PROGEOLAB

PROGEOLAB RESEARCH

robots.txt and AI

How Fortune 500 companies police AI crawlers · April 2026

PROGEOLAB

April 2026

progeolab.ai/research

Contents

CHAPTER 1

Executive Summary

This report presents the first large-scale analysis of robots.txt AI bot policies across the Fortune Global 500. By parsing the raw robots.txt files of 267 companies and cross-referencing declared policies against actual crawler access measured with four user agents, we reveal a landscape of absent policy, inconsistent enforcement, and untapped strategic opportunity.

Key Findings

Only 20 of 267 companies name any AI bot in their robots.txt. The remaining 247 companies (92.5%) have no explicit AI crawler policy. Their AI accessibility is governed by wildcard rules designed for search engines and by WAF configurations managed independently of content policy. (Chapter 3)

Amazon blocks 16 AI bots — the most comprehensive blocker. Amazon names and blocks every known AI crawler variant including training bots (GPTBot), retrieval bots (ChatGPT-User), and research bots (AI2Bot). No other Fortune 500 company approaches this thoroughness. (Chapter 4)

HP allows 10 AI bots with full access — the most welcoming host. HP explicitly allows all major AI crawlers and provides regional sitemaps specifically for AI systems. The Amazon-HP contrast illustrates two fundamentally different strategies: protecting the purchase funnel vs. maximizing cross-channel discovery. (Chapter 4)

GPTBot is the most recognized AI bot (named by 15 companies), followed by ClaudeBot (12) and Google-Extended (11). Only 8 companies distinguish between training crawlers and user-initiated retrieval crawlers — a critical distinction most policies miss. (Chapter 5)

robots.txt declarations do not predict actual access. Goldman Sachs allows GPTBot and ChatGPT-User in robots.txt but blocks at the WAF layer. Alibaba blocks GPTBot in robots.txt but is accessible to ChatGPT-User. Cross-team coordination between content policy (robots.txt) and security infrastructure (WAF) is lacking at most enterprises. (Chapter 6)

Four template strategies emerge from the data: Block All AI (Amazon), Allow All AI (HP), Allow Search but Block Training (nuanced), and Selective Per-Model (partnership-driven). Each serves a different business objective and requires consistent WAF configuration to be effective. (Chapter 7)

About This Study

This report was produced by PROGEOLAB as part of the Fortune Global 500 AI Accessibility audit series. robots.txt files were collected as raw HTTP response bodies during the Chrome UA probe run (April 18, 2026), parsed for User-agent directives, and cross-referenced against four-way user agent access testing. The complete dataset covers 267 parseable robots.txt files containing 518 distinct User-agent names across approximately 15,000 individual rules.

Chapter 2: Methodology

Data Collection

robots.txt files were collected as part of the Fortune Global 500 AI Accessibility audit (April 2026). Each company's `/robots.txt` endpoint was probed with the Chrome User-Agent string. The raw response body was saved for every successful request.

Parsing and Classification

Of 500 companies probed, 267 returned parseable robots.txt files. Files were excluded when they contained HTML (indicating a soft-404 or WAF challenge page), when the response body was less than 10 bytes, or when the file did not contain any `User-agent:` directive.

Each parseable file was parsed into User-agent sections. A section consists of a `User-agent:` directive followed by one or more `Allow:`, `Disallow:`, `Crawl-delay:`, or `Sitemap:` directives. Sections were classified by:

- **User-agent name** — compared against a directory of 19 known AI-specific crawler names
- **Action** — Blocked (`Disallow: /` without any `Allow:`), Allowed (explicit `Allow: /` or `Disallow:` empty), or Partial (`Disallow: /` with specific `Allow:` exceptions)
- **Wildcard policy** — the rules associated with `User-agent: *`, classified as allow-all, block-all, selective, or absent

Cross-Reference with Access Testing

The 20 companies with AI bot policies were cross-referenced against the four-way user agent access results from the broader audit. This comparison reveals whether robots.txt declarations match actual WAF behavior — a critical validation that no other published robots.txt study has performed.

