PROGEOLAB

PROGEOLAB RESEARCH

The GEO Visibility Gap

Fortune 500 AI Accessibility Report · April 2026

PROGEOLAB
April 2026
progeolab.ai/research

Contents

CHAPTER 1

Executive Summary

This report presents the first comprehensive measurement of AI crawler accessibility across the Fortune Global 500. By probing 500 of the world's largest companies with four different user agents — a research bot, Googlebot, a Chrome browser, and OpenAI's ChatGPT-User — we quantify the gap between what human visitors can access and what AI answer engines can read.

Key Findings

53 companies are visible to browsers but invisible to ChatGPT. These companies serve content to Chrome users but actively block ChatGPT-User, the user agent sent when someone asks ChatGPT to browse a URL. When users ask AI about these companies, the model cannot access current information. The GEO visibility gap spans Johnson & Johnson, AstraZeneca, Lockheed Martin, Oracle, IBM, AT&T, Toyota, Goldman Sachs, Salesforce, and 44 others. (Chapter 4)

AI-identified user agents perform worst. Chrome UA reached 352 companies (70.4%), the research bot reached 325 (65.0%), Googlebot reached 307 (61.4%), and ChatGPT-User reached only 300 (60.0%). ChatGPT-User receives 986 more HTTP 403 responses per run than Chrome — the quantifiable cost of identifying as an AI crawler. (Chapter 3)

Blocking operates at three distinct layers. Layer 1: 53 companies block based on the User-Agent string. Layer 2: 24 companies block all traffic from datacenter IP ranges, regardless of UA. Layer 3: 15 companies use TLS fingerprinting to reject non-browser HTTP clients before the request is even read. No single countermeasure addresses all three. (Chapter 5)

Impersonating Googlebot backfires. Counterintuitively, claiming to be Googlebot from a non-Google IP produces worse results (307 reachable) than using an honest unknown bot (325). Thirty-two companies verify Googlebot identity via reverse DNS and penalize impersonation attempts more aggressively than they penalize unknown bots. (Chapter 6)

F5 BIG-IP dominates the WAF landscape. Raw response body analysis reveals F5 BIG-IP on 232 Fortune 500 sites — more than Cloudflare (64), Akamai (59), and Imperva (23) combined. WAF vendor choice directly influences AI crawler treatment, with Cloudflare's TLS fingerprinting being the most difficult to bypass. (Chapter 7)

Status codes overstate adoption by 25×. Of 353 companies returning HTTP 200 on /llms.txt, only 14 have a real llms.txt file. The rest are soft-404 pages — HTML error pages, homepage redirects, and catch-all templates served with a 200 status. True llms.txt adoption among the Fortune 500 is 2.8%, not the 71% that status-code-only scans would suggest. One hundred sixty companies (41% of responding sites) serve catch-all pages on any URL path. (Chapter 8)

Telecommunications and software lead in AI-specific blocking. Twenty-seven percent of telecom companies and 50% of software companies (small sample) in the GEO gap block ChatGPT specifically. Metals, trading, and energy sectors show zero AI-specific blocking. (Chapter 9)

The United States accounts for 43% of the GEO gap. Twenty-three US companies block ChatGPT-User while serving Chrome — a 16.5% rate versus the 10.6% global average. China's 128 companies are

predominantly unreachable due to network infrastructure (Great Firewall), not AI-specific policy. (Chapter 10)

Implications

For enterprises, the most impactful action is auditing WAF bot management rules for AI-specific user agents. The distinction between AI training crawlers (GPTBot) and user-initiated retrieval (ChatGPT-User) should inform policy — blocking one need not mean blocking both. (Chapter 11)

For AI companies, structural barriers persist beyond policy: IP reputation scoring and TLS fingerprinting block AI crawlers independently of robots.txt compliance. Implementing reverse DNS verification (as Google has done for Googlebot) and working with WAF vendors on native AI crawler verification would reduce the friction that drives blanket blocking. (Chapter 12)

About This Study

This report was produced by PROGEOLAB, an AI visibility intelligence platform. The study probed 500 companies with 67 endpoints across 4 user agents — 134,000 individual requests — between April 16 and 19, 2026. Raw response bodies (approximately 8 GB across all runs) were analyzed for WAF signatures, soft-404 detection, and content validation. The methodology is documented in Chapter 2, and full company-level data is available in the appendix.

