Barkhausen AI

CENSUS · BA-D-2026-01

AI-crawler access across four sectors: a robots.txt census of 2,000 domains

On 2026-07-09 the robots.txt policy of 2,000 domains — four documented frames of 500 universities, news outlets, e-commerce sites, and U.S.

“Every leap begins as a crackle.”

DOCUMENT	BA-D-2026-01
SERIES	Census
DATE	2026
SUBJECTS	Crawlers & access; Corpora & training data
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

On 2026-07-09 the robots.txt policy of 2,000 domains — four documented frames of 500 universities, news outlets, e-commerce sites, and U.S. federal government domains, each built from a cited public source and ordered by a public ranking (Tranco traffic rank, or reported enrollment for the U.S. universities) — was fetched and evaluated against sixteen crawler tokens. Among domains with a determinable policy, news sites root-blocked AI crawlers far more than any other sector: 66.7% (288/432) blocked at least one AI-specific token, against 7.8% (33/423) of universities. A distinct pattern recurs in news: 45.8% (198/432) root-blocked a retrieval-class crawler that supplies AI answer engines while staying open to general search. Non-response was itself a finding — the policy was undeterminable for 40.6% (203/500) of government and 37.8% (189/500) of e-commerce domains. This census enumerates the four frames completely; it makes no sampling inference to all universities, news sites, stores, or agencies.

IN BRIEF

Websites can ask automated crawlers to stay away by publishing a small file, robots.txt, that names which crawlers may visit. This census read that file for 2,000 real websites on a single day — 500 each of universities, news outlets, online stores, and U.S. federal government sites, every list drawn from a named public source — and checked, for sixteen named crawlers, whether each site's file told that crawler to keep out of the whole site. News sites stood apart. Two-thirds of the news sites whose policy could be read (66.7%, 288 of 432) blocked at least one AI crawler, against fewer than one in twelve universities (7.8%, 33 of 423). More striking, nearly half the news sites (45.8%, 198 of 432) blocked a crawler whose job is to feed AI answer engines — the crawler that decides whether a site can appear in an AI-generated answer at all — while leaving ordinary search crawlers alone. The data cannot say whether that was intended or a mis-aimed attempt to opt out of training, only that the configuration removes those sites from specific AI answers while keeping them in general search. A large share of sites could not be read at all: on more than a third of government and e-commerce sites, the file returned an error page or a blocking challenge, so their policy is unknowable rather than absent. Every figure here counts a fully enumerated list of sites, not a sample, and describes those lists on one day only.

KEY FINDINGS

1. Across four documented frames of 500 domains each, fetched on 2026-07-09, news sites root-blocked AI crawlers far more than any other sector: 66.7% (288 of 432 news domains with a determinable robots.txt policy) blocked at least one AI-specific crawler token, against 7.8% (33 of 423) of universities, 25.1% (78 of 311) of e-commerce sites, and 8.4% (25 of 297) of U.S. federal government domains.
2. Among news domains with a determinable policy, 45.8% (198 of 432) root-blocked at least one retrieval-class crawler (OAI-SearchBot, Claude-SearchBot, or PerplexityBot) while leaving the default group open to general search — removing the site from specific AI answer engines while remaining generally crawlable. The same configuration appeared in 12.5% (39 of 311) of e-commerce, 1.0% (3 of 297) of government, and 0.7% (3 of 423) of university domains, as fetched on 2026-07-09.
3. OpenAI's GPTBot training crawler was root-blocked by 53.0% (229 of 432) of news domains with a determinable policy versus 5.0% (21 of 423) of universities; Common Crawl's CCBot was the single most-blocked token in three of the four sectors — news 57.9% (250 of 432), e-commerce 21.2% (66 of 311), and government 6.4% (19 of 297) — with Bytespider the most-blocked in universities at 6.1% (26 of 423), all as fetched on 2026-07-09.
4. A clean training-only opt-out — root-blocking at least one of the five operator-documented training tokens (GPTBot, ClaudeBot, CCBot, Google-Extended, Applebot-Extended) while allowing all three retrieval tokens — was uncommon even in news, at 18.3% (79 of 432); its strict form, blocking all five training tokens while allowing all three retrieval tokens, appeared in only 5.6% (24 of 432) of news domains with a determinable policy on 2026-07-09.
5. Blocking a crawler and naming it are different acts: of the 229 news domains that root-blocked GPTBot, 227 named it in an explicit user-agent group, and news domains named GPTBot 248 times in total; universities almost never named it, leaving 398 of 423 university domains without a GPTBot group and 394 of those neither naming nor blocking it, all as fetched on 2026-07-09.
6. A /llms.txt file returned HTTP 200 for 299 of the 2,000 domains probed (universities 57, news 75, e-commerce 91, government 76, each of 500 domains), but only 93 of those responses (universities 16, news 24, e-commerce 44, government 9) resembled a valid llms.txt file rather than an HTML page or soft-404 — most 'present' /llms.txt responses were not valid llms.txt files, as probed on 2026-07-09.
7. Non-response is a first-class result: the robots.txt policy was undeterminable for 40.6% (203 of 500) of U.S. federal government domains and 37.8% (189 of 500) of e-commerce domains — dominated by HTTP 403 responses and bot-challenge pages, plus registry domains that resolve but serve no website — so those sectors' block rates describe only the determinable remainder of each frame, not the full 500, as fetched on 2026-07-09.