Limitations

Our parser handles standard robots.txt syntax but may not correctly interpret non-standard extensions (e.g., `Crawl-delay`, `Request-rate`). Some companies serve different robots.txt content based on geographic location or requesting IP — our data represents the view from a single European datacenter. Companies using JavaScript-rendered robots.txt (extremely rare) would not be captured.

> ***Source data:*** *Raw robots.txt bodies from* `extracted_data.json` *(Chrome UA run, April 18, 2026). 267 parseable files out of 500 probed. AI bot directory: 19 known user agent names compiled from OpenAI, Anthropic, Google, Perplexity, Common Crawl, Meta, ByteDance, Apple, Amazon, Cohere, Diffbot, and Allen Institute documentation.*

CHAPTER 3

The Headline — Only 20 Companies Mention AI Bots

Of 267 parseable robots.txt files in the Fortune Global 500 dataset, only 20 (7.5%) contain a User-agent directive that names any AI-specific crawler. The remaining 247 companies rely entirely on their wildcard (`User-agent: *`) rules to govern all automated access — rules that predate the AI crawler era and were designed for search engine management.

This means that 92.5% of Fortune 500 companies have no explicit, documented policy for how AI systems should interact with their website. Their AI crawler access is determined by default WAF behavior and generic bot rules, not by deliberate decisions.

The 20 Companies with AI Bot Policies

Company	Domain	AI Bots Named	Action	Total UA Sections
NVIDIA	nvidia.com	17	Allow all	60
Amazon	amazon.com	16	Block all	48
SAP	sap.com	13	Allow all	53
EnBW	enbw.com	11	Allow all	18
Seatrade Maritime	seatrade-maritime.com	11	Allow all	24
HP	hp.com	10	Allow all	14
WSJ (News Corp)	wsj.com	10	Block all	24
Baking Business (Smucker)	bakingbusiness.com	9	Block all	10
BNY Mellon	bnymellon.com	8	Allow all	21
Meta	meta.com	8	Allow all	18
LA Times	latimes.com	5	Block all	13
Samsung	samsung.com	5	Allow all	12
Capital One	capitalone.com	3	Allow all	37
Cisco	cisco.com	3	Allow all	10
Alibaba	alibaba.com	2	Block all	6

Company	Domain	AI Bots Named	Action	Total UA Sections
Goldman Sachs	goldmansachs.com	2	Allow all	3
Dell	dell.com	1	Allow (Bytespider only)	9
Novartis	novartis.com	1	Allow (GPTBot only)	4
Airbus	airbus.com	1	Allow (Bytespider only)	139
LG Electronics	lge.co.kr	1	Allow (meta-externalagent only)	6

Block vs Allow

Of the 20 companies with AI bot policies:

- **14 companies allow** AI crawlers (NVIDIA, SAP, EnBW, Seatrade, HP, BNY Mellon, Meta, Samsung, Capital One, Cisco, Goldman Sachs, Dell, Novartis, Airbus, LG)
- **5 companies block** AI crawlers (Amazon, WSJ, Baking Business, LA Times, Alibaba)
- **1 company partially blocks** (not observed in current data)

The 14:5 ratio favoring allow over block is notable. Among the small number of companies that have made a deliberate decision about AI access, the majority choose openness. The blockers are concentrated in two sectors: media/publishing (WSJ, LA Times, Baking Business) and e-commerce (Amazon, Alibaba).

The Wildcard Problem

The remaining 247 companies govern all bot access through wildcard rules. These rules were designed for an era when the primary automated visitors were search engine crawlers (Googlebot, Bingbot, Yandex). They cannot distinguish between a search engine indexing content for web search, an AI system ingesting content for training, and an AI system fetching content to answer a user's real-time question.