PROGEOLAB plans to repeat this audit quarterly, tracking changes in AI accessibility, llms.txt adoption, and robots.txt AI bot policy across the Fortune Global 500.

CHAPTER 2

Methodology

Scope

This study audits the AI accessibility posture of all 500 companies on the 2024 Fortune Global 500 list. The company list was sourced from us500.com, which publishes the Fortune Global 500 rankings with company names, countries, industries, and financial data. Domain resolution — mapping each company name to its primary corporate website — was performed through a four-source cascade (Wikidata, Wikipedia, SEC EDGAR, manual curation), producing a verified list of 500 unique domains.

Probing Framework

A custom Python-based audit tool was developed for this study. The tool executes 67 probes per company, of which 64 are HTTP requests and 3 are network-level collectors:

- **Tier 1 (6 probes):** Core SEO — robots.txt, sitemap.xml (3 path variants), homepage, favicon
- **Tier 2 (4 probes):** AI-specific files — llms.txt, llms-full.txt, ai.txt (2 path variants)
- **Tier 3 (21 probes):** Agent readiness and .well-known — agents.json (4 variants), MCP endpoints (3 variants), security.txt (2 paths), well-known discovery files (6 paths), ads.txt (3 paths), humans.txt, webmaster verification (2 paths)
- **Tier 4 (33 probes):** Advanced signals — RSS/Atom feeds (7 paths), OpenAPI specs (7 paths), PWA manifest (3 paths), status pages (2 paths), content pages (14 paths including /about, /careers, /sustainability, /investors), DNS records (6 types), certificate transparency logs, TLS certificate inspection
- **Network (3 probes):** DNS (A, AAAA, MX, TXT, CAA, NS plus _dmarc and _bimi), certificate transparency via crt.sh, TLS certificate details via socket connection

Four User-Agent Runs

The complete probe set was executed four times, each with a different User-Agent string, against the same 500 domains:

Run	User-Agent	Date	Purpose
Baseline	`F500-Web-Maturity-Audit/1.0 (research)`	April 16–17, 2026	Honest unknown bot baseline
Googlebot	`Googlebot/2.1`	April 17–18, 2026	Search engine crawler test
Chrome	`Chrome/135.0.0.0` (full string)	April 18, 2026	Browser baseline (upper bound)
ChatGPT-User	`ChatGPT-User/1.0`	April 18–19, 2026	AI answer engine test

Total probes: 500 companies × 67 probes × 4 runs = **134,000 individual probe requests**.

Infrastructure

All probes were executed from a single datacenter environment using httpx (Python HTTP/2 client) with per-host connection limiting (3 concurrent connections per domain), 15-second timeouts, and HTTP/2 protocol preference. No JavaScript rendering was performed — all probes are raw HTTP requests, which is the same mechanism used by GPTBot, ClaudeBot, and PerplexityBot in production.

Raw Body Collection

In addition to HTTP status codes, the full response body was saved for every probe that returned content. This produced approximately 2 GB of raw data per UA run (~8 GB total), enabling:

- **Soft-404 detection:** Body fingerprinting across multiple paths to identify catch-all pages returning 200 status with error content
- **WAF vendor attribution:** Signature matching in response HTML, JavaScript, and headers
- **llms.txt validation:** Content-type analysis distinguishing real Markdown files from HTML error pages
- **JSON-LD extraction:** Parsing structured data from homepage HTML

Classification Methodology

Reachability: A company is classified as "reachable" for a given UA if at least one of its 64 HTTP probes returned an HTTP 200 response.

GEO gap: A company falls in the GEO visibility gap when Chrome UA returns ≥1 HTTP 200 AND ChatGPT-User UA returns 0 HTTP 200 responses.

Three-layer blocking: Layer 1 (UA string) = Chrome works, ChatGPT blocked. Layer 2 (IP reputation) = all 4 UAs return 403. Layer 3 (TLS fingerprint) = StreamReset errors in HTTP/2 framing.

