	3 (0.6%)	10 (0.6%)	44 (2.5%)	76 (4.2%)	133 (2.3%)
intermediary - only: every directive sits inside a block written by an intermediary	sites with rules	0 (0.0%)	4 (0.2%)	3 (0.2%)	5 (0.3%)	12 (0.2%)
intermediary - plus - own: a block written by an intermediary, and directives outside it	sites with rules	0 (0.0%)	0 (0.0%)	6 (0.4%)	3 (0.2%)	9 (0.2%)
platform - template: no intermediary block; contains the default text of a content-management system or site builder (verified fingerprint)	sites with rules	35 (6.7%)	195 (11.1%)	294 (17.3%)	401 (23.5%)	925 (16.3%)
generator - template: no intermediary block or platform text; contains a published block list or generator output (verified fingerprint)	sites with rules	4 (0.8%)	15 (0.9%)	5 (0.3%)	4 (0.2%)	28 (0.5%)
unverified - template: matches only template fingerprints that have not been checked against a source or a collected sample	sites with rules	47 (9.1%)	79 (4.5%)	83 (4.9%)	77 (4.5%)	286 (5.0%)
no - known - template: directives, and none of the known templates: written by the site, or by a tool this study does not recognise	sites with rules	433 (83.4%)	1,462 (83.3%)	1,308 (77.0%)	1,219 (71.3%)	4,422 (77.8%)
no - directives: comments or nothing at all, not from a known intermediary	sites that served a file	6 (1.1%)	26 (1.5%)	27 (1.5%)	35 (1.9%)	94 (1.6%)
Any intermediary - written text: the policy - only file, or a block	sites that served a file	3 (0.6%)	14 (0.8%)	53 (3.0%)	84 (4.6%)	154 (2.6%)
Every explicit full disallow of an AI token is intermediary - written	sites that explicitly refuse an AI token	0 (0.0%)	4 (1.2%)	7 (2.7%)	7 (3.6%)	18 (1.9%)

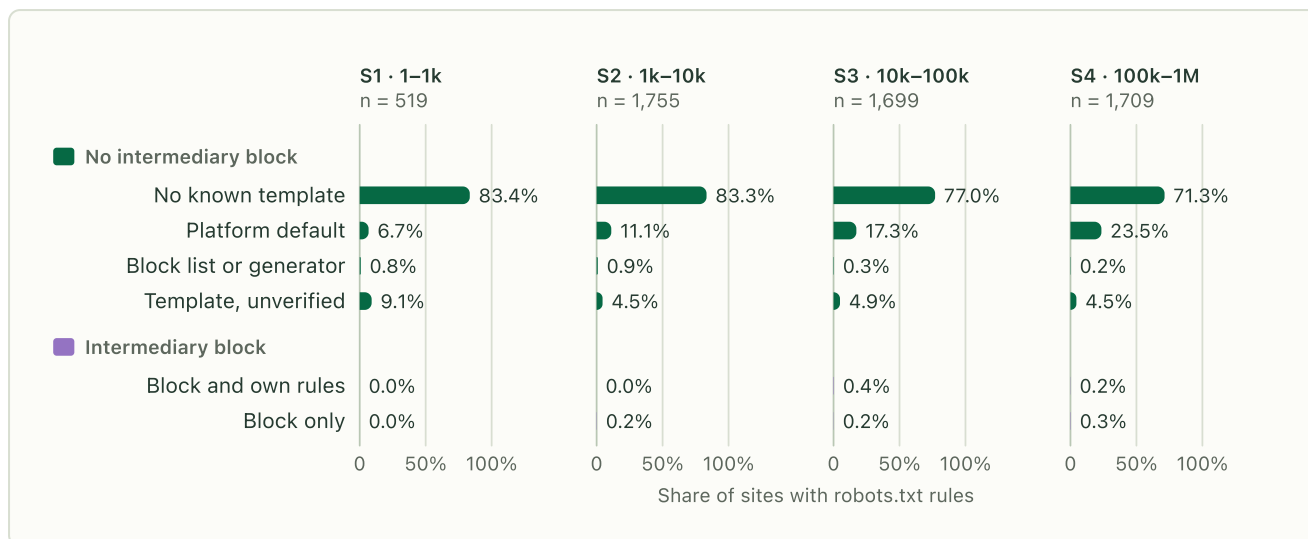


Figure 3. Who wrote the robots.txt text, among sites whose robots.txt with rules was obtained under the research identity (n in each panel), by rank stratum (S1: ranks 1–1,000; S2: 1,001–10,000; S3: 10,001–100,000; S4: 100,001–1,000,000). A file is classed by the first description that applies: it holds a block written by an intermediary, with or without rules of the site's own; it is the default text of a platform; it is a published block list or the output of a generator; it matches a template fingerprint that has not yet been verified against a source; or it matches none of the known templates. The six shares of a panel add up to 100%. Values are in Table 6, which also counts the files without any directive, among them the policy-only text an intermediary serves for sites that have no robots.txt.

The hand check that verified the platform fingerprints also showed their limits (counts from the verification log). "Platform template" means that the file carries a platform's default text: of the 20 files checked for the WordPress fingerprint, 5 were confirmed by both reviewers as core output as served and the other 15 were hand-written, plugin-generated or statically served files that kept the two default rules; of the files that match the Drupal fingerprint, about one in seven belong to other systems that ship the same lines. Four fingerprints for hosted builders were confirmed on all or all but one of the drawn files (19 of 20, 9 of 9, 13 of 13 and 4 of 4) but have no published vendor source for their text, so under the pre-registered rule they stay unverified. The 24, 9 and 13 sites of the first three are in `unverified-template`, and a reader who accepts the hand check alone would move them to `platform-template`; the 4 sites of the fourth also carry WordPress's two default rules, a verified fingerprint, and are in `platform-template` already, whichever way that fourth fingerprint is read. One fingerprint for a blogging platform was a false positive on 17 of 20 drawn files, large sites' hand-written files that happen to carry a group for an advertising crawler and a `Disallow: /search` line; the protocol forbids tightening a pattern after results have been seen, so it stays unverified and counts towards no verified class. One fingerprint meant to recognise a generator's per-agent output recognised no vendor: of 20 matching files none carried the generator's markers, and of 211 distinct matching bodies only 5 were that generator's output, so the generator class here recognises only two generators, a content-signals generator and the `ai.robots.txt` list, and the 238 sites the per-agent fingerprint matched are not called generator users.

Two platform defaults bear on the refusal figures; the verification log counts 22 sites whose default text names an AI token, 9 of one site builder and 13 of another. The first builder's default names 26 AI tokens in the same group as `*`, so it disallows for them only what the site disallows for everyone: on most of those sites the named tokens are restricted on some paths only, which this report does not count as refusal, and they are refused by name only where the site disallows everything for that group. The second builder's default disallows everything for one crawler on the codebook's AI list, with a crawl delay, so all 13

of its sites refuse that token by name. The token-level figures of Section 4.2 include such defaults, because the file is what the site states, whoever wrote it. Their share of named refusal is small: of the sites that refuse an AI token by name, 14 of 942 carry one of these two defaults (an exploratory count), fewer than the 22 of the log because the first builder's default does not by itself fully disallow the tokens it names.

4.5 The invitation layer

Counted strictly, a valid llms.txt is served by 860 of 10,000 sampled domains (8.6%; 8.8% of the top 10,000, 95% CI 8.1–9.5%; 8.7% of the top million). Among sites where the request received an HTTP answer at all, the figure is 12.1%. Status 200 alone would count 17.4% of domains, 2.0× the valid count; the difference is HTML pages, catch-all responses and empty bodies. llms-full.txt is valid on 2.2% of domains, with a status-200 overcount of 5.0×, and Markdown content negotiation on 1.1%, 1.6% of the top 10,000. Beyond those three, the invitation files are rare. The API catalog, one of two registered paths, is valid on 31 of 10,000 domains (0.3% of S1) while 8.7% answered 200, an overcount of 28.1×; the A2A Agent Card is valid on 3 of 10,000 domains against 7.4% answering 200, 247.3×. The MCP server card, mcp.json, the AI Catalog, Agentic Resource Discovery, the Agent Skills index and auth.md are valid on 17 of 10,000, 3 of 10,000, 16 of 10,000, 6 of 10,000, 23 of 10,000 and 8 of 10,000 domains, with overcounts between 32.8× and 250.0×. A valid MCP discovery file at any of the four probed paths, the three candidates and the proposed new name, is on 30 of 10,000 domains, and a valid Agent Card at either path on 4 of 10,000. The TDMRep file is valid on 17 of 10,000 domains and the RSL licence document on 3 of 10,000.

Table 7. The invitation layer: files a site can publish for AI systems and agents. "Status 200" is what a scan that looks only at the status line would count; "present" is the lenient tier, "valid" the strict tier that the paper reports. Counts over all sampled domains; "not requested" are sites whose robots.txt disallowed the request or that were left alone.

File	Group	Status 200	Present	Valid (share of sample)	Valid, top 10,000	Status 200 ÷ valid	Not requested
API catalog (RFC 9727)	registered	871	31	31 (0.3%)	0.6% (0.4–0.8%)	28.1×	2,838
A2A Agent Card	registered	742	3	3 (<0.1%)	<0.1%	247.3×	2,856
llms.txt	proposal	1,736	944	860 (8.6%)	8.8% (8.1–9.5%)	2.0×	2,832
llms-full.txt	proposal	1,090	226	220 (2.2%)	1.3% (1.0–1.6%)	5.0×	2,830
Markdown content negotiation	proposal	—	234	106 (1.1%)	1.6% (1.2–1.9%)		2,791
A2A Agent Card, legacy path	proposal	744	4	1 (<0.1%)	<0.1%	744.0×	2,860
MCP server card	proposal	755	22	17 (0.2%)	0.3% (0.2–0.5%)	44.4×	2,862

File	Group	Status 200	Present	Valid (share of sample)	Valid, top 10,000	Status 200 ÷ valid	Not requested
mcp.json	proposal	750	12	3 (<0.1%)	<0.1%	250.0×	2,858
AI Catalog	proposal	759	16	16 (0.2%)	0.3% (0.2–0.5%)	47.4×	2,858
Agentic Resource Discovery	proposal	746	7	6 (0.1%)	0.1%	124.3×	2,858
Agent Skills index	proposal	755	23	23 (0.2%)	0.4% (0.2–0.5%)	32.8×	2,862
auth.md	proposal	900	12	8 (0.1%)	0.1%	112.5×	2,844
agents.json	proposal	739	1	no strict rule			2,859
ai-plugin.json	proposal	748	10	no strict rule			2,858
ai.txt	proposal	873	15	no strict rule			2,834
TDMRep file	proposal	758	18	17 (0.2%)	0.2% (0.1–0.3%)	44.6×	2,856
RSL licence document	proposal	4	3	3 (<0.1%)	0.1%	1.3×	2,474
OAuth protected-resource metadata	infrastructure	1,111	173	173 (1.7%)	0.6% (0.4–0.8%)	6.4×	2,835
OAuth authorization-server metadata	infrastructure	1,163	224	223 (2.2%)	1.1% (0.9–1.4%)	5.2×	2,834
OpenID configuration	infrastructure	1,026	183	182 (1.8%)	0.8% (0.6–1.0%)	5.6×	2,837
security.txt	infrastructure	1,368	631	610 (6.1%)	10.0% (9.3–10.8%)	2.2×	2,845