Our analysis of the 267 wildcard policies:

Wildcard Policy	Companies	Meaning
Selective rules	195 (73%)	Allows some paths, blocks others (e.g., `/admin/`, `/cgi-bin/`, `/search?`)
Allow all	50 (19%)	No meaningful restrictions (`Disallow:` or no Disallow at all)
No wildcard section	18 (7%)	No `User-agent: *` entry (relies on named UA sections only)
Block all	4 (1.5%)	`Disallow: /` for all bots — blocks everything

The 195 companies with selective wildcard rules present a nuanced situation. Their rules block specific paths (admin panels, search result pages, user accounts) while allowing the rest of the site. AI crawlers are treated the same as any other unnamed bot — they can access public content but are blocked from

internal-facing paths. This is reasonable for search engines but may be insufficient for AI-specific concerns like training data or content attribution.

Source data: *robots.txt files from the Chrome UA run (`2026-04-18_chrome_ua`), stored in `extracted_data.json`. Files classified as parseable when they contain at least 10 bytes of non-HTML text content. AI bot identification based on a directory of 19 known AI-specific user agent names (GPTBot, ClaudeBot, PerplexityBot, Google-Extended, etc.). Wildcard classification based on the rules associated with the `User-agent: *` section.*

CHAPTER 4

Amazon vs HP — Two Extremes of AI Bot Strategy

The Fortune 500 robots.txt landscape has two pole positions occupied by two very different companies. Amazon names and blocks 16 AI bots — the most comprehensive AI blocker in the dataset. HP names and allows 10 AI bots with full access — the most welcoming enterprise to AI crawlers. Their approaches represent the two ends of the strategic spectrum.

Amazon: The Comprehensive Blocker

Amazon's robots.txt at amazon.com names 16 distinct AI-specific user agents, every one with `Disallow: /` — a complete block on all content. The file contains 48 total User-agent sections across 5,888 bytes.

The AI bots Amazon blocks by name:

#	Bot Name	Operator	Purpose
1	GPTBot	OpenAI	Training + retrieval
2	ChatGPT-User	OpenAI	User-initiated browsing
3	OAI-SearchBot	OpenAI	Search crawling
4	ClaudeBot	Anthropic	Training + indexing
5	Claude-User	Anthropic	User-initiated browsing
6	Claude-SearchBot	Anthropic	Search crawling
7	Claudebot	Anthropic	Alternate casing
8	PerplexityBot	Perplexity	Search + retrieval
9	Perplexity-User	Perplexity	User-initiated browsing
10	Google-Extended	Google/DeepMind	AI training
11	CCBot	Common Crawl	Training corpus
12	Bytespider	ByteDance	Training
13	meta-externalagent	Meta	Training
14	Diffbot	Diffbot	Knowledge graph extraction
15	Cohere-ai	Cohere	Training
16	AI2Bot	Allen Institute	Research training

Amazon's thoroughness is remarkable. They block not just the primary crawlers (GPTBot, ClaudeBot) but also the secondary variants (Claude-User, Claude-SearchBot, Perplexity-User) and the research-oriented crawlers (AI2Bot, CCBot, Diffbot). This is the most complete AI bot directory we found in any Fortune 500 robots.txt file.

The strategic implication is clear: Amazon does not want any AI system to access, cache, or redistribute amazon.com content. This includes user-initiated browsing — blocking ChatGPT-User and Claude-User means that when a customer asks an AI to look up a product on Amazon, the AI cannot comply. Amazon's content remains within Amazon's own ecosystem.

HP: The Welcoming Host

HP's robots.txt takes the opposite approach. It names 10 AI-specific user agents, every one with `Allow: /` — full access to all content. The file is 13,467 bytes with 14 User-agent sections.

The AI bots HP explicitly allows:

#	Bot Name	Operator	Rule
1	GPTBot	OpenAI	Allow: /
2	ChatGPT-User	OpenAI	Allow: /
3	OAI-SearchBot	OpenAI	Allow: /
4	ClaudeBot	Anthropic	Allow: /
5	Claude-User	Anthropic	Allow: /
6	Claude-SearchBot	Anthropic	Allow: /
7	Google-Extended	Google/DeepMind	Allow: /
8	PerplexityBot	Perplexity	Allow: /
9	Perplexity-User	Perplexity	Allow: /

HP's strategy is to maximize the availability of its product information, specifications, support documentation, and marketing content across all AI platforms. When a customer asks an AI to compare HP laptops, the AI can access current pricing, specs, and reviews directly from hp.com.