CONTENTS

1 1. Summary

On 2026-07-09, the `robots.txt` file of 2,000 web domains was fetched and evaluated against sixteen crawler tokens. The domains form four frames of 500 each — universities, news outlets, e-commerce sites, and U.S. federal government domains — and each frame is a top-N selection from a cited public source under a documented ordering — Tranco traffic rank [2], or reported enrollment for the U.S. university portion — so that a reader can re-derive the exact list. This is a census of those four documented frames, not a sample of any larger population: every figure below is an exact count and percentage of a fully enumerated frame, and no figure is generalized to "all universities" or "all news sites." The unit of analysis is the domain's root-level robots.txt policy toward each token, as served on one day.

The organizing finding is a sharp separation between sectors. Among domains whose policy could be determined, news sites root-blocked AI crawlers far more than the other three sectors: 66.7% (288 of 432 news domains with a determinable policy) blocked at least one AI-specific token, against 25.1% (78 of 311) of e-commerce sites, 8.4% (25 of 297) of government domains, and 7.8% (33 of 423) of universities. Within news, a specific configuration recurs: 45.8% (198 of 432) root-blocked at least one retrieval-class crawler — the class that supplies AI answer engines — while leaving the default group open, so the site remains generally crawlable but is removed from specific AI answers. The same configuration is rare elsewhere (e-commerce 12.5%, government 1.0%, universities 0.7%).

Two disciplines run through the report. First, the denominator is stated everywhere. The access figures are computed only over domains with a *determinable* policy — a real robots.txt file or a confirmed absence of one — because a domain that answered `/robots.txt` with an HTTP 403 or a bot-challenge page has an *unknowable* policy, which is neither "allows everything" nor "blocks AI." Second, that unknowable share is reported as a finding rather than discarded: the policy could not be determined for 40.6% (203 of 500) of government and 37.8% (189 of 500) of e-commerce domains, a rate high enough that those sectors' block figures describe only their determinable remainders. What follows reports the access matrix in full, the retrieval-block pattern in both directions, the asymmetry between blocking a crawler and naming it, the state of `/llms.txt` adoption, and the composition of the non-response.

2 2. Scope and method

2.1 The four frames

Each frame is a top-500 selection from an authoritative public source, imposed on a documented traffic ordering so the selection is reproducible. The universal ordering is the Tranco research list (list ID 46XZX, a 30-day list ending 2026-07-08), a manipulation-hardened aggregate of five ranking providers [2]; where a source list exceeds 500 and carries no intrinsic size metric, its members are intersected with Tranco and taken in rank order. Registrable domains (eTLD+1) are computed with the Public Suffix List [3]. The four frames are disjoint: 2,000 rows, 2,000 distinct domains. One domain that appeared in both the raw news and university source lists was assigned to universities and removed from news, with news backfilled to 500 from the next-ranked entries.

- **Universities (500).** 300 U.S. institutions ordered by reported enrollment from the U.S. Department of Education's College Scorecard [4], plus 200 non-U.S. institutions from the Hipolabs university-domains list [5] ordered by Tranco rank. Enrollment ordering surfaces large online and for-profit institutions first; the two orderings are not a single uniform ranking.

- **News** (**500**). The palewire news-homepages source list [6], reduced to registrable domains, de-duplicated, intersected with Tranco, and taken in rank order. The source list is US-heavy (about 65%), and the traffic cut favors larger outlets; some high-ranked entries are portals with news operations rather than pure publishers.
- **E-commerce** (**500**). The "shopping" category of the UT1 (Université Toulouse Capitole) web-filtering blacklists [7], restricted to entries that are themselves a bare registrable domain, intersected with Tranco, taken in rank order after an explicit CDN/infrastructure exclusion list. A web-filtering category is imperfect: a small residue of platforms and non-storefronts survives the filter.
- **Government** (**500**). The CISA `current-federal.csv` registry of U.S. federal executive, legislative, and judicial `.gov` domains [8], intersected with Tranco and taken in rank order. This frame is U.S.-only and federal-only; state, local, and tribal domains are excluded, and each agency domain is a separate row with no agency-level rollup.

The selection rules, source URLs, licenses, and every non-mechanical judgment call are recorded in the dataset's methodology and summarized in the appendix. The frames are traffic-biased by construction — a crawler-access census is concerned with the trafficked domains crawlers actually visit — so they are not uniform samples of their sectors, and their differing construction means cross-sector comparisons are comparisons of like method on unlike frames, not of representative populations.