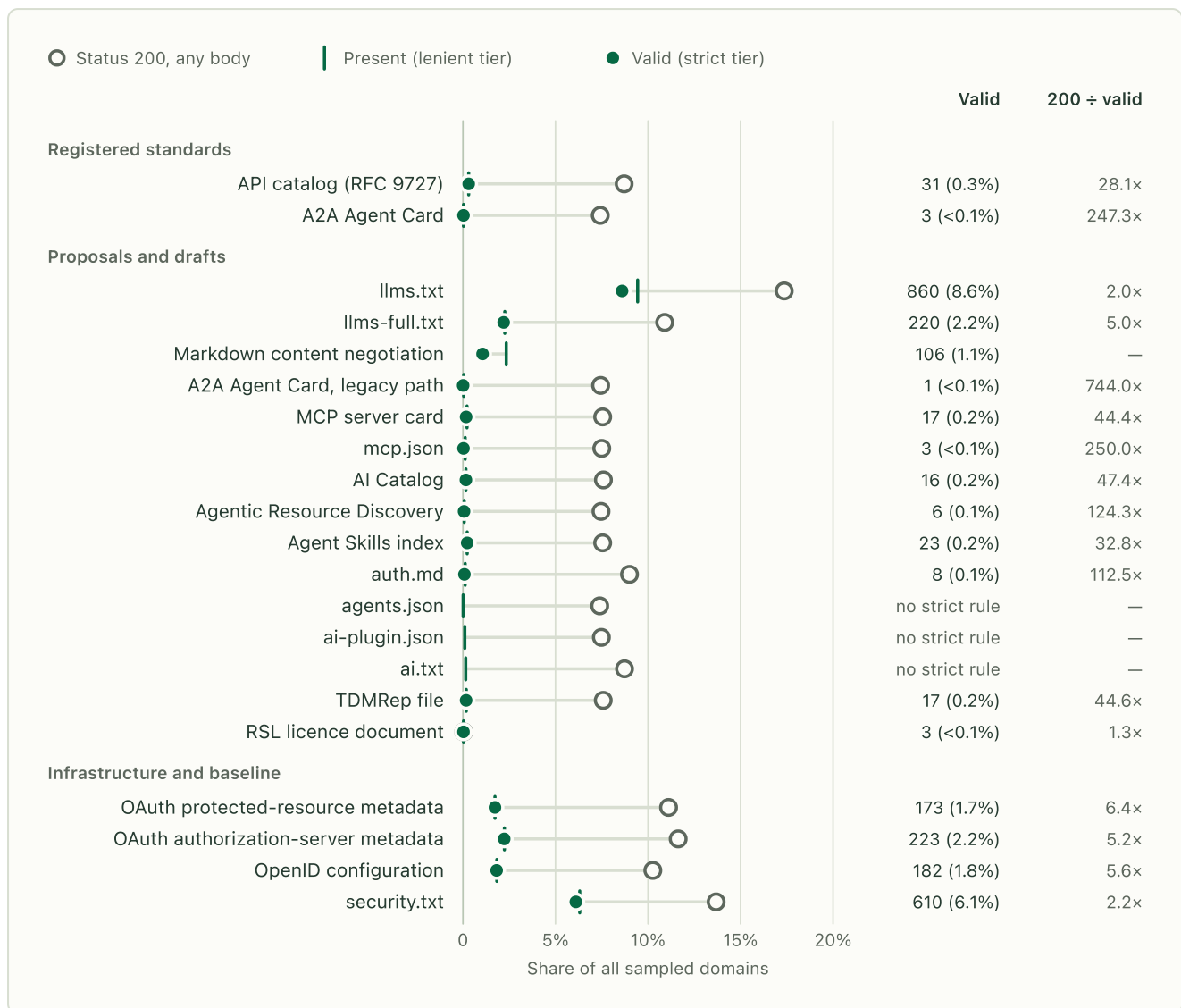


Figure 4. The invitation layer counted three ways, over all sampled domains (unweighted). Open ring: the request ended with status 200, whatever the body, which is what a scan of status lines counts. Vertical tick: the file is present under the lenient rule. Filled dot: the file is valid under the strict rule, the figure this paper reports. The last column is the first count divided by the third. Three files have no verified strict rule; the Markdown request goes to the home page, so a status count is not shown for it. Values are in Table 7.

The overcounts have one cause. A random path at the root answered 200 on 12.0% of sites whose control request was answered, and a random path under `/.well-known/` on 11.9%; a scanner that trusts status lines counts every one of those sites as publishing every file it asks for. The infrastructure group shows the same effect at a larger scale: OAuth protected-resource metadata is valid on 1.7% of domains, authorization-server metadata on 2.2% and OpenID configuration on 1.8%, with overcounts of 6.4×, 5.2× and 5.6×. `security.txt`, the baseline with the longest history, is valid on 6.1% of sampled domains and 17.6% of the top thousand, more than any AI invitation at the top of the list. Three files have no verified strict rule and are shown at the lenient tier only: `agents.json` is present on 1 of 10,000 domains, `ai-plugin.json` on 10 of 10,000 and `ai.txt` on 15 of 10,000.

Signals inside `robots.txt` are young. A `Content-Signal` record appears in 120 of 5,682 files with rules (2.1%; 2.7% in S1), 1.7% outside intermediary blocks; a site-wide `ai-train=no` on 1.4% and `ai-train=yes` on 0.5%; `search=yes` on 1.9% and `search=no` on 0.0%; `ai-input=yes` on 1.2% and `ai-input=no` on 0.2%.

The IETF vocabulary (`Content-Usage`) appears in 6 of 5,682 files, an RSL `License` line in 7 of 5,682, an `Agentmap` record in 4 of 5,682 and the records of the expired Microsoft draft in 3 of 5,682.

Put together, 943 of 10,000 sampled domains have at least one valid invitation (9.4%; 10.9% of S1, 95% CI 9.1–13.0%; 10.1% of the top 10,000; 9.2% of the top million, 95% CI 8.3–10.2%). Almost all of it is `llms.txt` and its companions: a valid file of the proposal group is on 9.4% of domains, a valid registered file on 0.3% (0.7% of S1, 0.1% of the top million), and the infrastructure group, which invites no one, on 8.1% (19.0% of S1). Unlike refusal, invitation has almost no rank gradient: 10.9% of S1 and 9.3% of S4 have one (Table A2 gives every file by stratum).

4.6 Invitation against refusal

Of the 943 of 10,000 sampled domains with a valid invitation, 61 of 943 (6.5%) refuse at least one listed crawler of the audience their invitation addresses under the first-party scheme (7.3% of S1's invited sites, 95% CI 3.2–14.0%; 5.5% weighted to the top million, 95% CI 3.1–7.9%). Under the community scheme, which assigns fewer tokens to the inference audience, the rate is 1.4%. A void invitation, one whose every addressed crawler is refused, occurs on 1 of 943 invited sites. The broad count, a valid invitation beside any fully disallowed AI token, is 12.8% of invited sites (15.6% of S1's). As a share of all sampled domains, same-purpose conflict is 0.6%; as a share of sites with rules, 1.1%.

Table 8. Invitation against refusal. "Same-purpose conflict": a valid invitation, and at least one listed crawler of the kind the invitation addresses is refused. "Void invitation": every such crawler is refused. "Co-occurrence": a valid invitation next to any fully disallowed AI token. Counts of sites, with their share of the sites that have at least one valid invitation.

	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)	All
Sites with at least one valid invitation	109	301	254	279	943
Same-purpose conflict, first-party scheme	8 (7.3%)	22 (7.3%)	16 (6.3%)	15 (5.4%)	61 (6.5%)
Same-purpose conflict, community scheme	4 (3.7%)	5 (1.7%)	3 (1.2%)	1 (0.4%)	13 (1.4%)
Void invitation, first-party scheme	1 (0.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.1%)
Void invitation, community scheme	1 (0.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	1 (0.1%)
Co-occurrence	17 (15.6%)	41 (13.6%)	34 (13.4%)	29 (10.4%)	121 (12.8%)

By artifact, `llms.txt` is in same-purpose conflict on 54 of 860 of the sites where it is valid (6.3%): the file is published and its own path is disallowed for at least one AI-search or user-triggered crawler. `llms-full.txt` is in conflict on 3.6% of its sites and Markdown negotiation on 5.7%. A site-wide `ai-input=yes` sits beside a full disallow of an inference crawler on 11 of 70 of the sites that grant it, and `ai-train=yes` beside a refused training crawler on 1 of 31. None of the agent-facing files is in conflict (the API catalog, 0 of 31; the Agent Skills index, 0 of 23). A site-wide signal that grants search beside a fully disallowed Googlebot or

bingbot occurs on 0 of 10,000 domains. The opposite pattern is common and is not a conflict: among the 120 of 5,682 files with rules that carry a `Content-Signal` record, 40.0% say `ai-train=no` and also refuse some training tokens, and 21.7% say `ai-train=no` and refuse none, a signal with no rule behind it (Table A3).

These rates are a floor: the crawler obeyed robots.txt, so 28.3% of sampled domains were never asked for llms.txt, because a `*` rule or the file's own path was disallowed for unnamed crawlers or because the site was left alone.

4.7 Two identities, one file

Of the 10,000 collected domains, the robots.txt request was answered for both identities on 73.5% and for neither on 23.3%. It was answered for the browser string and not for the research identity on 237 of 10,000 (2.4%; 3.2% of S1, 95% CI 2.3–4.5%; 3.4% of the top 10,000, 95% CI 2.9–3.8%; 1.4% of the top million, 95% CI 1.0–1.7%), and for the research identity only on 84 of 10,000 (0.8%; 1.1% of S1, 0.7% of the top million). The asymmetry holds in every stratum (Table 9) and is largest in S1 and S2. The browser identity obtained a file with rules where the research identity obtained none on 210 of 10,000 collected domains (2.1%), against 0.6% the other way: those are the sites whose policy file this study could read only by pretending to be a browser, and their files are excluded from every refusal figure above.

Table 9. The two robots.txt requests of each site, identical but for the User-Agent header: was the request answered (file served, or definitively absent)? Counts of collected sites, with their share.

	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)	All
Answered for both identities	669 (66.9%)	2,120 (70.7%)	2,233 (74.4%)	2,332 (77.7%)	7,354 (73.5%)
Answered for the browser identity only	32 (3.2%)	101 (3.4%)	66 (2.2%)	38 (1.3%)	237 (2.4%)
Answered for the research identity only	11 (1.1%)	37 (1.2%)	16 (0.5%)	20 (0.7%)	84 (0.8%)
Answered for neither	288 (28.8%)	742 (24.7%)	685 (22.8%)	610 (20.3%)	2,325 (23.3%)
Both received a body, and the bytes differ	130 (16.7%)	423 (16.9%)	381 (14.8%)	347 (13.1%)	1,281 (15.1%)

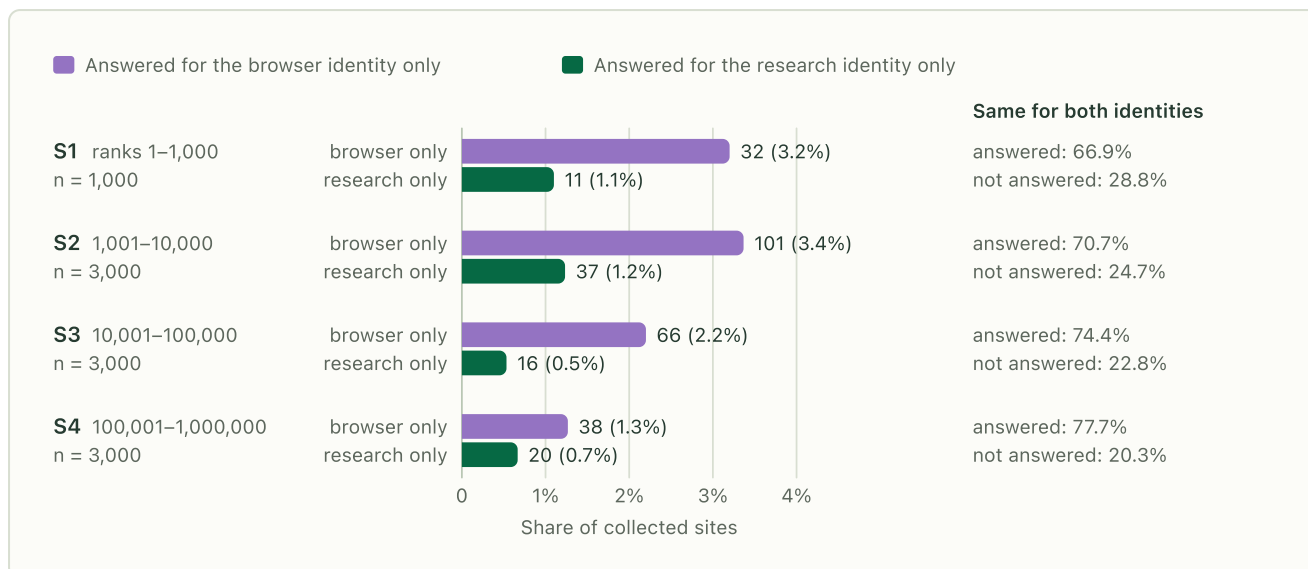


Figure 5. The two robots.txt requests of each site, identical but for the User-Agent header: one declares the research crawler, the other carries a common desktop-browser string. A request is “answered” when the file was served or was definitively absent. The bars show only the sites where the two requests ended differently; the text column gives the sites where they ended alike. Shares are of the collected sites of each stratum (n). Both requests come from one HTTP client without JavaScript, so the comparison isolates the effect of the header alone. Values are in Table 9; the test is C6 in Table 10.

Two checks accompany the pair. Order did not drive it: the research identity was refused on 8.2% of sites where it asked first and 7.0% where it asked second; the browser identity was refused on 5.9% where it asked first and 6.6% where it asked second. Where both identities received a body, the bytes differed on 1,281 of 8,506 (15.1%; 16.7% in S1, 13.1% in S4). Where both identities obtained a file with rules, the two files fully disallow different AI tokens on 0 of 5,621: no site in the sample showed one policy to the browser string and another to the research crawler. The rest of the differing pairs include every case where one identity received a block, challenge or error page; six sites served the same challenge page to both identities with a different token in each and are counted among the differing bodies for that reason.