Soft-404 classification: A site is flagged as soft-404 when ≥3 of 7 fingerprint paths return identical body MD5 hashes. These paths (agents.json, agents.txt, mcp.json, mcp-server, llms.txt, llms-full.txt, changelog) serve different purposes and should never contain identical content.

Limitations

Single vantage point: All probes originated from one datacenter location. Results may differ from other geographic locations, particularly for sites with geo-specific CDN configurations.

No JavaScript rendering: Sites that rely on client-side rendering (SPAs) may serve minimal or empty HTML to our probes. This is a realistic representation of what AI crawlers see, since GPTBot and ClaudeBot also do not execute JavaScript.

Point-in-time snapshot: Data was collected over April 16–19, 2026. WAF configurations, robots.txt files, and website content change over time. The quarterly replication schedule (Chapter 20) will track these changes.

Domain resolution quality: Despite manual curation, some domain mappings may point to regional subsidiaries or investor relations subdomains rather than the primary corporate website. These cases were minimized through validation but may affect individual company results.

> ***Reproducibility:*** *The audit tool source code, domain resolution CSV, configuration files, and analysis scripts are documented in the project repository. The raw SQLite databases (4 × ~50 MB) and extracted body analysis (246 MB JSON) are available for independent verification.*

CHAPTER 3

The Four-Way Reachability Matrix

To measure how AI crawler access differs from traditional web access, we probed all 500 companies with four distinct User-Agent strings, each representing a different actor in the web ecosystem. The same 64 HTTP endpoints were tested per company, per user agent — a total of 128,000 HTTP probe requests across all four runs.

The Four User Agents

Each run used an identical probe set (67 probes per company, of which 64 are HTTP requests and 3 are network-level collectors). Only the User-Agent header differed:

Label	User-Agent String	What It Represents
Research UA	`F500-Web-Maturity-Audit/1.0 (research; contact@progeolab.ai)`	An honest, unknown bot identifying itself and its purpose. The baseline for how sites treat unrecognized crawlers.
Googlebot	`Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)`	Google's search crawler. Tests whether sites grant preferential access to search engines — and whether they verify Googlebot by reverse DNS.
Chrome	`Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/135.0.0.0 Safari/537.36`	A standard desktop browser. Represents what a human visitor sees. The upper bound of accessibility for any non-JavaScript-dependent probe.
ChatGPT-User	`Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko); compatible; ChatGPT-User/1.0; +https://openai.com/bot`	The user agent sent when a ChatGPT user asks the model to browse a URL. This is the GEO-critical fetch — it directly determines whether a company's content appears in AI-generated answers.

The Headline Numbers

The results reveal that AI-identified user agents receive materially worse treatment than browser user agents — and that impersonating a search engine crawler can backfire.

User Agent	Reachable (≥1 HTTP 200)	WAF-Blocked (all 403)	Unreachable (connection errors)	Total Successful Probes
Research UA	325 / 500 (65.0%)	51	124	4,382
Googlebot	307 / 500 (61.4%)	60	133	4,460
Chrome	**352 / 500 (70.4%)**	**32**	116	**4,688**

User Agent	Reachable (≥1 HTTP 200)	WAF-Blocked (all 403)	Unreachable (connection errors)	Total Successful Probes
ChatGPT-User	300 / 500 (60.0%)	64	136	4,047

Three patterns emerge from this data:

Chrome outperforms all other user agents. 352 companies returned at least one successful response to a Chrome-identified request, compared to 325 for our research UA, 307 for Googlebot, and only 300 for ChatGPT-User. WAF blocks dropped from 51 (research) to just 32 (Chrome), confirming that many enterprise WAFs are configured to pass browser-like traffic while challenging or blocking bot-like traffic regardless of the bot's stated purpose.

ChatGPT-User performs worst overall. Despite being a legitimate, documented user agent from OpenAI, ChatGPT-User reached fewer companies (300) than even our obscure research bot (325). This is not a side effect of general bot blocking — it is evidence that companies are actively identifying and rejecting AI-specific user agents. The 64 companies returning 403 to ChatGPT-User (vs. 32 for Chrome) represent explicit AI-targeted blocking.