2.2 Fetch and category operationalization

Each domain's `https://<domain>/robots.txt` was requested once (GET, up to five redirects, 15-second timeout, user-agent `Mozilla/5.0 (compatible; research-fetch)`), with a single retry over `http://` on a connect-level failure. Every response's exact bytes and headers were stored. The single most consequential decision is the category assigned to each domain, because it defines what counts as a readable policy:

CATEGORY	TRIGGER	MEANING	TOKEN FIELDS
`parsed`	HTTP 200, body is not HTML-leading, and is texty or contains `user-agent`	A real robots file was evaluated	populated
`no_robots_file`	HTTP 404 or 410	No file → crawling allowed by default; evaluated as an empty ruleset	populated (all allowed)
`unparseable_blocked`	401/403/5xx, any other non-200, or a 200 whose body is HTML-leading or non-texty without robots markers	**Unknown** — not "allows all" and not "blocks AI"	empty
`fetch_error`	Network failure (timeout/DNS/TLS/other) after the retry	Could not fetch	empty

The HTML-leading guard matters: several domains return a 34 KB HTML challenge page with HTTP 200 at `/robots.txt`; counting that as "empty robots, therefore allows all" would be a false positive for AI access, so a body beginning with an HTML doctype or tag is classified `unparseable_blocked` even if the substring `user-agent` appears inside the HTML. The **knowable denominator** used throughout sections 3–6 is therefore the count of `parsed` plus `no_robots_file` domains; `unparseable_blocked` and `fetch_error` are excluded from access denominators and reported on their own in section 7. Per sector, the knowable denominators are: news 432, universities 423, e-commerce 311, government 297.

2.3 The block definition and the token classes

The headline measure for each token is `root_blocked`: the token cannot fetch `/` — full-site exclusion at the root path — evaluated with the `protego` library (version 0.6.2), which implements the RFC 9309 reference semantics [1]. This measures the root path only: a domain that allows `/` but disallows a deep path such as `/news/` is counted as *not* root-blocked, so path-level partial blocks are outside scope. RFC 9309 is explicit that robots.txt rules "are not a form of access authorization" [1]; a block here is a published request a crawler is asked to honor, not an enforced barrier, and this census measures the request, not any crawler's compliance.

The sixteen tokens are classified by function, following the crawler taxonomy in BA-C-6, because the consequence of blocking a token depends on its class. Fourteen tokens are **AI-specific** (every token except Googlebot and bingbot); "blocks ≥1 AI" means root-blocking at least one of those fourteen. The three **retrieval-class** tokens — OAI-SearchBot, Claude-SearchBot, PerplexityBot — are the load-bearing ones for AI answer visibility: BA-C-6 §2.2 defines a retrieval crawler as one whose index an AI system queries at answer time, so blocking it tends to remove a site from that system's answers. The five operator-documented **training** tokens with per-token opt-outs are GPTBot, ClaudeBot, CCBot, Google-Extended, and Applebot-Extended [9][10][11]; Amazonbot, Bytespider, and meta-externalagent are also training-class but are not among the five with a distinct documented opt-out. Googlebot and bingbot are **search**-class. This census reports crawler-access configuration only; it makes no visibility measurement and no engine-behavior claim (visibility metrics are defined in BA-C-2, and the sampling that would estimate them in BA-C-3).

3 3. Results: the access matrix

The full matrix reports, for each of the sixteen tokens and each sector, the share of that sector's determinable domains that root-block the token, with the count in parentheses over the sector's knowable denominator (news 432, e-commerce 311, government 297, universities 423). The row order groups tokens by BA-C-6 class.

TOKEN	CLASS	NEWS (N=432)	E-COMMERCE (N=311)	GOVERNMENT (N=297)	UNIVERSITIES (N=423)
GPTBot	training	53.0% (229)	19.0% (59)	6.1% (18)	5.0% (21)
ClaudeBot	training	53.0% (229)	18.3% (57)	5.7% (17)	3.8% (16)
CCBot	training	57.9% (250)	21.2% (66)	6.4% (19)	4.5% (19)
Google-Extended	training	50.2% (217)	13.2% (41)	5.7% (17)	3.8% (16)
Applebot-Extended	training	45.6% (197)	9.6% (30)	4.4% (13)	3.5% (15)
Amazonbot	training	39.1% (169)	10.6% (33)	5.4% (16)	4.0% (17)
Bytespider	training	51.9% (224)	19.3% (60)	5.1% (15)	6.1% (26)
meta-externalagent	training	42.1% (182)	17.0% (53)	4.7% (14)	4.0% (17)
OAI-SearchBot	retrieval	28.2% (122)	8.7% (27)	2.4% (7)	1.4% (6)
Claude-SearchBot	retrieval	28.7% (124)	8.7% (27)	2.0% (6)	0.9% (4)
PerplexityBot	retrieval	46.5% (201)	14.1% (44)	2.7% (8)	1.4% (6)
ChatGPT-User	user-fetch	38.2% (165)	9.0% (28)	4.0% (12)	1.4% (6)
Claude-User	user-fetch	28.7% (124)	8.7% (27)	2.0% (6)	0.9% (4)
Perplexity-User	user-fetch	28.5% (123)	9.0% (28)	2.4% (7)	1.4% (6)
Googlebot	search	0.0% (0)	0.3% (1)	1.7% (5)	0.2% (1)
bingbot	search	0.0% (0)	0.3% (1)	2.0% (6)	0.2% (1)