The verification log shows whom the refusals were aimed at: three sites behind one cloud firewall challenged every path after a few requests, whichever identity sent them, a request-count trigger; six sites behind one hosting platform challenged the opening request of the session, which was the browser-string request, and every later one; two domains on that platform applied the path rule glossed in Section 4.1. None of these was aimed at the research string. The 237 of 10,000 sites that answered the browser string and did not answer the research identity, by a refusal or by a failure, are the ones whose outcome did turn on the string, and they are 2.4% of the sample.

4.8 The six tests and the categories

All six pre-registered tests fall below the Bonferroni threshold of 0.0083. C1 ($p < 0.001$): among the 5,682 sites with rules, 398 refuse GPTBot and admit OAI-SearchBot under the effective rule and 4 do the reverse. C2 ($p < 0.001$): 287 refuse GPTBot and admit ChatGPT-User, 16 the reverse. C3 ($p < 0.001$): the linear trend of explicit training refusal over the four strata has a statistic of 72.85 on one degree of freedom. C4 ($p < 0.001$): news sites refuse a training token by name more often than other sites in every stratum, with a common odds ratio of 9.02 (95% CI 6.85–11.89). C5 ($p < 0.001$): sites under an EU top-level domain refuse one more often than the rest, common odds ratio 2.14 (95% CI 1.74–2.65). C6 (p

<0.001): 237 sites answered the policy request for the browser identity only and 84 for the research identity only, of 10,000 pairs. Every direction matches what the protocol recorded as the expectation from earlier work, except C5, for which none was registered.

Table 10. The 6 pre-specified tests. Each is two-sided and judged at 0.0083 (0.05 divided by 6). No other comparison in the paper is a confirmatory test.

Test	Comparison	Counts	Statistic	p	Below the threshold
C1	Training against AI search, same operator. Sites with robots.txt rules: GPTBot fully disallowed against OAI-SearchBot fully disallowed (effective rule), paired within site	398 first only, 4 second only, of 5,682 pairs	exact	<0.001	yes
C2	Training against user-triggered fetching, same operator. Sites with robots.txt rules: GPTBot fully disallowed against ChatGPT-User fully disallowed (effective rule), paired within site	287 first only, 16 second only, of 5,682 pairs	exact	<0.001	yes
C3	Refusal of training crawlers by rank stratum. Sites with robots.txt rules: at least one training token explicitly fully disallowed, linear trend over strata S1 to S4 (scores 1 to 4)	S1 109/519, S2 259/1,755, S3 200/1,699, S4 137/1,709	$\chi^2 = 72.85$	<0.001	yes
C4	News publishers against all other sites. Sites with robots.txt rules: at least one training token explicitly fully disallowed, news against not news, stratified by rank stratum	S1 31/39 vs 78/480, S2 69/135 vs 190/1620, S3 28/59 vs 172/1640, S4 3/8 vs 134/1701	$\chi^2 = 322.20$; OR 9.02 (6.85–11.89)	<0.001	yes
C5	EU top-level domains against all other sites. Sites with robots.txt rules: at least one training token explicitly fully disallowed, EU proxy against the rest, stratified by rank stratum	S1 15/26 vs 94/493, S2 54/179 vs 205/1576, S3 37/199 vs 163/1500, S4 31/312 vs 106/1397	$\chi^2 = 52.32$; OR 2.14 (1.74–2.65)	<0.001	yes
C6	Declared research identity against browser identity. Collected sites: robots.txt request answered (file served, or definitively absent) for the research identity against the browser identity, paired within site	237 first only, 84 second only, of 10,000 pairs	exact	<0.001	yes

The same comparison for other operators, which is descriptive and not a test, gives the same shape wherever the operator documents a split (Table A5): Anthropic's and Amazon's training tokens are refused alone on hundreds of sites and their search or user-triggered tokens alone on a handful. Mistral AI is the exception: its training and index tokens are refused alone about equally often, and its user-triggered token is refused alone far more often than its training token.

The categories describe where refusal and invitation concentrate. News publishers that have a robots.txt with rules refuse a training token by name on 31 of 39 in S1 (79.5%), 51.1% in S2 and 47.5% in S3, 54.4% pooled, against 10.8% of sites with no category label; they also have a policy file more often than any other group (89.3%) and refuse the research identity less often (4.8%, against 7.4% of unlabelled sites). Government sites have rules on 51.8%, refuse a training token on 3.4% and refuse the research identity on 11.6%, the highest of the three reported categories, though retail, shown for description only, is higher still (20.7%); education sites have rules on 63.9% and refuse a training token on 3.2%. Valid invitations are rarest on government (1.1%) and education sites (4.6%) and commonest, among the reported groups, on

unlabelled sites (9.8%); news sites publish one on 6.3%. Sites under an EU top-level domain, 9.9% of the sample, refuse a training token by name on 57.7% in S1 and 9.9% in S4, 19.1% pooled, against 11.4% for the rest; the gap is widest at the top.

Table 11. Site categories from open sources, by rank stratum. A site can carry several labels. Rates are printed only where a category has at least 30 sampled domains in the stratum; below that only counts are given, and no comparison is made.

Category	Stratum	Sampled	robots.txt rules obtained	Refuses a training token, of sites with rules	robots.txt refused to the research identity	At least one valid invitation
Education	S1	9	9/9	1/9	0/9	0/9
Education	S2	73	76.7% (56/73)	1.8% (1/56)	5.5% (4/73)	1.4% (1/73)
Education	S3	113	60.2% (68/113)	4.4% (3/68)	7.1% (8/113)	8.0% (9/113)
Education	S4	46	45.7% (21/46)	0.0% (0/21)	4.3% (2/46)	2.2% (1/46)
Government	S1	20	11/20	0/11	4/20	0/20
Government	S2	126	48.4% (61/126)	1.6% (1/61)	15.1% (19/126)	0.8% (1/126)
Government	S3	97	50.5% (49/97)	6.1% (3/49)	7.2% (7/97)	2.1% (2/97)
Government	S4	41	63.4% (26/41)	3.8% (1/26)	7.3% (3/41)	0.0% (0/41)
News publishing	S1	42	92.9% (39/42)	79.5% (31/39)	4.8% (2/42)	9.5% (4/42)
News publishing	S2	152	88.8% (135/152)	51.1% (69/135)	5.3% (8/152)	5.9% (9/152)
News publishing	S3	67	88.1% (59/67)	47.5% (28/59)	4.5% (3/67)	6.0% (4/67)
News publishing	S4	9	8/9	3/8	0/9	0/9
Retail and e-commerce	S1	25	22/25	8/22	2/25	4/25
Retail and e-commerce	S2	74	59.5% (44/74)	6.8% (3/44)	27.0% (20/74)	13.5% (10/74)
Retail and e-commerce	S3	33	54.5% (18/33)	5.6% (1/18)	18.2% (6/33)	15.2% (5/33)
Retail and e-commerce	S4	18	12/18	1/12	3/18	2/18
Community and user-generated content	S1	46	84.8% (39/46)	23.1% (9/39)	8.7% (4/46)	8.7% (4/46)
Community and user-generated content	S2	33	81.8% (27/33)	18.5% (5/27)	6.1% (2/33)	9.1% (3/33)

Category	Stratum	Sampled	robots.txt rules obtained	Refuses a training token, of sites with rules	robots.txt refused to the research identity	At least one valid invitation
Community and user-generated content	S3	7	4/7	2/4	1/7	0/7
Community and user-generated content	S4	1	0/1	0/0	0/1	0/1
No category label	S1	862	46.8% (403/862)	15.1% (61/403)	6.6% (57/862)	11.3% (97/862)
No category label	S2	2,548	56.4% (1,438/2,548)	12.5% (180/1,438)	8.4% (215/2,548)	10.9% (277/2,548)
No category label	S3	2,689	56.0% (1,507/2,689)	10.9% (164/1,507)	8.0% (216/2,689)	8.7% (234/2,689)
No category label	S4	2,886	56.9% (1,643/2,886)	8.0% (132/1,643)	6.0% (174/2,886)	9.6% (276/2,886)
EU top-level domain	S1	34	76.5% (26/34)	57.7% (15/26)	5.9% (2/34)	11.8% (4/34)
EU top-level domain	S2	227	78.9% (179/227)	30.2% (54/179)	6.6% (15/227)	7.5% (17/227)
EU top-level domain	S3	288	69.1% (199/288)	18.6% (37/199)	8.3% (24/288)	8.3% (24/288)
EU top-level domain	S4	440	70.9% (312/440)	9.9% (31/312)	5.9% (26/440)	8.0% (35/440)
Other top-level domain	S1	966	51.0% (493/966)	19.1% (94/493)	6.9% (67/966)	10.9% (105/966)
Other top-level domain	S2	2,773	56.8% (1,576/2,773)	13.0% (205/1,576)	9.1% (253/2,773)	10.2% (284/2,773)
Other top-level domain	S3	2,712	55.3% (1,500/2,712)	10.9% (163/1,500)	8.0% (217/2,712)	8.5% (230/2,712)
Other top-level domain	S4	2,560	54.6% (1,397/2,560)	7.6% (106/1,397)	6.1% (156/2,560)	9.5% (244/2,560)

4.9 Beside other trackers

Table A8 places this study's weighted estimate for the top 10,000 beside the figures that Cloudflare Radar published for the same week. The frames differ (a Tranco rank range against Radar's own ranking of domains), the clients differ (a declared research crawler against Cloudflare's scanner), the denominators differ (sites whose file with rules was obtained against files parsed), and Radar's rule counts a file as fully disallowing an agent on the same explicit basis as this study's primary figure. This study's figures for all four tokens are above Radar's, and the intervals do not overlap. Nothing in the comparison can say which difference in frame, client or denominator carries the gap, and Cloudflare, which publishes the comparator, also writes robots.txt text for its customers; the table is given so that the two can be read side by side, not reconciled.

Table A8. Same-month figures published by other trackers, beside this study's estimate for the top 10,000 (explicit full disallow among sites with robots.txt rules). Frames, clients and counting rules differ; the rows are placed side by side, not on one scale.

Source	Data date	Frame	Rule and denominator	GPTBot	CCBot	Google - Extended	ClaudeBot
This study	2026-10-10	Tranco 647LX, ranks 1–10,000 (weighted)	explicit full disallow / robots.txt with rules	11.7% (10.6–12.8%)	13.2% (12.0–14.3%)	9.1% (8.1–10.1%)	10.7% (9.6–11.7%)
Cloudflare Radar (robots.txt, AI user agents)	2026-10-05	Radar top 10,000 domains	fully disallowed / robots.txt files parsed	8.4% (360 of 4,261; 7.7–9.3%)	8.0% (339 of 4,261; 7.2–8.8%)	6.2% (266 of 4,261; 5.6–7.0%)	7.3% (311 of 4,261; 6.6–8.1%)

4.10 Sensitivity

The pre-registered sensitivity analyses move the headline little. Reading each file as Google's parser reads it changes some verdict for some token on 4.0% of sites with rules (8.7% in S1), most of it under the narrower reading that concerns how a token is matched against a `User-agent` value (3.8%) and some under the one that concerns extension lines (1.9%), but the explicit figures for the four headline tokens are unchanged under it. Requiring a stricter full block, every path disallowed and no `Allow` rule in the group, lowers GPTBot's explicit figure from 9.9% to 9.0% of the sample's sites with rules. Using the file of whichever identity obtained rules adds 210 of 10,000 sites to the denominator and leaves GPTBot at 9.8%. The other registered sensitivity analyses (own text, the two purpose schemes, the three tiers, the answered-only denominator, redirected files and soft 404s, request order, the classic-search signal) appear in the sections above.

Table A4. Sensitivity of the explicit full-disallow figures to how a file is read. Share of sites with robots.txt rules, sample pooled. "Stricter full block": every path disallowed and no Allow rule in the group. "Either identity": the file of whichever identity obtained rules.