HP goes further than a simple allow — their robots.txt includes extensive sitemap declarations specifically for AI crawlers. The PerplexityBot section alone lists dozens of regional sitemaps covering EMEA, LATAM, and APJ markets, directing the crawler to the most important content for each region.

The Strategic Calculus

The Amazon-HP contrast illustrates the fundamental strategic question every enterprise must answer:

Amazon's calculation: Amazon's competitive advantage is its marketplace. Product information on Amazon drives purchases on Amazon. Letting AI systems extract and redistribute this information (pricing,

reviews, availability) would benefit competitors and reduce the incentive for consumers to visit Amazon directly. Amazon controls the end-to-end customer journey and has no interest in being cited — it wants to be visited.

HP's calculation: HP sells through multiple channels. When a customer asks an AI to recommend a laptop, HP benefits from the AI having access to current, accurate HP product information. HP does not own the purchase journey — consumers buy from Best Buy, Amazon, and HP.com. Being cited by AI systems drives awareness and consideration across all channels.

Neither strategy is objectively correct. They reflect fundamentally different business models and competitive positions. The value of this analysis is not to judge which approach is better, but to demonstrate that these decisions should be intentional — not left to default WAF configurations.

What Most Companies Do Instead

The Amazon-HP contrast highlights the gap with the other 247 companies that have no AI-specific robots.txt policy. These companies have made no deliberate choice. Their AI accessibility is determined by:

1. Their wildcard `User-agent: *` rule (which may allow or restrict)
2. Their WAF configuration (which may block AI UAs regardless of robots.txt)
3. Default settings inherited from their hosting provider or CDN

The recommendation from this analysis is not necessarily to follow Amazon's or HP's model — it is to make any deliberate choice at all.

> ***Source data:*** *robots.txt files from* `extracted_data.json`*, Chrome UA run. Amazon:* `www.amazon.com/robots.txt` *(5,888 bytes, 48 UA sections). HP:* `hp.com/robots.txt` *(13,467 bytes, 14 UA sections). Both files validated as genuine robots.txt content (no HTML, proper User-agent/Disallow syntax).*

CHAPTER 5

The AI Bot Directory — 19 Named Bots Found

Across the 20 companies that mention AI bots in their robots.txt, we identified 19 distinct AI-specific user agent names. This directory represents the most comprehensive catalog of AI crawlers extracted from Fortune 500 robots.txt files.

The Complete AI Bot Directory

Bot Name	Operator	Primary Purpose	Named By (companies)	Blocked	Allowed
GPTBot	OpenAI	Training + retrieval	15	4	11
ClaudeBot	Anthropic	Training + indexing	12	4	8
Claudebot	Anthropic	Alternate casing	12	4	8
ChatGPT-User	OpenAI	User-initiated browsing	8	1	7
OAI-SearchBot	OpenAI	Search crawling	8	1	7
PerplexityBot	Perplexity	Search + retrieval	9	2	7
Google-Extended	Google/DeepMind	AI training	11	4	7
Perplexity-User	Perplexity	User browsing	5	1	4
Claude-User	Anthropic	User browsing	5	1	4
Claude-SearchBot	Anthropic	Search crawling	4	1	3
CCBot	Common Crawl	Training corpus	8	3	5
Amazonbot	Amazon	Product search	7	1	6
Bytespider	ByteDance	Training	8	4	4
meta-externalagent	Meta	Training	7	3	4
Applebot-Extended	Apple	AI training	5	2	3
FacebookBot	Meta	Social crawling	4	1	3
Diffbot	Diffbot	Knowledge graph	3	2	1
Cohere-ai	Cohere	Training	4	2	2
AI2Bot	Allen Institute	Research training	2	1	1

Notable Patterns

GPTBot is the most recognized AI bot. Fifteen companies name it — more than any other AI crawler. This reflects OpenAI's market visibility and the early publication of GPTBot's user agent string.