Three features of the matrix are worth stating in prose, because each survives excerpting. First, the sector gradient is large and consistent across tokens: for almost every AI-specific token, the news column exceeds the e-commerce column, which exceeds government and universities. Second, the search-class tokens are the mirror image — Googlebot and bingbot are root-blocked by essentially no one (0 of 432 news domains block Googlebot; the highest search-class figure anywhere is 2.0% (6 of 297) of government domains blocking bingbot). A site that blocks AI crawlers while leaving Googlebot open is choosing to stay in general search, and the matrix shows that is the near-universal choice. Third, the retrieval and user-fetch classes are blocked at roughly half the rate of the training class within each sector — in news, CCBot (training) is blocked by 57.9% (250 of 432) while OAI-SearchBot (retrieval) is blocked by 28.2% (122 of 432) — which is the raw material for the pattern examined next.

Two summary rows follow from the matrix. The share blocking at least one AI-specific token is 66.7% (288 of 432) for news, 25.1% (78 of 311) for e-commerce, 8.4% (25 of 297) for government, and 7.8% (33 of 423) for universities (Figure 1). CCBot is the most-blocked single token in news, e-commerce, and government; Bytespider is the most-blocked in universities at 6.1% (26 of 423).

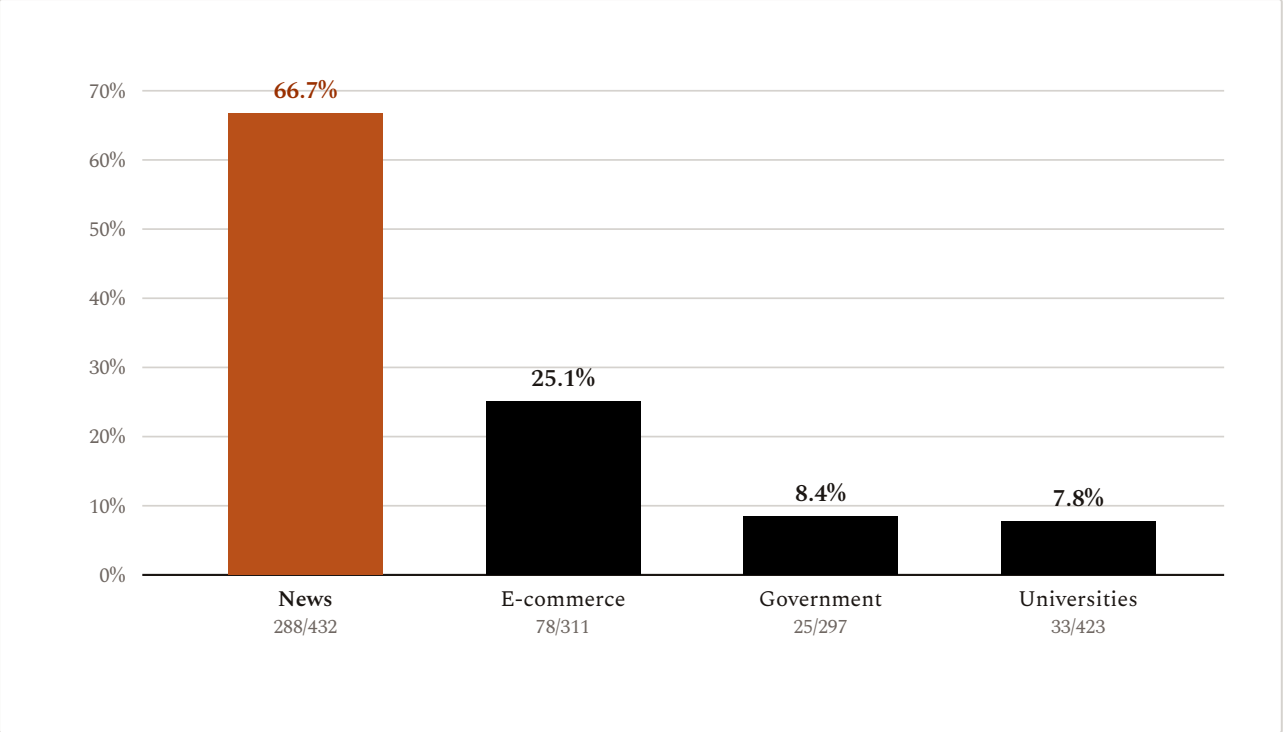

Figure 1. Share of each sector's domains with a determinable robots.txt policy that root-block at least one AI-specific crawler token (any of the fourteen tokens other than Googlebot and bingbot). Bars are exact frame counts, not estimates; denominators are each sector's knowable domains.

BA-D-2026-01, dataset crawler-access-2026 · news n=432; e-commerce n=311; government n=297; universities n=423 · 2026-07-09

4 4. The retrieval-block pattern, in both directions

The matrix shows that retrieval-class tokens are blocked at roughly half the rate of training tokens. The consequential question is what the retrieval blocks *co-occur with*, because a site's posture toward AI is legible only from the combination of tokens it blocks. BA-C-6 §5 names the failure this exposes — the mis-block, in which a rule aimed at one class silently governs another. This section operationalizes it in both directions and reports the counts.