Token	Primary reading	Google's parser	Stricter full block	Either identity
GPTBot	561 (9.9%)	561 (9.9%)	511 (9.0%)	577 (9.8%)
CCBot	575 (10.1%)	575 (10.1%)	523 (9.2%)	592 (10.0%)
Google - Extended	391 (6.9%)	391 (6.9%)	333 (5.9%)	398 (6.8%)
ClaudeBot	479 (8.4%)	479 (8.4%)	437 (7.7%)	492 (8.4%)

5. Discussion

5.1 What the two voices add up to

On the five questions the data give five short answers. Refusal of AI crawlers by name is a minority stance that grows with rank: 27.0% of the top thousand's sites with rules and 11.9% of the top million's, and it is selective by purpose in one direction only, training out and search in (RQ1). The refusals are the sites' own words: intermediary-written blocks carry every AI refusal on 1.9% of refusing sites, but a platform's default text is in 22.8% of files weighted to the top million, and on 22 sites, by the verification log, a default is what names an AI token (RQ2). The invitation layer is one file and its companions: a valid llms.txt on 8.6% of sampled domains, everything else at or below 2.2%, and any valid invitation on 9.4%, with no rank gradient to speak of (RQ3). Among invited sites, 6.5% refuse a crawler of the audience they invite, and one in eight refuses some AI crawler while inviting another (RQ4). A crawler that says what it is was shown the policy file less often than a browser string, on 2.4% of sites net of 0.8% the other way, and no site in the sample showed the two identities different rules (RQ5).

5.2 For standards bodies

Three findings bear on the standards themselves. The RFC reads a 401 or 403 at `/robots.txt` as "no restrictions". In this sample 7.4% of domains turned a declared crawler's request for the policy file away with such a status, and the standard tells that crawler it may fetch anything. A refusal at the policy file is a statement in its own right, and a common one. Second, the file has no vocabulary for purpose, so purpose lives in token names that eight operators split and the rest do not, and in two signal vocabularies, Cloudflare's `Content-Signal` and the IETF's `Content-Usage`, which are on 2.1% and 0.1% of files with rules and are not translations of each other. Third, a registered path is not a validity rule. The overcounts of Section 4.5, from 2.0× for llms.txt to 744.0× for the legacy Agent Card path, come from sites that answer 200 for any path, 12.0% of those that answered a control request at all. A specification that names a path and a media type but no minimal structural test, or a registry that records paths without one, invites counts that measure catch-all servers. Path churn does the rest: the Agent Card path was renamed in 2025, MCP discovery has three candidate paths and a proposal to rename the newest, and a valid MCP file at any of the four is on 0.3% of domains.

5.3 For AI companies

Sites use the purpose splits that operators offer, with one operator's tokens as the exception (Table A5). A company that wants its search or user-triggered crawler admitted has reason to give it a token of its own and to document what it does; a company that mixes purposes under one token, as Google does for Search and its AI features, leaves sites no choice but the setting it keeps in its own console. Five operators say such a fetcher may not follow robots.txt; ChatGPT-User is nonetheless refused by name on 5.0% of sites with rules. A line that the operator has said it may ignore is still a preference the site has written down, and this report counts it as one. The identity result: a crawler that declares itself, with a product token and a contact link, was refused the policy file more often than a plain browser string. A company's declared crawler may fare differently from this study's; the data cannot say how, only that the policy file is not equally visible to everyone who asks for it.

5.4 For site owners

Three things in the data are within a site's control. The default text of a platform is a statement the site makes; on 22 sites, by the verification log, it names AI tokens that the operator may never have chosen, and on some of them it refuses them outright, and the welcome that an explicit `Allow` gives is rare (2.4% of sites with rules for GPTBot). A `*` group that disallows everything, present on 5.8% of files with rules, refuses every crawler that is not named, including the ones an `llms.txt` on the same site is written for. And a signal without a rule behind it is only a signal: 21.7% of the 120 of 5,682 files with rules that carry a `Content-Signal` record say `ai-train=no` and refuse no training crawler. A site that wants to be read at answer time and not used for training can say so consistently in one file; a site that wants the opposite can check that the path of its invitation is open to the crawler it invites, which 6.3% of `llms.txt` publishers have not.

5.5 Intermediaries, and one intermediary's two roles

Cloudflare appears in this report in three ways: as the author of `robots.txt` text, as a comparator, and as an edge that refuses. As an author its footprint in the files is small. Its comment-only policy file is 4.0% of served files weighted to the top million, its marked blocks appear in 0.2% and 0.2% of files with rules, and every figure computed on own text differs from the as-served figure by a few tenths of a point. That is far below the millions of domains on which the company reported its managed `robots.txt` enabled in September 2025^{3,103}, and the gap has an explanation that the data support rather than prove: a domain with the feature on and no file of its own receives the comment-only text, which states no rule, and a domain whose edge refuses the research identity never shows its file at all. The policy request was refused on 760 of 10,000 sampled domains and recognised as a verified vendor's block or challenge page on 519 of 10,000; most of those carry one of Cloudflare's three signatures. As a comparator it publishes the weekly series that Table A8 sets beside this study's estimate, and its figures are lower than this study's for all four tokens; the company also proposed the `Content-Signal` vocabulary, publishes the largest series on its adoption, and its managed `robots.txt` text carries the customer's choice as a `Content-Signal` line⁹. A site-wide `ai-train=no` is on 1.4% of files with rules as served and 1.0% in own text; the difference is files in which the value sits only inside such a block, and none of this measures what the edge enforces.

5.6 The Google blind spot, and after 3 December 2026

Everything above about Google concerns `Google-Extended`, which is refused by name on 6.9% of the sample's sites with rules and 3.5% weighted to the top million, less than GPTBot though refusing it costs a site nothing in Search. Googlebot itself is welcomed by name on 4.3% of sites and fully disallowed after the fallback on 3.1%. What no site can say in a file is whether its pages may be used in AI Overviews and AI Mode; that choice sits in Search Console¹⁰, and the UK requirement that takes effect on 3 December 2026 orders a comparable control³⁵. A re-measurement after that date, with the same instrument and a fresh draw from a later list, is planned as follow-up work; it is not a condition of this report, which stands as the October 2026 reading.

5.7 Beside the figures that came before

The report's numbers can be placed beside earlier ones only with their differences stated. Liu et al.'s 14 per cent of the top 10,000 concerned pages served to a browser and refused to an AI crawler's string in late 2024 ¹²; this study's 3.4% concerns the policy file and a research identity, a different object and a different string. Bouchaud and Ramaciotti's 25.2 per cent of the top thousand fully blocking GPTBot in September 2025 came from a browsing-population frame and an undeclared client ¹⁴; this study's 15.8% of S1 is a Tranco census read by a declared crawler, and nothing licenses reading the difference as a decline. For llms.txt, the Web Almanac's 2.1 per cent of Chrome UX Report origins in July 2025 ⁶⁴ and the 8.3 per cent of a Tranco top 10,000 in September 2026 ⁶⁶ are of the same order as this study's 8.8% of the top 10,000, which lies above both; agreement on the order of magnitude is the strongest statement the comparison allows. AI Visibility's 10.2 per cent of about 1,900 sites with any file of its own set ¹⁷ and this study's 9.4% with any valid invitation count different sets of files on different frames. None of these pairs is a time series, and the report makes no claim of trend.

6. Limitations

The fourteen limitations registered in the protocol are kept, reordered by how much they bear on the results, with what the collection added to each.

1. **The study measures statements, not enforcement.** A site that blocks AI crawlers in a firewall without a word in robots.txt is counted as open; earlier work found such blocking common ^{12,60}. The one enforcement this study saw was the one aimed at itself (Section 4.7), and an intermediary's edge can enforce a choice that its robots.txt text only states (Section 5.5).
2. **One vantage point, one pass, one day per site.** A residential connection in the United States, IPv4 only, on 2026-10-10. Sites that answer differently by country, network or hour were seen once, and nothing here measures change over time.
3. **The browser identity is a header.** Both requests of the pair came from one HTTP/1.1 client with one TLS fingerprint and no JavaScript, as crawlers and browsers are known to be told apart by more than their string ^{104,105}. A site that blocks on those grounds blocked both, and the pair cannot see it; the asymmetry reported is a lower bound on identity-dependent treatment of the policy file.
4. **A self-identified research crawler may be treated unlike any crawler a site has heard of.** RQ5 measures this for the policy file, and it also conditions RQ3: an invitation file served only to known agents is missed, and one such case, a file served only to certain user-agent strings, is documented elsewhere ¹⁷. The challenges recorded in the verification log (Section 4.7) were not aimed at the research string; the 237 of 10,000 sites that answered only the browser string were.
5. **The list is of domains, not of websites.** No origin could be reached for 13.5% of sampled domains, 20.5% in the top thousand, which is why every rate carries a second denominator and why the all-domains figures are lower than the same rates would be on a list of websites ^{98,99}.

6. **Registrable domains and root paths only.** Policy files on subdomains, path-suffixed OAuth metadata and an llms.txt below the root were not probed.
7. **Conflicts hidden behind a * rule are not observable** by a crawler that obeys robots.txt. 28.3% of domains were never asked for llms.txt, and the conflict rates of Section 4.6 are floors.
8. **Purposes are what operators declare.** A token labelled "search" may feed other uses, twenty-four AI tokens have no usable declaration, and the two schemes disagree on twenty-seven tokens; the training figure moves from 17.9% to 21.5% of sites with rules between them.
9. **Google's AI-search control is invisible from outside** (Section 5.6).
10. **Authorship is recognised only for known text, and a verified fingerprint shows less than its name suggests.** "Platform template" means the file carries a platform's default text, not that the platform wrote it; Section 4.4 gives the hand-check counts. Four hosted builders' templates passed the hand check but have no published vendor source, so under the pre-registered rule they stay unverified: the sites of three of them, 24, 9 and 13 by the log, stay in `unverified-template`, and the fourth's 4 sites are in `platform-template` by a verified WordPress fingerprint they also match; two of those templates name AI tokens by default, so the token-level figures include up to 22 sites whose named refusal is a default, fewer under this report's definition of refusal. One fingerprint was a false positive on 17 of 20 files and counts towards no verified class; one meant to recognise a generator recognised none of its 238 matches as that generator's output, so the generator class covers only two generators; one framework template has no directives and cannot be recognised at all. All counts are from the published verification log.
11. **Edge pages are identified only by verified signatures, and two signature defects were found and not corrected.** Ten signatures are verified and are named by vendor in Section 4.1, with, for the seven raised after the collection, their all-probe matches from the verification log; six are unverified and are reported only as unidentified, inside the challenge and block-page shares and never under a vendor's name. One unverified signature recorded the real robots.txt of 6 sites and the 404 pages of 2 as refused, and stopped 3 sites early; the author chose disclosure over recomputation after the results had been seen, and the upper bound on the error is 8 sites, 0.08 per cent of the sample. One signature confirmed on 18 of 20 files, with 2 that could not be judged, was not raised, so the policy-file refusals among its 55 all-probe matches from 11 domains are counted as challenge pages of an unverified signature, among the unidentified refusals. The CloudFront signature stayed unverified because the two reviewers agreed on 16 of its 20 drawn files, and includes a "Bad request" variant on 3 of the 34 domains it matched; Vercel's deny header records a path rule on 2 domains; both are glossed in Section 4.1.
12. **Categories cover a tenth of the sample and contain errors.** Labels come from Wikidata and domain-name rules, a construction known to be lossy¹⁰⁶; news rests on one source because the open directory planned as a second could not be retrieved; the EU proxy is a top-level domain, not a jurisdiction, and misses an EU publisher under `.com`. Tests C4 and C5 inherit these labels.

13. **"Valid" is a small structural test taken from a specification**, not conformance and not evidence that anything consumes the file.
14. **The specifications are moving**. Several probed paths are drafts or vendor conventions that may be renamed; the codebook records each one's status on 2026-10-04, and a reader in 2027 should expect some of the rows of Table 7 to have changed name.

Two further points concern execution rather than design. The author saw a rehearsal's headline table while collection was running, before the fingerprint checks, and the protocol's rule against verifying fingerprints after seeing results by stratum was therefore met in letter and only partly in spirit; the log records it. And the comparator rows of Table A8 were pulled on the days of the collection and the search of Appendix B was not re-run before publication, so a study published in the week of the collection could have been missed.

7. Ethics

The study had no institutional review. It was designed against the Menlo Report's four principles¹⁰¹ and the norms set for internet-wide measurement, a crawler that names itself, links to an explanation and honours opt-outs¹⁰⁷, and it stops at the lines that recent work on web measurement draws: no credentials, no challenge solved, nothing circumvented¹⁰⁸.

Respect for persons. The subjects are websites, reached through files they publish for machines. Every request carried a `User-Agent` naming the study and linking to a page that states who is asking, the full list of requests, the limits and two ways to opt out; during the run the link led to the author's contact page, because the study page could not be deployed in time, and the study page is published with this report. One request per site carried a browser string instead, disclosed on that page with its reason. The crawler never used another organisation's bot name.