Googlebot impersonation backfires. Counterintuitively, presenting as Googlebot produced worse results (307 reachable) than our honest research UA (325). As detailed in Chapter 6, this occurs because 32 companies verify Googlebot identity via reverse DNS lookup. When the request originates from a non-Google IP range, these sites flag it as impersonation and block more aggressively than they would block an unknown bot.

Response Distribution

Beyond the binary reachable/unreachable classification, the distribution of HTTP response codes reveals the mechanisms at work:

Status Code	Research UA	Googlebot	Chrome	ChatGPT-User
200 (OK)	4,382	4,460	4,688	4,047
403 (Forbidden)	5,026	5,131	4,412	5,398
404 (Not Found)	13,560	12,583	14,220	12,714
Connection Error	5,351	5,417	3,781	5,289
Timeout	1,586	1,843	1,829	1,764
Other	2,095	2,066	1,570	2,288

The 404 responses are the most numerous across all runs — this is expected and healthy, as most companies do not host files like `/llms.txt`, `/.well-known/agents.json`, or `/swagger.json`. A clean 404 is informative data, not a failure.

The 403 responses are the diagnostic signal. ChatGPT-User receives 5,398 forbidden responses — 986 more than Chrome (4,412). Those ~1,000 additional 403s are the quantifiable cost of identifying as an AI

crawler rather than a browser.

Reachability by Bucket

Every company falls into one of four categories based on its behavior across the four user agents:

Category	Count	Description
Broadly accessible	293	Returns content to all four user agents. No AI-specific blocking detected.
Selectively blocking	59	Accessible to some user agents but not others. Includes the 53-company GEO gap (Chrome ✓, ChatGPT ✗) detailed in Chapter 4.
WAF-hardened	29	Returns 403 to all four user agents including Chrome. Deep bot detection beyond UA string — likely IP reputation, TLS fingerprinting, or JavaScript challenges.
Unreachable	109	Connection/timeout errors across all four user agents. Primarily Chinese state-owned enterprises behind the Great Firewall (~70), wrong or expired domains (~15), and sites with aggressive geo-blocking (~24).
Mixed errors	10	Inconsistent behavior — some UAs get errors, others get 403s, none get 200s.

The 293 broadly accessible companies represent the corpus where meaningful GEO analysis can occur — these sites serve content to AI crawlers and are therefore candidates for structured data analysis, llms.txt adoption, and robots.txt AI bot policy evaluation described in subsequent chapters. The 53 selectively blocking companies are the core of the GEO visibility gap explored in Chapter 4.

> ***Source data:*** *Four SQLite databases containing 33,500 probe results each. Research UA:* `2026-04-16_baseline` *run. Googlebot:* `2026-04-17_googlebot_ua` *run. Chrome:* `2026-04-18_chrome_ua` *run. ChatGPT-User:* `2026-04-18_chatgpt_user_ua` *run. Cross-database comparison performed on HTTP probes only (excluding DNS, CT log, and TLS certificate probes which are UA-independent). Raw response bodies stored in* `raw/` *subdirectories per run, per domain, per probe.*