A **retrieval mis-block** is defined here as a domain that root-blocks at least one retrieval-class token (OAI-SearchBot, Claude-SearchBot, or PerplexityBot) while *not* blocking the default group (`star_root_blocked` is false, so the site is not simply blocking everyone). The complement of interest is a **clean training-only opt-out**: a domain that root-blocks at least one of the five documented training tokens while allowing all three retrieval tokens — the configuration that expresses "do not train on me, but do keep me eligible for AI answers." Its strict form blocks all five training tokens while still allowing all three retrieval tokens. Both are computed over the knowable denominator.

SECTOR	RETRIEVAL MIS-BLOCK	CLEAN TRAINING-ONLY OPT-OUT	CLEAN OPT-OUT (STRICT)
News (n=432)	45.8% (198)	18.3% (79)	5.6% (24)
E-commerce (n=311)	12.5% (39)	8.4% (26)	1.6% (5)
Government (n=297)	1.0% (3)	5.1% (15)	2.4% (7)
Universities (n=423)	0.7% (3)	3.8% (16)	2.4% (10)

In news, the mis-block (45.8%, 198 of 432) is more than twice as common as the clean training-only opt-out (18.3%, 79 of 432), and the strict clean opt-out is rare (5.6%, 24 of 432). Read against the "no one blocks Googlebot" result from section 3, the dominant news configuration is a site that stays in general search but has removed itself from at least one AI answer engine's retrieval index. This census measures the *configuration*, not the *intent* behind it. Two readings are consistent with the same bytes: a deliberate decision to exit AI answers while remaining in search, or a training opt-out that reached a retrieval token because the operator did not distinguish the classes. The data cannot separate them — that separation would require asking the operators — and BA-C-6 §5 is careful for the same reason: it describes the mechanism by which one rule binds another, not the state of mind of whoever wrote the rule (Figure 2).

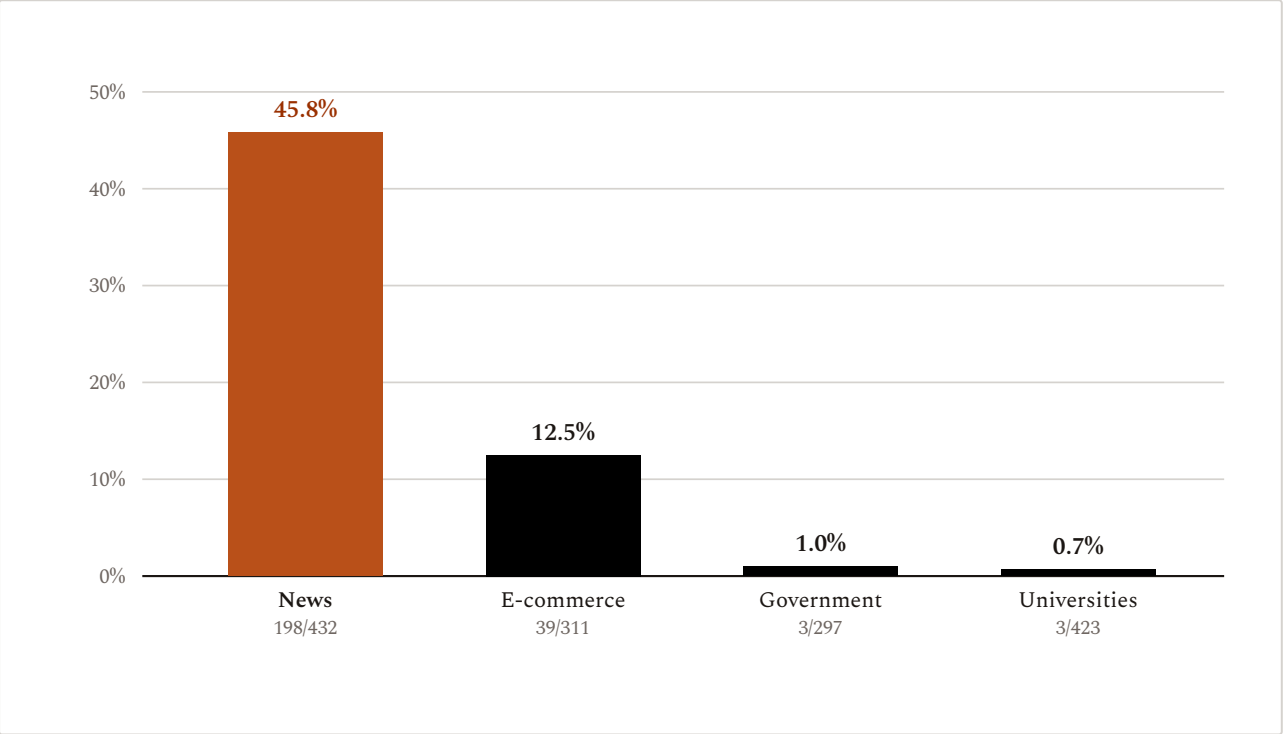

Figure 2. Share of each sector's determinable domains that root-block at least one retrieval-class crawler (OAI-SearchBot, Claude-SearchBot, or PerplexityBot) while leaving the default group open — the retrieval mis-block. Exact frame counts over each sector's knowable denominator.