Beneficence. Load was bounded by construction: at most twenty-six `GET` requests per site in one pass, at most one per second per site, bodies cut at one megabyte, one retry, no request after a site refused the crawler or signalled overload. The collector obeyed robots.txt for its own token, and 447 of 10,000 sites disallowed at least one probe through a group it fell under; those probes were not sent. The three sites stopped early by the signature defect of Section 4.1 received fewer requests than planned, not more.

Justice. Sites were selected by rank and a public hash. The text names no site as refusing, inviting or contradicting itself, and the published per-site table holds derived fields only, no file bodies.

Respect for law and public interest. Only `GET` was used; no cookie was stored, no form submitted, nothing written to any site; code, sample and per-site data are published.

To the four questions Partridge and Allman ask of a measurement paper¹⁰²: the requests were few, slow and addressed to public policy files, so collection could not harm a person; the Tranco list, the Public Suffix List and Wikidata are public datasets used under their terms with versions and retrieval times recorded; the collector sought no private information, kept home pages only as extracted signals and a hash, and a `security.txt` contact address stays out of the published table; and the one request that could

be questioned, a browser string from a client that is not a browser, was one request per site for a file meant for crawlers, disclosed in advance, and the only way to learn whether a declared crawler is shown the same policy as everyone else.

What sites did with the crawler's name: no robots.txt in the sample named it (0 of 10,000); no operator asked by message to be left out (0 of 10,000 excluded); and no complaint had been received when this report was prepared.

8. Conclusion

In October 2026 the web's answer to "who may read me?" is written in two places that do not agree with each other and are not read by the same parties. In robots.txt, a minority of popular sites refuse AI crawlers by name, 11.9% of those with rules weighted to the top million and 27.0% of the top thousand, and when they distinguish purposes they keep search and turn training away. The text is theirs: intermediaries wrote almost none of the refusals, though platforms supplied the default lines of a growing share of files down the list. In a newer set of files, a smaller minority invites models and agents in, 9.4% of sampled domains with any valid invitation and almost all of it llms.txt, and one in fifteen of them refuses the crawler it invites. A scan that trusts status lines would count several times as many invitations as exist. And the policy file itself was shown to a browser string and withheld from a crawler that said what it was on 2.4% of sites, so that any count made by a declared crawler, this one included, is a count of what sites are willing to tell such a crawler.

Authorship, status, and materials

Max Song designed the study, wrote the protocol, ran the collection and wrote this report. The collector, the analysis code, the research notes behind Section 2 and the text of this manuscript were prepared with substantial assistance from an AI system (Claude, Anthropic), which drafted code and prose under the author's direction; the author is responsible for every claim. The report is a self-published technical report, version 1.0, dated 2026-10-10. It has not been peer reviewed and had no institutional ethics review; AI assistance is not an independent validation of its claims. No DOI, conference acceptance, institutional affiliation or external endorsement is claimed. The protocol, the codebook, the collector, the sample, the per-site table, the analysis outputs and the verification log are published beside this page.

References

- [1] Martijn Koster; Gary Illyes; Henner Zeller; Lizzi Sassman. 2022. Robots Exclusion Protocol. RFC Editor. Published September 2022; read in full.
- [2] Jeremy Howard. 2026. The /llms.txt file, version 2. Proposal page, llmstxt.org. Published 2024-09-03, revised 2026-08-10. The page does not define llms-full.txt. Retrieved 2026-10-04.
- [3] Will Allen; et al.. 2025. Giving users choice with Cloudflare's new Content Signals Policy. Cloudflare Blog, 24 September 2025. Modified 2026-07-15. Retrieved 2026-10-04.

- [4] Kevin Smith. 2025. api-catalog: A Well-Known URI and Link Relation to Help Discovery of APIs. RFC Editor. Published June 2025; the path is a permanent IANA registration.
- [5] A2A Project. 2026. Agent2Agent (A2A) Protocol Specification, version 1.0. Specification, Linux Foundation project; releases v1.0.0 (2026-03-12) and v1.0.1 (2026-05-28). The Agent Card path was renamed from agent.json in v0.3.0 (2025-07-30). Retrieved 2026-10-04.
- [6] United States District Court for the Southern District of New York. 2025. In re OpenAI, Inc. Copyright Infringement Litigation, No. 25-md-3143: Opinion and Order, Dkt. 968 (Stein, J.). Filed opinion, 15 December 2025, RECAP mirror of the PACER original. Retrieved 2026-10-04.
- [7] European Commission, AI Office. 2025. Code of Practice for General-Purpose AI Models: Copyright Chapter. European Commission. Published 10 July 2025; the Commission page lists the signatories. Retrieved 2026-10-04.
- [8] Jin-Hee Lee. 2026. Say it once: introducing Bot Preference Sync. Cloudflare Blog, 21 August 2026. Retrieved 2026-10-04.
- [9] Bryan Becker. 2026. Have it both ways: stay discoverable in search while disallowing AI training. Cloudflare Blog, 15 September 2026. Retrieved 2026-10-04.
- [10] Mrinalini Loew. 2026. New opportunities, control and insights for website owners. Google Keyword blog, 3 June 2026, with an update dated 31 August 2026; and Search Console Help, Search generative AI control. Retrieved 2026-10-04.
- [11] Victor Le Pochat; Tom Van Goethem; Samaneh Tajalizadehkhoob; Maciej Korczyński; Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019).
- [12] Enze Liu; Elisa Luo; Shawn Shan; Geoffrey M. Voelker; Ben Y. Zhao; Stefan Savage. 2025. Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers. Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25). pp. 78–99.
- [13] Felix Hoffmann; Tobias Goergens; Sonal Khosla; Vaibhav Bajpai. 2026. From robots.txt to ai.txt: Mapping the Evolution of Web Permissions in the Age of AI. ACM SIGCOMM Computer Communication Review. vol. 56(2). pp. 29–41. Only the abstract and the published code and data artefact (github.com/felixhoffmann/robots-ai-permissions) were read for this report.
- [14] Paul Bouchaud; Pedro Ramaciotti. 2025. Web Crawler Restrictions, AI Training Datasets & Political Biases. arXiv:2510.09031, preprint. Version 1, 10 October 2025.
- [15] Cloudflare. 2025. Expanded AI insights in Cloudflare Radar: robots.txt directives for AI user agents. Cloudflare developer changelog, 4 February 2025, and the Radar API endpoint `radar/robots_txt/top/user_agents/directive`. Weekly series; the comparison table uses the scan dated 2026-10-05. Retrieved 2026-10-04.
- [16] André Jesus; Vance Morrison. 2026. Introducing the Agent Readiness score. Is your site agent-ready?. Cloudflare Blog, 17 April 2026, with the Radar agent-readiness API. Cloudflare proposed the Content Signals vocabulary and writes robots.txt text for its customers. Retrieved 2026-10-04.
- [17] Mark McNeece. 2026. AI Discovery File Adoption Research, Q4 2026. AI Visibility. Published 1 October 2026 with a companion post titled One in Ten Top Websites Now Has an AI Discovery File. The vendor defines the file set it measures. Retrieved 2026-10-04.
- [18] Google. 2026. How Google interprets the robots.txt specification. Google Search Central documentation. Page last updated 2026-08-31. Retrieved 2026-10-04.
- [19] Google. 2026. Robots.txt parser and matcher library ([google/robotstxt](https://github.com/google/robotstxt)). Source repository, Apache-2.0 licence. Commit 22b355ff used for the parser comparison. Retrieved 2026-10-04.

- [20] Taein Kim; Karstan Bock; Claire Luo; Amanda Liswood; Chloe Poroslay; Emily Wenger. 2025. Scrapers Selectively Respect robots.txt Directives: Evidence from a Large-Scale Empirical Study. Proceedings of the 2025 ACM Internet Measurement Conference (IMC '25). pp. 541–557.
- [21] TollBit. 2026. State of the Bots, 2026 Q1 & Q2: The Bad Bots. TollBit. A vendor of bot paywalls reporting on its own publisher customers. Retrieved 2026-10-04.
- [22] Yang Sun; Ziming Zhuang; C. Lee Giles. 2007. A Large-Scale Study of Robots.txt. Proceedings of the 16th International Conference on World Wide Web (WWW 2007). pp. 1123–1124.
- [23] Santanu Kolay; Paolo D'Alberto; Ali Dasdan; Arnab Bhattacharjee. 2008. A Larger Scale Study of Robots.txt. Proceedings of the 17th International Conference on World Wide Web (WWW 2008). pp. 1171–1172.
- [24] OpenAI. 2026. Overview of OpenAI Crawlers. OpenAI developer documentation. Undated page. Retrieved 2026-10-04.
- [25] Anthropic. 2026. Does Anthropic crawl data from the web, and how can site owners block the crawler?. Claude Help Center article. Page dated 2026-04-07. Retrieved 2026-10-04.
- [26] Amazon. 2026. About AmazonBot. Amazon Developer documentation. Undated page. Retrieved 2026-10-04.
- [27] Mistral AI. 2026. Mistral crawlers. Mistral AI documentation. Undated page. Retrieved 2026-10-04.
- [28] Perplexity. 2026. Perplexity Crawlers. Perplexity documentation. Undated page. Retrieved 2026-10-04.
- [29] Meta. 2026. Meta Web Crawlers. Meta for Developers documentation. Undated page. Retrieved 2026-10-04.
- [30] Apple. 2026. About Applebot. Apple Support. Published 2026-09-04. Retrieved 2026-10-04.
- [31] Google. 2026. Google's common crawlers. Google Search Central documentation. Page last updated 2026-07-14. Retrieved 2026-10-04.
- [32] Google. 2026. Google User-Triggered Fetchers. Google Search Central documentation. Page last updated 2026-08-19. Retrieved 2026-10-04.
- [33] Common Crawl. 2026. CCBot. Common Crawl website. Undated page. Retrieved 2026-10-04.
- [34] Google. 2025. AI features and your website. Google Search Central documentation. Page last updated 2025-12-10. Retrieved 2026-10-04.
- [35] UK Competition and Markets Authority. 2026. Google Search: Publisher Conduct Requirement, notice under section 19(1) of the Digital Markets, Competition and Consumers Act 2024, and final decision. Competition and Markets Authority. Imposed 3 June 2026; in force from 3 December 2026. Retrieved 2026-10-04.
- [36] ai.robots.txt project. 2026. ai.robots.txt: a list of AI agents and robots to block. Source repository, release v2.0. Release v2.0 dated 2026-10-03, 181 tokens. Retrieved 2026-10-04.
- [37] Known Agents. 2026. Automatic Robots.txt. Known Agents documentation: generated per-agent robots.txt and REST API. Retrieved 2026-10-04.
- [38] Known Agents. 2026. The Agentic Web Index. Known Agents (formerly Dark Visitors), live page. Blocking rule not documented. Retrieved 2026-10-04.

- [39] Cloudflare. 2026. Block AI Bots: Search, Agent and Training presets. Cloudflare developer documentation, Bots. Page last updated 2026-07-01. Retrieved 2026-10-04.
- [40] Internet Assigned Numbers Authority. 2026. Well-Known URIs registry. IANA protocol registry. Registry last updated 2026-09-16, 105 entries. Retrieved 2026-10-04.
- [41] Cloudflare. 2026. Markdown for Agents. Cloudflare developer documentation. Page last updated 2026-07-13. Retrieved 2026-10-04.
- [42] Model Context Protocol. 2026. Model Context Protocol: specification revision 2026-07-28, and the server-card discovery proposal SEP-2127. Specification repository and proposal, under review on 2026-10-04. The specification defines no well-known discovery file; the proposal routes discovery through `/.well-known/ai-catalog.json`. Retrieved 2026-10-04.
- [43] Junjie Bu; R. V. Guha; Shaun Smith. 2026. Agentic Resource Discovery Specification. Version 0.91, status Proposal, 26 August 2026. Retrieved 2026-10-04.
- [44] Cloudflare. 2026. Agent Skills Discovery via Well-Known URIs. Draft specification, version 0.2.0, source repository. Retrieved 2026-10-04.
- [45] Wildcard. 2025. agents.json specification, version 0.1.0. Source repository README. Repository last pushed 2025-08-21. Retrieved 2026-10-04.
- [46] OpenAI. 2026. Plugin development for ChatGPT and Codex. OpenAI developer documentation. The current pages do not describe ai-plugin.json; the earlier plugin documentation redirects here. Retrieved 2026-10-04.
- [47] Spawning. 2023. ai.txt generator. Spawning website. The generator page carries a 2023 footer; the project's main site showed a maintenance page and answered `/ai.txt` with an HTML page on 2026-10-04.
- [48] Cloudflare. 2026. Content Signals. contentsignals.org. Undated page describing the Content-Signal record syntax. Retrieved 2026-10-04.
- [49] Michael Tremante; Leah Romm. 2025. Vocabulary For Expressing Content Signals. Internet-Draft draft-romm-aipref-contentsignals-00, individual submission; dated 2025-10-01; expired 2026-04-04 with no successor revision. Defines the three labels only; the Content-Signal line syntax is not in any IETF document. Retrieved 2026-10-04; re-checked 2026-10-10.
- [50] Paul Keller; Martin Thomson. 2026. A Vocabulary For Expressing AI Usage Preferences. Internet-Draft draft-ietf-aipref-vocab-08, IETF AIPREF working group, work in progress, not an RFC. Revision 08 dated 2026-09-14. Retrieved 2026-10-04.
- [51] Gary Illies; Martin Thomson. 2026. Associating AI Usage Preferences with Content in HTTP. Internet-Draft draft-ietf-aipref-attach-05, IETF AIPREF working group, work in progress, not an RFC. Revision 05 dated 2026-08-19. Retrieved 2026-10-04.
- [52] RSL Technical Steering Committee. 2025. Really Simple Licensing (RSL) 1.0 Specification. RSL-SPEC-1.0, industry specification. Published 2025-12-10. Retrieved 2026-10-04.
- [53] Laurent Le Meur. 2024. TDM Reservation Protocol (TDMRep). W3C Text and Data Mining Reservation Protocol Community Group. Retrieved 2026-10-04.
- [54] Fabrice Canel; Krishna Madhavan. 2025. Robots Exclusion Protocol Extension to communicate AI preferences vocabulary. Internet-Draft draft-canel-robots-ai-control-01, individual submission; latest revision 2025-04-04; expired, datatracker state last updated 2025-10-06. Retrieved 2026-10-04; re-checked 2026-10-10.
- [55] Michael B. Jones; Phil Hunt; Aaron Parecki. 2025. OAuth 2.0 Protected Resource Metadata. RFC Editor. Published April 2025.