The casing problem. Both `ClaudeBot` and `Claudebot` appear in the data, each named by 12 companies. robots.txt user agent matching is case-insensitive per the specification, but this inconsistency in how companies reference Anthropic's crawler suggests copy-paste from different sources.

Training vs retrieval confusion. Only 8 companies distinguish between training crawlers (GPTBot) and retrieval crawlers (ChatGPT-User, OAI-SearchBot). The remaining 12 companies that name OpenAI's crawlers only name GPTBot — blocking training but leaving the retrieval path unaddressed. This matters because blocking GPTBot alone does not prevent ChatGPT from browsing your site when a user asks it to.

Bytespider is the most polarizing. ByteDance's crawler is blocked by 4 companies and allowed by 4 — an even split. This reflects uncertainty about ByteDance's data practices and the geopolitical dimension of AI training data.

Chapter 6: Declaration vs Reality — When robots.txt Lies

A critical finding of this study is that robots.txt declarations do not predict actual AI crawler access. Companies can declare `Allow: /` for GPTBot in their robots.txt while their WAF blocks the same crawler at the network layer. Conversely, companies with restrictive robots.txt may be fully accessible because AI crawlers do not always comply.

The Gap Between Policy and Practice

Cross-referencing the 20 companies with AI bot policies against our 4-UA access testing reveals significant inconsistencies:

Company	robots.txt Policy	Actual ChatGPT-User Access	Match?
HP	Allow all 10 AI bots	Accessible (10/64 probes)	✓
Meta	Allow 8 AI bots	Accessible (13/64)	✓
Samsung	Allow 5 AI bots	Accessible (10/64)	✓
Goldman Sachs	Allow GPTBot + ChatGPT-User	Blocked (0/64)	✗
NVIDIA	Allow all 17 AI bots	Accessible (10/64)	✓
Amazon	Block all 16 AI bots	Blocked (0/64)	✓
Cisco	Allow 3 AI bots	Blocked (0/64)	✗
Alibaba	Block 2 AI bots	Accessible (3/64)	✗

The mismatches are significant:

Goldman Sachs allows GPTBot and ChatGPT-User in its robots.txt but blocks ChatGPT-User at the WAF layer (0/64 probes succeed). The declared policy says "come in" but the infrastructure says "access denied." This creates a false promise — AI companies see the permissive robots.txt and expect access, but the WAF blocks them regardless.

Cisco allows GPTBot, Google-Extended, and CCBot in its robots.txt but blocks ChatGPT-User at the WAF layer. The policy intention and the technical reality are misaligned.

Alibaba blocks GPTBot and Google-Extended in its robots.txt but is actually accessible to ChatGPT-User (3/64 probes succeed). The opposite mismatch — the declared policy is restrictive but the enforcement is permissive.

Why This Happens

The robots.txt file and the WAF operate at different layers and are managed by different teams:

- **robots.txt** is a text file managed by webmasters or marketing teams. It expresses the organization's content access policy.
- **The WAF** is infrastructure managed by security or DevOps teams. It enforces bot management rules based on technical signals (UA string, IP, TLS fingerprint).

These two systems are rarely synchronized. A marketing team may add `User-agent: GPTBot / Allow: /` to robots.txt without informing the security team, whose WAF continues to block all non-browser user agents. Conversely, a security team may add AI bot blocking to the WAF without checking whether the content team has expressed a different policy in robots.txt.

Implications

Any AI visibility assessment that relies solely on robots.txt analysis will produce inaccurate results. A permissive robots.txt does not guarantee access, and a restrictive robots.txt does not guarantee blocking. Both signals must be tested empirically — which is what our 4-UA comparison methodology achieves.

For enterprises, this finding underscores the need for cross-team coordination between content/marketing (who manage robots.txt) and security/DevOps (who manage WAF rules). A coherent AI access policy requires both layers to express the same intent.