`BA-D-2026-01, dataset crawler-access-2026 · news n=432; e-commerce n=311; government n=297; universities n=423 · 2026-07-09`

5 5. Named-but-allowed versus blocked-without-naming

Blocking a crawler and naming it are separate acts, and the census records both. `root_blocked` measures effective access under RFC 9309's prefix-based user-agent matching; a second field, `has_specific_group`, records whether the file contains a `user-agent:` line that is an exact, case-insensitive match for the token — whether the operator named the crawler explicitly. The two can diverge: a group line `User-agent: Claude` prefix-matches and therefore blocks ClaudeBot, Claude-User, and Claude-SearchBot without naming any of them exactly, while `User-agent: Googlebot` does not match `Google-Extended` at all. The crosstab of the two fields separates a deliberate, targeted rule from a block inherited by prefix or breadth. The census stores this crosstab for GPTBot and OAI-SearchBot.

GPTBot (knowable denominators as in section 3):

SECTOR	NAMED + BLOCKED	NAMED + ALLOWED	UNNAMED + BLOCKED	UNNAMED + ALLOWED
News (n=432)	227	21	2	182
E-commerce (n=311)	54	20	5	232
Government (n=297)	13	1	5	278
Universities (n=423)	17	8	4	394

OAI-SearchBot:

SECTOR	NAMED + BLOCKED	NAMED + ALLOWED	UNNAMED + BLOCKED	UNNAMED + ALLOWED
News (n=432)	121	23	1	287
E-commerce (n=311)	21	19	6	265
Government (n=297)	2	0	5	290
Universities (n=423)	2	5	4	412

The asymmetry between news and universities is the finding. In news, GPTBot is named in a specific group 248 times (227 + 21), and of the 229 domains that block it, 227 name it — the news blocks are explicitly named at the token level, not side effects of a broad rule. Universities are the opposite: 398 of 423 university domains carry no GPTBot group at all, and 394 of those neither name nor block it — the default posture is silence, and silence resolves to access. The named-but-allowed cells document a third, distinct decision: 21 news domains and 8 university domains name GPTBot and then allow it, and for OAI-SearchBot the named-but-allowed count in news (23) exceeds the number that name it and block it in some smaller sectors — an explicit token-level admission of a crawler the file's author named. Reading the two tables together, retrieval-class naming is rarer than training-class naming even in news (OAI-SearchBot is named 144 times against GPTBot's 248), consistent with retrieval tokens being newer and less widely templated into robots.txt files than the training tokens.

6 6. llms.txt: present versus valid

A `/llms.txt` file — a proposed convention for a site to publish guidance addressed to language models — was probed on every domain regardless of its robots.txt outcome, so its denominator is the full 500 per sector. The probe records two fields: `present` (an HTTP 200 at `/llms.txt`) and `looks_valid`, a content-sanity heuristic that requires the response to be non-HTML-leading and to look like markdown (a leading `#`, a `text/markdown` or `text/plain` body with content, or a markdown heading, link, or bullet). The heuristic exists precisely because a 200 is not evidence of a real file: many sites answer `/llms.txt` with an HTML soft-404.

The gap between the two fields is the result. A `/llms.txt` returned HTTP 200 for 299 of the 2,000 domains — universities 57, news 75, e-commerce 91, government 76, each over 500 — but only 93 of those responses looked like a valid llms.txt file: universities 16 of 57, news 24 of 75, e-commerce 44 of 91, and government just 9 of 76. Most "present" responses are not valid llms.txt files. The false-positive rate is highest in government, where 76 domains return a 200 at `/llms.txt` but only 9 (11.8% of the 76) pass the content check — the rest are HTML pages served for any path. Presence of the file, moreover, is not evidence that any model consumes it: BA-W-2026-01 assembles the public record on this point, which finds no measurable association between llms.txt and AI citation and that AI crawlers rarely request the

file. This census adds only that even the raw presence figure is inflated: counting HTTP 200 responses overstates real adoption roughly three-to-one (299 present versus 93 that look valid).

7 7. Non-response and unknowables

A domain whose `/robots.txt` returned a 403, a bot-challenge page, or a network error has an *unknowable* policy: the census cannot say whether it blocks AI crawlers, because it never received a readable file. Treating these as "allows all" would be a systematic false positive, and treating them as "blocks all" a false negative; both would corrupt the access figures. They are therefore excluded from the access denominators and reported here as a first-class result, because the *rate* and *composition* of non-response is itself informative.