- [56] Michael B. Jones; Nat Sakimura; John Bradley. 2018. OAuth 2.0 Authorization Server Metadata. RFC Editor. Published June 2018.
- [57] Edwin Foudil; Yakov Shafranovich. 2022. A File Format to Aid in Security Vulnerability Disclosure. RFC Editor. Published April 2022.
- [58] Thibault Meunier; Sandor Major. 2026. HTTP Message Signatures for automated traffic. Internet-Draft draft-ietf-webbotauth-httpsig-protocol-00, IETF, work in progress, not an RFC. Dated 2026-09-01. Retrieved 2026-10-04.
- [59] Microsoft. 2025. NLWeb: source repository and REST API documentation. GitHub repository nlweb-ai/NLWeb, formerly microsoft/NLWeb. Announced 19 May 2025. Retrieved 2026-10-04.
- [60] Nicolas Steinacker-Olsztyn; Devashish Gosain; Ha Dao. 2026. Is Misinformation More Open? A Study of robots.txt Gatekeeping on the Web. arXiv:2510.10315, version 4; the PDF carries an ACM block for The Web Conference 2026 whose DOI could not be resolved through Crossref on 2026-10-04.
- [61] Shayne Longpre; Robert Mahari; Ariel Lee; Campbell Lund; et al.. 2024. Consent in Crisis: The Rapid Decline of the AI Data Commons. Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track.
- [62] Richard Fletcher. 2024. How Many News Websites Block AI Crawlers?. Reuters Institute for the Study of Journalism.
- [63] Michael Dinzinger; Florian Heß; Michael Granitzer. 2024. A Survey of Web Content Control for Generative AI. arXiv:2404.02309, preprint.
- [64] Amaka Chulwuma; Chris Green; Sophie Brannon. 2026. SEO. The 2025 Web Almanac, chapter 5. Published 15 January 2026, updated 9 June 2026. Retrieved 2026-10-04.
- [65] Christian Liebel; Yash Vekaria; Jonathan Pagel. 2026. Generative AI. The 2025 Web Almanac, chapter 4. Published 15 January 2026. Retrieved 2026-10-04.
- [66] Rankability. 2026. IImS.txt Adoption Research Report. Vendor research page, 18 September 2026. An SEO vendor's scan of Tranco list V3YPN. Retrieved 2026-10-04.
- [67] SEOMator. 2026. IImS.txt: We Scanned the Web's Top 100,000 Domains. Vendor blog, 13 September 2026. An SEO vendor's scan of the Cloudflare Radar top 100,000. Retrieved 2026-10-04.
- [68] Frank Sauerburger. 2026. IImS.txt adoption, one year later. Personal blog, 8 September 2026. Retrieved 2026-10-04.
- [69] Kin Lane. 2026. A2A Adoption Is 421 of 27,840, and Most of Them Arrived With the Card. API Evangelist, 21 September 2026. The earlier post of 29 July 2026, Most Published Agent Cards Are Not Actually A2A, was read as background. Retrieved 2026-10-04.
- [70] Jonathan Gillham. 2024. Block AI Bots from Crawling Websites Using Robots.txt. Originality.ai, 22 August 2024, with a live dashboard. The list and the blocking rule behind the dashboard are not documented. Retrieved 2026-10-04.
- [71] Ben Welsh. 2026. Who blocks OpenAI, Google AI and Common Crawl?. News Homepages project documentation, live page. Live page; its footer read Last updated October 4, 2026. Retrieved 2026-10-04.
- [72] United States District Court for the Southern District of New York. 2026. Reddit, Inc. v. SerpApi LLC et al., No. 25 Civ. 8736: Opinion and Order, Dkt. 104 (Engelmayer, J.). Filed opinion, 31 July 2026, RECAP mirror of the PACER original. Retrieved 2026-10-04.

- [73]** United States Court of Appeals for the Ninth Circuit. 2026. *Amazon.com Services, LLC v. Perplexity AI, Inc.*, No. 26-1444: Opinion (for publication). Filed opinion, 4 August 2026; rehearing en banc denied 10 September 2026. Retrieved 2026-10-04.
- [74]** United States District Court for the Northern District of California. 2025. *Bartz v. Anthropic PBC*, No. C 24-05417 WHA: Order on Fair Use, Dkt. 231 (Alsup, J.). Filed order, 23 June 2025, RECAP mirror of the PACER original. Retrieved 2026-10-04.
- [75]** United States District Court for the Northern District of California. 2025. *Kadrey v. Meta Platforms, Inc.*, No. 23-cv-03417-VC: Order, Dkt. 598 (Chhabria, J.). Filed order, 25 June 2025, RECAP mirror of the PACER original. Retrieved 2026-10-04.
- [76]** United States Court of Appeals for the Third Circuit. 2026. *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, No. 25-2153: Opinion (precedential). Filed opinion, 29 September 2026, unsealed 30 September 2026. Retrieved 2026-10-04.
- [77]** United States Copyright Office. 2025. *Copyright and Artificial Intelligence, Part 3: Generative AI Training (Pre-Publication Version)*. United States Copyright Office. Report of the Register of Copyrights, May 2025. Retrieved 2026-10-04.
- [78]** European Parliament and Council of the European Union. 2019. Directive (EU) 2019/790 of 17 April 2019 on copyright and related rights in the Digital Single Market, Article 4. Official Journal of the European Union, L 130; text read in the official UK reproduction. Retrieved 2026-10-04.
- [79]** European Union Intellectual Property Office. 2025. *The Development of Generative Artificial Intelligence from a Copyright Perspective*. European Union Intellectual Property Office. May 2025, ISBN 978-92-9156-369-2, DOI 10.2814/3893780. Retrieved 2026-10-04.
- [80]** European Commission. 2025. *Consultation on protocols for reserving rights from text and data mining under the AI Act and the GPAI Code of Practice*. Consultation page, Shaping Europe's digital future. Consultation open 1 December 2025 to 23 January 2026; page last updated 16 December 2025. Retrieved 2026-10-04.
- [81]** UK Department for Science, Innovation and Technology and Intellectual Property Office. 2026. *Report on Copyright and Artificial Intelligence*. UK Government. Policy paper laid under the Data (Use and Access) Act 2025, 18 March 2026. Retrieved 2026-10-04.
- [82]** Athena Chapekis; Anna Lieb. 2025. *Google users are less likely to click on links when an AI summary appears in the results*. Pew Research Center short read, 22 July 2025. Retrieved 2026-10-04.
- [83]** Jim Egan. 2026. *Overview and key findings of the 2026 Digital News Report*. Reuters Institute for the Study of Journalism, University of Oxford. Published 16 June 2026. Retrieved 2026-10-04.
- [84]** Jason Kint. 2025. *Facts: Google's push to AI hurts publisher traffic*. Digital Content Next, 14 August 2025. A publishers' trade body. Retrieved 2026-10-04.
- [85]** Liz Reid. 2025. *AI in Search is driving more queries and higher quality clicks*. Google Keyword blog, 6 August 2025. Retrieved 2026-10-04.
- [86]** João Tomé. 2025. *The crawl-to-click gap: Cloudflare data on AI bots, training, and referrals*. Cloudflare Blog, 29 August 2025. Retrieved 2026-10-04.
- [87]** Arielle Weiss; Zach Albertson; Emily Lanfear. 2026. *Content Independence Day, one year on: building the business model for the agentic Internet*. Cloudflare Blog, 1 July 2026. Retrieved 2026-10-04.
- [88]** The Associated Press. 2023. *AP and OpenAI agree to share select news content and technology in new collaboration*. Press release, 13 July 2023. Retrieved 2026-10-04.
- [89]** Reddit. 2024. *Expanding our Partnership with Google*. Reddit Upvoted blog, 22 February 2024. Retrieved 2026-10-04.

- [90] Tim Frank. 2026. Building Toward a Sustainable Content Economy for the Agentic Web. Microsoft Advertising Blog, 3 February 2026. Retrieved 2026-10-04.
- [91] Birgit Mueller; Chris Danis; Giuseppe Lavagetto. 2025. How crawlers impact the operations of the Wikimedia projects. Diff, Wikimedia Foundation, 1 April 2025. Retrieved 2026-10-04.
- [92] Eric Holscher. 2024. AI crawlers need to be more respectful. Read the Docs blog, 25 July 2024. Retrieved 2026-10-04.
- [93] Xe Iaso. 2025. Block AI scrapers with Anubis. Personal blog, 19 January 2025. Retrieved 2026-10-04.
- [94] Gabriel Corral; Vaibhav Singhal; Brian Mitchell; Reid Tatoris. 2025. Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives. Cloudflare Blog, 4 August 2025. An allegation by one party; the named company's response was not read for this report. Retrieved 2026-10-04.
- [95] Stavan Parikh; Rao Surapaneni. 2025. Powering AI commerce with the new Agent Payments Protocol (AP2). Google Cloud Blog, 16 September 2025. Retrieved 2026-10-04.
- [96] Jeff Weinstein; Steve Kaliski. 2025. Developing an open standard for agentic commerce. Stripe Blog, 29 September 2025. Retrieved 2026-10-04.
- [97] Adobe. 2025. Adobe to Acquire Semrush. Press release, 19 November 2025. Retrieved 2026-10-04.
- [98] Quirin Scheitle; et al.. 2018. A Long Way to the Top: Significance, Structure, and Stability of Internet Top Lists. Proceedings of the Internet Measurement Conference 2018 (IMC '18). pp. 478–493.
- [99] Kimberly Ruth; Deepak Kumar; Brandon Wang; Luke Valenta; Zakir Durumeric. 2022. Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists. Proceedings of the 22nd ACM Internet Measurement Conference (IMC '22). pp. 374–387.
- [100] Lawrence D. Brown; T. Tony Cai; Anirban DasGupta. 2001. Interval Estimation for a Binomial Proportion. Statistical Science. vol. 16(2). pp. 101–133.
- [101] David Dittrich; Erin Kenneally. 2012. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. U.S. Department of Homeland Security. August 2012.
- [102] Craig Partridge; Mark Allman. 2016. Ethical considerations in network measurement papers. Communications of the ACM. vol. 59(10). pp. 58–64.
- [103] Cloudflare. 2026. Managed robots.txt (robots.txt setting). Cloudflare developer documentation, Bots. Page last updated 2026-08-03. Retrieved 2026-10-04.
- [104] Syed Suleman Ahmad; Muhammad Daniyal Dar; Muhammad Fareed Zaffar; Narseo Vallina-Rodriguez; Rishab Nithyanand. 2020. Apophanies or Epiphanies? How Crawlers Impact Our Understanding of the Web. Proceedings of The Web Conference 2020 (WWW '20). pp. 271–280.
- [105] Jordan Jueckstock; et al.. 2021. Towards Realistic and Reproducible Web Crawl Measurements. Proceedings of the Web Conference 2021 (WWW '21). pp. 80–91.
- [106] Pelayo Vallina; Victor Le Pochat; Álvaro Feal; et al.. 2020. Mis-shapes, Mistakes, Misfits: An Analysis of Domain Classification Services. Proceedings of the ACM Internet Measurement Conference (IMC '20). pp. 598–618.
- [107] Zakir Durumeric; Eric Wustrow; J. Alex Halderman. 2013. ZMap: Fast Internet-Wide Scanning and Its Security Applications. Proceedings of the 22nd USENIX Security Symposium.