Chapter 7: Four Template Strategies

Based on the patterns observed across 267 Fortune 500 robots.txt files and 20 companies with explicit AI bot policies, we identify four archetypal strategies for managing AI crawler access. Each serves a different business objective.

Strategy 1: Block All AI (The Amazon Model)

Block every known AI crawler from accessing any content. Used when the competitive value of your content exceeds the value of AI citation.

``` User-agent: GPTBot Disallow: /
```

User-agent: ChatGPT-User Disallow: /

User-agent: ClaudeBot Disallow: /

User-agent: PerplexityBot Disallow: /

User-agent: Google-Extended Disallow: /

User-agent: Bytespider Disallow: /

User-agent: CCBot Disallow: /

User-agent: meta-externalagent Disallow: / ```

Best for: E-commerce platforms, content publishers with paywalls, companies with proprietary pricing data.

Strategy 2: Allow All AI (The HP Model)

Explicitly welcome all AI crawlers with full access. Used when AI citation drives customer awareness across multiple channels.

``` User-agent: GPTBot Allow: /

User-agent: ChatGPT-User Allow: /

User-agent: ClaudeBot Allow: /

User-agent: PerplexityBot Allow: /

User-agent: Google-Extended Allow: / ```

**Best for:** B2B companies, technology vendors, companies with multi-channel distribution.

### Strategy 3: Allow Search, Block Training

The nuanced approach: allow AI systems to retrieve content for user queries (ChatGPT-User, OAI-SearchBot) while blocking training crawlers (GPTBot, Google-Extended, CCBot). This preserves AI visibility while withholding content from model training.

``` User-agent: GPTBot Disallow: /

User-agent: Google-Extended Disallow: /

User-agent: CCBot Disallow: /

User-agent: Bytespider Disallow: /

User-agent: ChatGPT-User Allow: /

User-agent: OAI-SearchBot Allow: /

User-agent: PerplexityBot Allow: /

User-agent: ClaudeBot Allow: / ```

**Best for:** Companies concerned about AI training data usage but wanting to remain visible in AI-generated answers.

### Strategy 4: Selective Per-Model

Allow specific AI systems while blocking others. Used when a company has partnerships or preferences for particular AI platforms.

``` User-agent: GPTBot Allow: /

User-agent: ClaudeBot Allow: /

User-agent: Bytespider Disallow: /

User-agent: CCBot Disallow: /

User-agent: meta-externalagent Disallow: / ```

**Best for:** Companies with strategic AI partnerships, or those with specific concerns about particular operators.

### Critical Caveat

All four strategies assume that the WAF is configured consistently with the robots.txt policy. As Chapter 6 demonstrates, a permissive robots.txt is ineffective if the WAF blocks the same crawlers at the network layer. Any robots.txt change should be accompanied by a corresponding WAF rule review.

> ***Source data:*** *Template strategies derived from observed patterns across 20 Fortune 500 companies with explicit AI bot policies. Bot names from the AI bot directory (Chapter 5). Strategy classifications based on the dominant pattern in each company's robots.txt rules.*
```

About PROGEOLAB

PROGEOLAB is an AI-native visibility intelligence platform.

This report is part of the PROGEOLAB Fortune 500 AI Accessibility Audit — a series of research studies on how large enterprises appear to (or disappear from) AI answer engines. All measurements are from live HTTP probes across four user agents: a research bot, Googlebot, Chrome, and ChatGPT-User. No estimates, no third-party data sources.

Methodology in brief

500 companies · 67 probes each · 4 user agents · 134,000 probe requests. Data collected April 16–19, 2026. Response bodies stored and re-validated with MD5-hash soft-404 detection to eliminate the ~25× inflation that status-code-only scans produce.

Contact & next steps

Visit **progeolab.ai/research** or request a demo for a complimentary AI visibility analysis of your organisation.

For press enquiries, data requests, or syndication, write to **research@progeolab.ai**.

PROGEOLAB