SECTOR	UNKNOWABLE	RATE (OF 500)	UNPARSEABLE_BLOCKED	FETCH_ERROR
News	68	13.6%	50	18
Universities	77	15.4%	38	39
E-commerce	189	37.8%	167	22
Government	203	40.6%	128	75

The two high-unknowable sectors are unknowable for different reasons. E-commerce is dominated by active blocking at the HTTP layer: of its 167 `unparseable_blocked` domains, 129 returned HTTP 403 and 26 returned an HTML challenge page at HTTP 200 — bot-protection infrastructure that refuses the `/robots.txt` request itself. Government is a mix: 128 `unparseable_blocked` (76 of them 403, 45 HTML-challenge) plus a large `fetch_error` count of 75, of which 38 were DNS failures — a transport classification recorded at collection time; the published CSV carries the `fetch_error` category, and the transport split lives in the collection record. A `fetch_error` bounds unreachability *at the apex host as fetched*, not an institution's absence from the web: the census requested `https://<domain>/robots.txt` at the apex registrable domain (with an `http://` retry) and did not fall back to a `www.` host, so a DNS failure records only that the apex did not resolve. Many federal `.gov` entries are registry records whose apex carries no address, an artifact of enumerating a domain registry rather than a set of live apex web servers. The consequence for interpretation is explicit: government's 8.4% "blocks ≥1 AI" and e-commerce's 25.1% are computed over 297 and 311 determinable domains respectively, not over 500, and a policy that is unknowable is not a policy that is absent. Where the unknowable share is this large, the determinable remainder may not resemble the whole frame — the domains that answer cleanly can differ systematically from those behind a challenge wall — which bounds how far even a frame-level figure can be pushed.

8 8. Reproducibility appendix

This is a public-data census, so full reproduction from public sources is the intent. The frames, the fetch, and the parser are specified here at the level needed to rebuild the dataset.

Frames. Universal ordering: Tranco list 46XZX (30-day list ending 2026-07-08, Dowdall combination, pay-level domains, providers CrUX, Farsight, Majestic, Cloudflare Radar, Cisco Umbrella) [2]. Registrable domains via the Public Suffix List snapshot [3]. Sources and selection: universities = College Scorecard top-300 by enrollment (retrieved through the Scorecard public API at api.data.gov/ed/collegescorecard/v1/schools, ordered by latest reported enrollment) [4] plus Hipolabs top-200 non-U.S. by Tranco rank [5]; news = palewire news-homepages ∩ Tranco, top 500 [6]; e-commerce =

UT1 "shopping" bare-registrable entries ∩ Tranco, top 500 after excluding eleven CDN/infrastructure domains (media-amazon.com, ssl-images-amazon.com, images-amazon.com, amazon-adsystem.com, alicdn.com, shopifycdn.com, espncdn.com, ebaystatic.com, ebayimg.com, amaicdn.com, booztcdn.com) [7]; government = CISA `current-federal.csv` ∩ Tranco, top 500 [8]. Frames were de-duplicated within and across sectors (2,000 rows, 2,000 distinct domains; one cross-list duplicate was assigned to universities and news backfilled).

Fetch. `GET https://<domain>/robots.txt`, up to five redirects, 15-second timeout, user-agent `Mozilla/5.0 (compatible; research-fetch)`, one `http://` retry on a connect-level failure; HTTP/1.1 only. Exact response bytes and headers were stored, and each result carries the SHA-256 of the stored bytes, so any derived field can be recomputed offline without re-fetching. Category assignment follows the table in section 2.2, including the HTML-leading guard that reclassifies an HTML page served at `/robots.txt` as `unparseable_blocked`.

Parser. `protego` 0.6.2, whose RFC 9309 conformance was checked against the eight reference correctness vectors. Two semantics matter for interpretation. User-agent matching is **prefix-based** (RFC 9309 §2.2.1): a group value is matched as a case-insensitive prefix of the crawler token, so `User-agent: Claude` sets `root_blocked` for ClaudeBot, Claude-User, and Claude-SearchBot, while their `has_specific_group` stays false because `Claude` is not an exact match; and `User-agent: Googlebot` does not reach `Google-Extended`. Rule precedence uses **longest-match with Allow winning ties**: `Disallow: /` combined with `Allow: /` leaves the root *not* blocked. `root_blocked` is `NOT can_fetch("/")`; the `*`-group decision is probed separately with a token guaranteed to have no specific group (`star_root_blocked`). One non-conformance is not repaired: a comma-joined line such as `User-agent: GPTBot, CCBot` is treated as a single token by both protego and the naming scan, matching neither GPTBot nor CCBot — the behavior a conformant crawler exhibits. The `/llms.txt` probe is `HEAD` then `GET`, with `present = (status == 200)` and the `looks_valid` heuristic described in section 6.

9 Limitations

Frame, not population. Every figure is an exact count over a documented frame of 500 domains, not an estimate for a sector. The frames are traffic-biased top-N selections from specific public sources — a US-heavy news list, an enrollment-ordered then traffic-ordered university list, a web-filtering shopping category, a U.S. federal `.gov` registry — so no figure here should be read as "X% of universities" or "X% of news sites." Cross-sector differences are differences between these particular frames built by like method, not between representative populations, and the government and university frames combine two orderings rather than one uniform ranking.

Single-day snapshot. This census records each frame's robots.txt policy as served on 2026-07-09, within a collection window of 23:07–23:13 UTC. Sites revise robots.txt without notice, and results should be assumed perishable; any figure here should be checked against a current fetch before it is relied upon.