[108] Florian Hantke; Sebastian Roth; Rafael Mrowczynski; Christine Utz; Ben Stock. 2024. Where Are the Red Lines? Towards Ethical Server-Side Scans in Security and Privacy Research. Proceedings of the 2024 IEEE Symposium on Security and Privacy (S&P 2024). pp. 4405–4423.

[109] DeviantArt. 2022. UPDATE All Deviations Are Opted Out of AI Datasets. DeviantArt team journal, 11 November 2022. Defines the noai and noimageai meta and header values. Retrieved 2026-10-04.

Appendix

Appendix A. Supplementary tables

Table A1 gives, for every one of the fifty-four tokens of the codebook, how often it is named, explicitly refused, explicitly welcomed and effectively refused among the sample's sites with rules, with and without intermediary-written blocks.

Table A1. All crawler tokens of the codebook, among sites with robots.txt rules, sample pooled without weights. "Named": some group names the token. "Own text": after intermediary-written blocks are removed.

Token	First-party purpose	Community purpose	Named	Explicit full disallow	Explicit, own text	Explicit allow	Effective full disallow	Effective, own text
GPTBot	training	training	872 (15.3%)	561 (9.9%)	542 (9.5%)	137 (2.4%)	859 (15.1%)	840 (14.8%)
OAI-SearchBot	search-index	search-index	447 (7.9%)	157 (2.8%)	157 (2.8%)	130 (2.3%)	465 (8.2%)	465 (8.2%)
ChatGPT-User	user-fetch	user-fetch	570 (10.0%)	286 (5.0%)	278 (4.9%)	126 (2.2%)	588 (10.3%)	580 (10.2%)
OAI-AdsBot	undeclared	—	23 (0.4%)	7 (0.1%)	7 (0.1%)	11 (0.2%)	339 (6.0%)	340 (6.0%)
ClaudeBot	training	training	752 (13.2%)	479 (8.4%)	460 (8.1%)	120 (2.1%)	784 (13.8%)	765 (13.5%)
Claude-SearchBot	search-index	search-index	262 (4.6%)	131 (2.3%)	131 (2.3%)	46 (0.8%)	448 (7.9%)	449 (7.9%)
Claude-User	user-fetch	user-fetch	269 (4.7%)	140 (2.5%)	131 (2.3%)	50 (0.9%)	456 (8.0%)	448 (7.9%)
anthropic-ai	undeclared	training	491 (8.6%)	352 (6.2%)	334 (5.9%)	48 (0.8%)	670 (11.8%)	652 (11.5%)
Claude-Web	undeclared	training	375 (6.6%)	290 (5.1%)	272 (4.8%)	30 (0.5%)	608 (10.7%)	590 (10.4%)
Googlebot	search-index	search-index	593 (10.4%)	10 (0.2%)	10 (0.2%)	245 (4.3%)	174 (3.1%)	175 (3.1%)
Google-Extended	mixed	training	646 (11.4%)	391 (6.9%)	372 (6.5%)	110 (1.9%)	687 (12.1%)	668 (11.8%)

Token	First-party purpose	Community purpose	Named	Explicit full disallow	Explicit, own text	Explicit allow	Effective full disallow	Effective, own text
GoogleOther	undecleared	training	168 (3.0%)	112 (2.0%)	94 (1.7%)	16 (0.3%)	425 (7.5%)	407 (7.2%)
GoogleOther - Image	undecleared	—	51 (0.9%)	26 (0.5%)	26 (0.5%)	1 (<0.1%)	355 (6.2%)	356 (6.3%)
GoogleOther - Video	undecleared	—	50 (0.9%)	26 (0.5%)	26 (0.5%)	1 (<0.1%)	355 (6.2%)	356 (6.3%)
Google - CloudVertexBot	undecleared	training	169 (3.0%)	141 (2.5%)	122 (2.1%)	12 (0.2%)	463 (8.1%)	445 (7.8%)
Google - Agent	user - fetch	—	33 (0.6%)	23 (0.4%)	14 (0.2%)	0 (0.0%)	352 (6.2%)	344 (6.1%)
Google - GeminiNotebook	user - fetch	—	10 (0.2%)	9 (0.2%)	0 (0.0%)	0 (0.0%)	341 (6.0%)	333 (5.9%)
Google - NotebookLM	user - fetch	—	71 (1.2%)	49 (0.9%)	40 (0.7%)	11 (0.2%)	370 (6.5%)	362 (6.4%)
bingbot	search - index	search - index	470 (8.3%)	36 (0.6%)	36 (0.6%)	213 (3.7%)	221 (3.9%)	222 (3.9%)
Applebot	mixed	search - index	298 (5.2%)	61 (1.1%)	61 (1.1%)	128 (2.3%)	280 (4.9%)	281 (4.9%)
Applebot - Extended	training	training	485 (8.5%)	353 (6.2%)	333 (5.9%)	44 (0.8%)	663 (11.7%)	643 (11.3%)
meta - externalagent	mixed	training	509 (9.0%)	371 (6.5%)	351 (6.2%)	47 (0.8%)	680 (12.0%)	660 (11.6%)
meta - webindexer	search - index	—	129 (2.3%)	98 (1.7%)	98 (1.7%)	6 (0.1%)	425 (7.5%)	426 (7.5%)
meta - externalfetcher	user - fetch	user - fetch	213 (3.7%)	154 (2.7%)	146 (2.6%)	13 (0.2%)	471 (8.3%)	463 (8.1%)
meta - externalads	undecleared	—	37 (0.7%)	12 (0.2%)	12 (0.2%)	9 (0.2%)	336 (5.9%)	337 (5.9%)
facebookexternalhit	undecleared	—	259 (4.6%)	22 (0.4%)	22 (0.4%)	118 (2.1%)	297 (5.2%)	298 (5.2%)
FacebookBot	undecleared	training	322 (5.7%)	251 (4.4%)	251 (4.4%)	24 (0.4%)	571 (10.0%)	571 (10.0%)
Amazonbot	training	training	539 (9.5%)	396 (7.0%)	376 (6.6%)	53 (0.9%)	704 (12.4%)	684 (12.0%)
Amzn - SearchBot	search - index	—	76 (1.3%)	61 (1.1%)	61 (1.1%)	4 (0.1%)	391 (6.9%)	392 (6.9%)
Amzn - User	user - fetch	—	67 (1.2%)	57 (1.0%)	48 (0.8%)	3 (0.1%)	386 (6.8%)	378 (6.7%)

Token	First-party purpose	Community purpose	Named	Explicit full disallow	Explicit, own text	Explicit allow	Effective full disallow	Effective, own text
PerplexityBot	search-index	search-index	604 (10.6%)	306 (5.4%)	306 (5.4%)	135 (2.4%)	608 (10.7%)	608 (10.7%)
Perplexity-User	user-fetch	user-fetch	261 (4.6%)	144 (2.5%)	136 (2.4%)	49 (0.9%)	470 (8.3%)	462 (8.1%)
CCBot	undeclared	training	796 (14.0%)	575 (10.1%)	556 (9.8%)	123 (2.2%)	816 (14.4%)	797 (14.0%)
DuckAssistBot	search-index	user-fetch	231 (4.1%)	150 (2.6%)	150 (2.6%)	24 (0.4%)	470 (8.3%)	470 (8.3%)
MistralAI-Training	training	training	25 (0.4%)	25 (0.4%)	6 (0.1%)	0 (0.0%)	356 (6.3%)	338 (5.9%)
MistralAI-Index	search-index	—	23 (0.4%)	19 (0.3%)	19 (0.3%)	0 (0.0%)	350 (6.2%)	351 (6.2%)
MistralAI-User	user-fetch	user-fetch	165 (2.9%)	117 (2.1%)	108 (1.9%)	19 (0.3%)	445 (7.8%)	437 (7.7%)
AI2Bot	training	—	187 (3.3%)	155 (2.7%)	155 (2.7%)	2 (<0.1%)	485 (8.5%)	486 (8.6%)
ImagesiftBot	search-index	—	262 (4.6%)	245 (4.3%)	245 (4.3%)	1 (<0.1%)	566 (10.0%)	566 (10.0%)
webzio	undeclared	—	28 (0.5%)	26 (0.5%)	26 (0.5%)	0 (0.0%)	358 (6.3%)	359 (6.3%)
webzio-extended	training	—	169 (3.0%)	152 (2.7%)	152 (2.7%)	0 (0.0%)	483 (8.5%)	484 (8.5%)
omgili	undeclared	training	353 (6.2%)	324 (5.7%)	305 (5.4%)	0 (0.0%)	644 (11.3%)	626 (11.0%)
omgilibot	undeclared	—	303 (5.3%)	273 (4.8%)	273 (4.8%)	1 (<0.1%)	594 (10.5%)	594 (10.5%)
YouBot	search-index	—	331 (5.8%)	257 (4.5%)	257 (4.5%)	24 (0.4%)	574 (10.1%)	574 (10.1%)
PetalBot	search-index	training	454 (8.0%)	345 (6.1%)	327 (5.8%)	80 (1.4%)	596 (10.5%)	579 (10.2%)
Diffbot	search-index	training	340 (6.0%)	298 (5.2%)	280 (4.9%)	10 (0.2%)	622 (10.9%)	604 (10.6%)
Diffbot-User	user-fetch	—	16 (0.3%)	16 (0.3%)	16 (0.3%)	0 (0.0%)	348 (6.1%)	349 (6.1%)
Coherebot	undeclared	—	9 (0.2%)	8 (0.1%)	8 (0.1%)	0 (0.0%)	340 (6.0%)	341 (6.0%)
cohere-ai	undeclared	training	416 (7.3%)	328 (5.8%)	310 (5.5%)	22 (0.4%)	646 (11.4%)	628 (11.1%)

Token	First-party purpose	Community purpose	Named	Explicit full disallow	Explicit, own text	Explicit allow	Effective full disallow	Effective, own text
cohere-training-data-crawler	undisclosed	—	148 (2.6%)	126 (2.2%)	126 (2.2%)	2 (<0.1%)	458 (8.1%)	459 (8.1%)
Bytespider	undisclosed	training	703 (12.4%)	542 (9.5%)	524 (9.2%)	96 (1.7%)	784 (13.8%)	766 (13.5%)
DeepSeekBot	undisclosed	—	105 (1.8%)	88 (1.5%)	88 (1.5%)	5 (0.1%)	414 (7.3%)	415 (7.3%)
xAI-SearchBot	undisclosed	—	5 (0.1%)	4 (0.1%)	4 (0.1%)	0 (0.0%)	336 (5.9%)	337 (5.9%)
Timpibot	undisclosed	—	227 (4.0%)	207 (3.6%)	207 (3.6%)	1 (<0.1%)	534 (9.4%)	534 (9.4%)

Table A2 gives every valid invitation file and signal by rank stratum, as a share of all sampled domains with its interval.

Table A2. Valid invitation files and signals by rank stratum: share of all sampled domains, with 95% intervals.

File or signal	S1	S2	S3	S4
API catalog (RFC 9727)	0.7% (0.3–1.4%)	0.6% (0.4–0.9%)	0.1% (0.0–0.3%)	0.1% (0.0–0.2%)
A2A Agent Card	0.1% (0.0–0.6%)	<0.1% (0.0–0.2%)	0.0% (0.0–0.1%)	<0.1% (0.0–0.2%)
llms.txt	9.9% (8.2–11.9%)	8.7% (7.7–9.7%)	7.9% (7.0–9.0%)	8.8% (7.8–9.8%)
llms-full.txt	1.5% (0.9–2.5%)	1.3% (1.0–1.8%)	1.9% (1.5–2.5%)	3.6% (3.0–4.4%)
Markdown content negotiation	1.4% (0.8–2.3%)	1.6% (1.2–2.1%)	0.8% (0.6–1.2%)	0.6% (0.4–1.0%)
A2A Agent Card, legacy path	0.1% (0.0–0.6%)	0.0% (0.0–0.1%)	0.0% (0.0–0.1%)	0.0% (0.0–0.1%)
MCP server card	0.6% (0.2–1.3%)	0.3% (0.1–0.6%)	<0.1% (0.0–0.2%)	<0.1% (0.0–0.2%)
mcp.json	0.2% (0.0–0.7%)	0.0% (0.0–0.1%)	0.0% (0.0–0.1%)	<0.1% (0.0–0.2%)
AI Catalog	0.2% (0.0–0.7%)	0.3% (0.2–0.6%)	0.1% (0.0–0.3%)	<0.1% (0.0–0.2%)
Agentic Resource Discovery	0.0% (0.0–0.4%)	0.2% (0.1–0.4%)	<0.1% (0.0–0.2%)	0.0% (0.0–0.1%)
Agent Skills index	0.6% (0.2–1.3%)	0.4% (0.2–0.7%)	0.2% (0.1–0.4%)	<0.1% (0.0–0.2%)
auth.md	0.3% (0.1–0.9%)	0.1% (0.0–0.3%)	<0.1% (0.0–0.2%)	0.0% (0.0–0.1%)
TDMRep file	0.5% (0.2–1.2%)	0.2% (0.1–0.4%)	0.2% (0.1–0.4%)	<0.1% (0.0–0.2%)
RSL licence document	0.2% (0.0–0.7%)	<0.1% (0.0–0.2%)	0.0% (0.0–0.1%)	0.0% (0.0–0.1%)
OAuth protected-resource metadata	0.7% (0.3–1.4%)	0.6% (0.4–0.9%)	1.3% (1.0–1.8%)	3.6% (3.0–4.4%)

File or signal	S1	S2	S3	S4
OAuth authorization - server metadata	2.8% (1.9–4.0%)	0.9% (0.6–1.3%)	1.7% (1.3–2.3%)	3.8% (3.2–4.6%)
OpenID configuration	2.3% (1.5–3.4%)	0.6% (0.4–1.0%)	1.1% (0.8–1.6%)	3.5% (2.9–4.3%)
security.txt	17.6% (15.4–20.1%)	9.2% (8.2–10.3%)	3.8% (3.1–4.5%)	1.5% (1.2–2.0%)
agents.json (present; no strict rule)	0.1% (0.0–0.6%)	0.0% (0.0–0.1%)	0.0% (0.0–0.1%)	0.0% (0.0–0.1%)
ai-plugin.json (present; no strict rule)	0.3% (0.1–0.9%)	0.1% (0.0–0.3%)	0.1% (0.0–0.2%)	<0.1% (0.0–0.2%)
ai.txt (present; no strict rule)	0.0% (0.0–0.4%)	0.2% (0.1–0.4%)	0.1% (0.0–0.3%)	0.2% (0.1–0.4%)
Content - Signal record in robots.txt	1.4% (0.8–2.3%)	1.9% (1.5–2.5%)	0.9% (0.6–1.3%)	0.7% (0.5–1.1%)
Content - Signal record, own text only	1.4% (0.8–2.3%)	1.8% (1.4–2.3%)	0.6% (0.4–0.9%)	0.5% (0.3–0.8%)
Content - Usage record in robots.txt	0.1% (0.0–0.6%)	0.2% (0.1–0.4%)	0.0% (0.0–0.1%)	0.0% (0.0–0.1%)
License record in robots.txt (RSL)	0.4% (0.1–1.0%)	0.1% (0.0–0.2%)	<0.1% (0.0–0.2%)	0.0% (0.0–0.1%)
Any valid invitation	10.9% (9.1–13.0%)	10.0% (9.0–11.2%)	8.5% (7.5–9.5%)	9.3% (8.3–10.4%)

Table A3 sets the site-wide `ai-train` value of the 120 of 5,682 files that carry a `Content-Signal` record against what the same file does to training crawlers.

Table A3. Content-Signal ai-train against what the same robots.txt does to training crawlers (first-party scheme, effective rule). Counts of sites whose robots.txt carries a Content-Signal record, sample pooled.

Site - wide ai - train	No training token refused	Some refused	All refused
yes	30	1	0
no	26	48	3
unstated	10	2	0

Table A5 repeats the comparison of tests C1 and C2 for the other operators that document a training token beside a search or user-triggered token; only the OpenAI rows are tests, and the other p-values are unadjusted and descriptive.

Table A5. Training token against the search and user-triggered tokens of the same operator, among sites with robots.txt rules (effective rule). Only the OpenAI rows are pre-specified tests; the others are descriptive, with unadjusted exact McNemar p-values.

Operator	Comparison	Only the training token refused	Only the other token refused	Both refused	p
OpenAI	GPTBot against OAI - SearchBot	398	4	461	<0.001
OpenAI	GPTBot against ChatGPT - User	287	16	572	<0.001
Anthropic	ClaudeBot against Claude - SearchBot	339	3	445	<0.001
Anthropic	ClaudeBot against Claude - User	333	5	451	<0.001
Mistral AI	MistralAI - Training against MistralAI - Index	24	18	332	0.441
Mistral AI	MistralAI - Training against MistralAI - User	18	107	338	<0.001
Amazon	Amazonbot against Amzn - SearchBot	324	11	380	<0.001
Amazon	Amazonbot against Amzn - User	328	10	376	<0.001

Table A6 counts the signals found on the one home page fetched per site; these home-page signals are exploratory, registered as such in the protocol, and none of them is a primary result or a test. Among them are the `noai` and `noimageai` values that an art-sharing platform introduced in 2022¹⁰⁹ and the `Content-Signal` header that Cloudflare's Markdown rendering adds by default, which granted every use on 62 of 106 of the Markdown responses received.

Table A6. Exploratory, as registered in protocol section 6.3: signals on the home page (one page per site; says nothing about the rest of a site). Share of sites whose home page was obtained as HTML, sample pooled.

Signal	Sites
<meta name="robots"> of the home page carries noai. Denominator: domains whose home page was obtained as HTML	7 (0.1%)
<meta name="robots"> of the home page carries noimageai. Denominator: domains whose home page was obtained as HTML	6 (0.1%)
<meta name="robots"> of the home page carries noarchive. Denominator: domains whose home page was obtained as HTML	65 (0.9%)
<meta name="robots"> of the home page carries nosnippet. Denominator: domains whose home page was obtained as HTML	2 (<0.1%)
X-Robots-Tag of the home page carries noai or noimageai. Denominator: domains whose home page was obtained as HTML	3 (<0.1%)
The home page reserves text-and-data-mining rights (TDMRep meta element or header set to 1). Denominator: domains whose home page was obtained as HTML	14 (0.2%)
The home page declares rel="api-catalog" (link element or Link header). Denominator: domains whose home page was obtained as HTML	24 (0.3%)
The home page declares rel="describedby" pointing at an llms.txt (link element or Link header). Denominator: domains whose home page was obtained as HTML	18 (0.2%)
The home page declares a Markdown alternate (link element or Link header). Denominator: domains whose home page was obtained as HTML	36 (0.5%)
The home page declares an RSL licence link (link element or Link header). Denominator: domains whose home page was obtained as HTML	0 (0.0%)
The HTML home page response carries a Content-Signal header. Denominator: domains whose home page was obtained as HTML	1 (<0.1%)
The Markdown response carries Content-Signal with ai-train=yes, search=yes and ai-input=yes (the default Cloudflare adds). Denominator: domains that answered the Markdown request with Markdown	62 (58.5%)
The Markdown response grants ai-train while the site-wide Content-Signal in robots.txt withholds it. Denominator: domains that answered the Markdown request with Markdown	4 (3.8%)

Table A7 gives the counts behind Figure 1: sixteen tokens by stratum under both counting rules.

Table A7. Full disallow of 16 crawler tokens by rank stratum, under the explicit and the effective rule. Counts of sites, with their share of the sites with robots.txt rules.

Token	Rule	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)
GPTBot	explicit	82 (15.8%)	198 (11.3%)	169 (9.9%)	112 (6.6%)
GPTBot	effective	115 (22.2%)	255 (14.5%)	249 (14.7%)	240 (14.0%)
ClaudeBot	explicit	81 (15.6%)	179 (10.2%)	140 (8.2%)	79 (4.6%)
ClaudeBot	effective	116 (22.4%)	239 (13.6%)	221 (13.0%)	208 (12.2%)
CCBot	explicit	90 (17.3%)	224 (12.8%)	158 (9.3%)	103 (6.0%)

Token	Rule	S1 (1–1,000)	S2 (1,001–10,000)	S3 (10,001–100,000)	S4 (100,001–1,000,000)
CCBot	effective	132 (25.4%)	283 (16.1%)	237 (13.9%)	164 (9.6%)
Google - Extended	explicit	68 (13.1%)	153 (8.7%)	116 (6.8%)	54 (3.2%)
Google - Extended	effective	101 (19.5%)	210 (12.0%)	194 (11.4%)	182 (10.6%)
Bytespider	explicit	90 (17.3%)	204 (11.6%)	158 (9.3%)	90 (5.3%)
Bytespider	effective	133 (25.6%)	263 (15.0%)	239 (14.1%)	149 (8.7%)
Amazonbot	explicit	50 (9.6%)	135 (7.7%)	131 (7.7%)	80 (4.7%)
Amazonbot	effective	89 (17.1%)	195 (11.1%)	213 (12.5%)	207 (12.1%)
Applebot - Extended	explicit	61 (11.8%)	135 (7.7%)	107 (6.3%)	50 (2.9%)
Applebot - Extended	effective	98 (18.9%)	196 (11.2%)	190 (11.2%)	179 (10.5%)
meta - externalagent	explicit	67 (12.9%)	132 (7.5%)	106 (6.2%)	66 (3.9%)
meta - externalagent	effective	108 (20.8%)	190 (10.8%)	188 (11.1%)	194 (11.4%)
PerplexityBot	explicit	61 (11.8%)	122 (7.0%)	84 (4.9%)	39 (2.3%)
PerplexityBot	effective	94 (18.1%)	180 (10.3%)	166 (9.8%)	168 (9.8%)
OAI - SearchBot	explicit	32 (6.2%)	61 (3.5%)	43 (2.5%)	21 (1.2%)
OAI - SearchBot	effective	67 (12.9%)	117 (6.7%)	128 (7.5%)	153 (9.0%)
Claude - SearchBot	explicit	35 (6.7%)	60 (3.4%)	23 (1.4%)	13 (0.8%)
Claude - SearchBot	effective	74 (14.3%)	119 (6.8%)	108 (6.4%)	147 (8.6%)
ChatGPT - User	explicit	43 (8.3%)	97 (5.5%)	88 (5.2%)	58 (3.4%)
ChatGPT - User	effective	78 (15.0%)	153 (8.7%)	170 (10.0%)	187 (10.9%)
Claude - User	explicit	40 (7.7%)	61 (3.5%)	24 (1.4%)	15 (0.9%)
Claude - User	effective	78 (15.0%)	120 (6.8%)	109 (6.4%)	149 (8.7%)
Perplexity - User	explicit	38 (7.3%)	63 (3.6%)	24 (1.4%)	19 (1.1%)
Perplexity - User	effective	83 (16.0%)	124 (7.1%)	110 (6.5%)	153 (9.0%)
Googlebot	explicit	3 (0.6%)	3 (0.2%)	2 (0.1%)	2 (0.1%)
Googlebot	effective	23 (4.4%)	44 (2.5%)	59 (3.5%)	48 (2.8%)
bingbot	explicit	3 (0.6%)	10 (0.6%)	11 (0.6%)	12 (0.7%)
bingbot	effective	25 (4.8%)	56 (3.2%)	75 (4.4%)	65 (3.8%)

Appendix B. The searches behind the negative claims

Every statement in this report that something was "not found" refers to the following searches, all run on 2026-10-04, and means no more than that these queries did not find it. On arXiv's listing search, all fields: "robots.txt", "llms.txt", "ai.txt", "AI crawlers", "Robots Exclusion Protocol", "agent card", "agent cards" discovery, "Model Context Protocol" measurement, "agent readiness", "agentic web" measurement, "Content Signals", "Really Simple Licensing", and several longer phrases combining AI crawlers, opt-out, Tranco and measurement, with author searches for the two groups whose work was closest. On OpenAlex, works from 2024 onward: "llms.txt", "robots.txt" AI crawlers websites block, "robots.txt" "generative AI" opt-out adoption, llms.txt adoption websites measurement, AI crawlers robots.txt large-scale measurement websites, agent card well-known discovery websites measurement adoption and robots.txt user agents AI bots Tranco, the longer result lists screened to their first fifteen or twenty entries. On Crossref, DOI and title lookups for verification. On a general web search, title queries for the academic papers and reports named in Section 2 and queries for llms.txt, RSL, Content-Signal and Agent Card adoption scans. Semantic Scholar answered 429 to every query, the ACM Digital Library answered 403, dblp presented a bot challenge, and Google Scholar was not queried; ACM-published items were located through Crossref, OpenAlex and web search. The queries were in English only, industry scans were found by web search and are therefore biased towards well-ranked vendor pages, and works published after 2026-10-04, theses not indexed by these services and paywalled reports are not covered. Hoffmann et al.'s article was located by these searches, but only its abstract and code were read.