Root-block operationalization. The block measure is root-path exclusion only. A site that allows `/` but disallows a deep path is counted as not blocking, so path-level rules — a real and common way to govern crawlers — are invisible to this census, and the true restriction on some domains is understated. Relatedly, `root_blocked` measures a published request under RFC 9309, not any crawler's compliance; whether a crawler honors the file is a separate empirical question this census does not address.

Prefix-matching nuance. Because user-agent matching is prefix-based, a broad group value can block several tokens at once without naming any, and the block and naming fields answer genuinely different

questions. The crosstabs in section 5 are the correct reading of "named," and `root_blocked` the correct reading of "blocked"; conflating them would misstate both.

Unknowable concentration. Non-response is uneven and heavy in two sectors — 40.6% of government and 37.8% of e-commerce domains — so those sectors' access figures rest on determinable remainders (297 and 311 domains) that may differ systematically from the domains behind a 403 or a challenge page. An unknowable policy is not an absent one; it is missing data with a plausibly non-random mechanism, and the block rates should be read with that selection in mind.

Apex-only fetch. The fetch targeted the apex registrable domain over `https` with an `http://` retry and no `www.` fallback — a design choice recorded in the collection tooling. A DNS-class `fetch_error` therefore means the apex host did not resolve, which for some domains understates web presence rather than establishing its absence. A post-collection DNS diagnostic (2026-07-09, not reproducible from the published dataset) found that 18 of the 19 university DNS failures and 16 of the 38 government DNS failures resolved on the corresponding `www` host at review time, while the remaining 22 government entries were unresolved on both and are consistent with decommissioned registry records. These figures are diagnostic context only: no published count was adjusted, and the dataset stands exactly as collected.

Ordering bias. The Tranco and enrollment orderings favor higher-traffic and larger institutions within each source. This is appropriate for a census of the domains crawlers actually visit, but it means low-traffic domains are under-represented, and any behavior that correlates with size is entangled with the selection.

llms.txt validity is a heuristic. `looks_valid` is a content-sanity screen, not a specification validator; it can admit a malformed file that happens to start with a heading or reject an unusual but legitimate one. The present-versus-valid gap is robust to the heuristic's exact threshold — the soft-404 pages it screens out are plainly not llms.txt files — but the valid counts should be read as "resembles a valid llms.txt file," not "conforms to the llms.txt convention."

REFERENCES

1. M. Koster, G. Illyes, H. Zeller, and L. Sassman, IETF. *RFC 9309: Robots Exclusion Protocol* (2022). https://www.rfc-editor.org/rfc/rfc9309.html Accessed 2026-07-09. [archived]
2. V. Le Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, Proceedings of the Network and Distributed System Security Symposium (NDSS). *Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation* (2019). https://doi.org/10.14722/ndss.2019.23386 Accessed 2026-07-09.
3. Public Suffix List (Mozilla Foundation). *Public Suffix List (public_suffix_list.dat)* (2026). https://publicsuffix.org/list/public_suffix_list.dat Accessed 2026-07-09. [archived]
4. U.S. Department of Education, College Scorecard. *College Scorecard institution-level data (public API)* (2026). https://collegescorecard.ed.gov/data/ Accessed 2026-07-09. [archived]
5. Hipolabs. *university-domains-list (world_universities_and_domains.json)* (2026). https://raw.githubusercontent.com/Hipo/university-domains-list/master/world_universities_and_domains.json Accessed 2026-07-09. [archived]
6. B. Welsh, palewire, news-homepages project. *news-homepages source list (sites.csv)* (2026). https://raw.githubusercontent.com/palewire/news-homepages/main/newshomepages/sources/sites.csv Accessed 2026-07-09. [archived]
7. F. Prigent, Université Toulouse 1 Capitole, web-filtering blacklists. *UT1 blacklists, shopping category (shopping.tar.gz)* (2026). https://dsi.ut-capitole.fr/blacklists/download/shopping.tar.gz Accessed 2026-07-09. [archived]
8. Cybersecurity and Infrastructure Security Agency (CISA), dotgov-data. *Federal .gov domain registry (current-federal.csv)* (2026). https://raw.githubusercontent.com/cisagov/dotgov-data/main/current-federal.csv Accessed 2026-07-09. [archived]
9. OpenAI. *Overview of OpenAI Crawlers*. https://developers.openai.com/api/docs/bots Accessed 2026-07-09. [archived]

10. Google, Crawling infrastructure documentation. *Google's common crawlers.* https://developers.google.com/crawling/docs/crawlers-fetchers/google-common-crawlers Accessed 2026-07-09. [archived]

11. Common Crawl Foundation. *CCBot*. https://commoncrawl.org/ccbot Accessed 2026-07-09. [archived]

HOW TO CITE

Barkhausen AI (2026). AI-crawler access across four sectors: a robots.txt census of 2,000 domains. https://barkhausen.ai/research/crawler-access-census-2026/

Barkhausen AI · Est. MMXXVI · *Every leap begins as a crackle.*