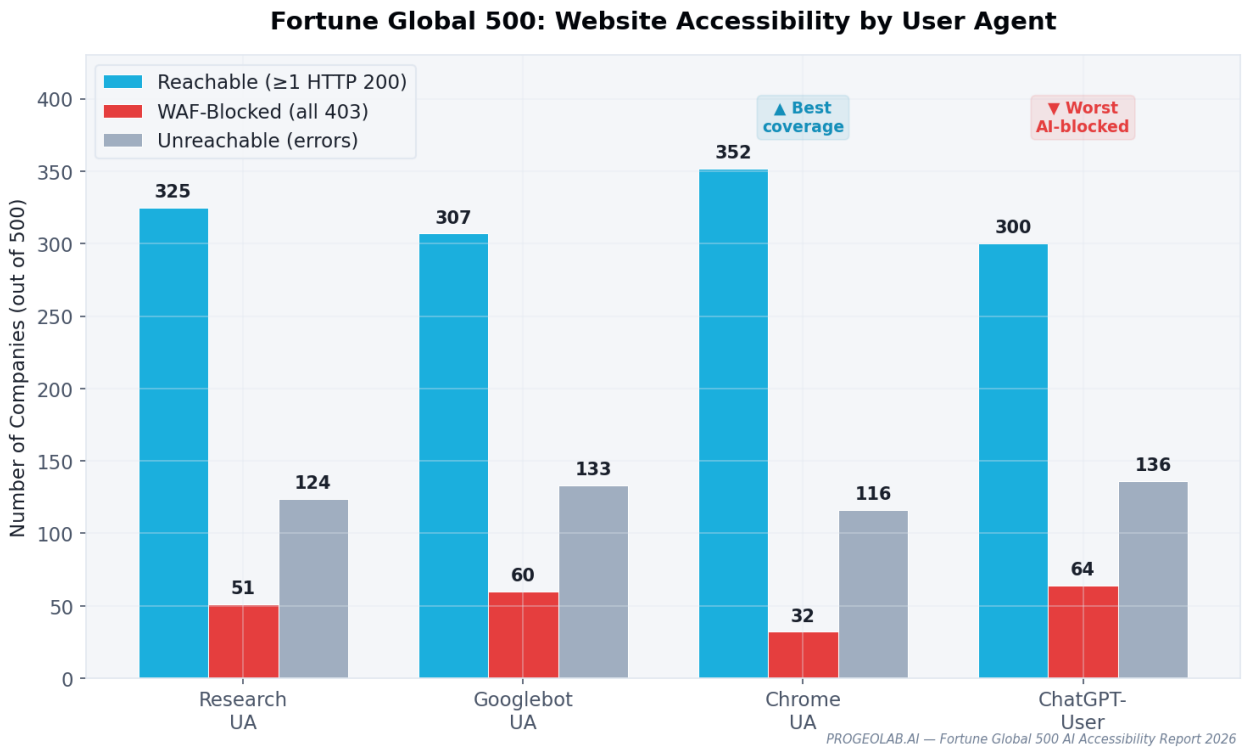

Figure 3.1 · Reachability bars

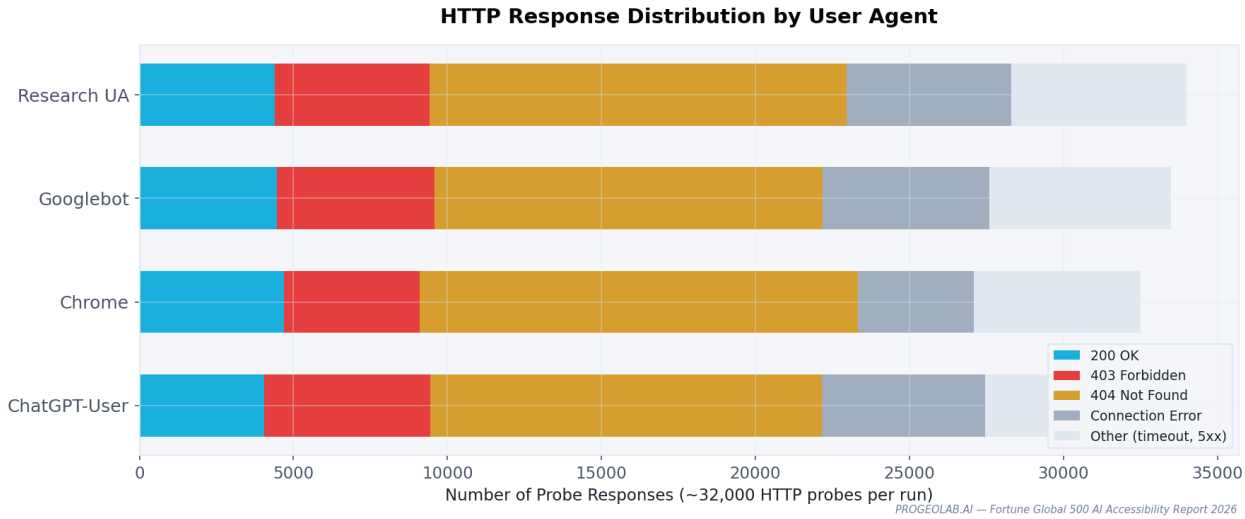

Figure 3.2 · Response distribution

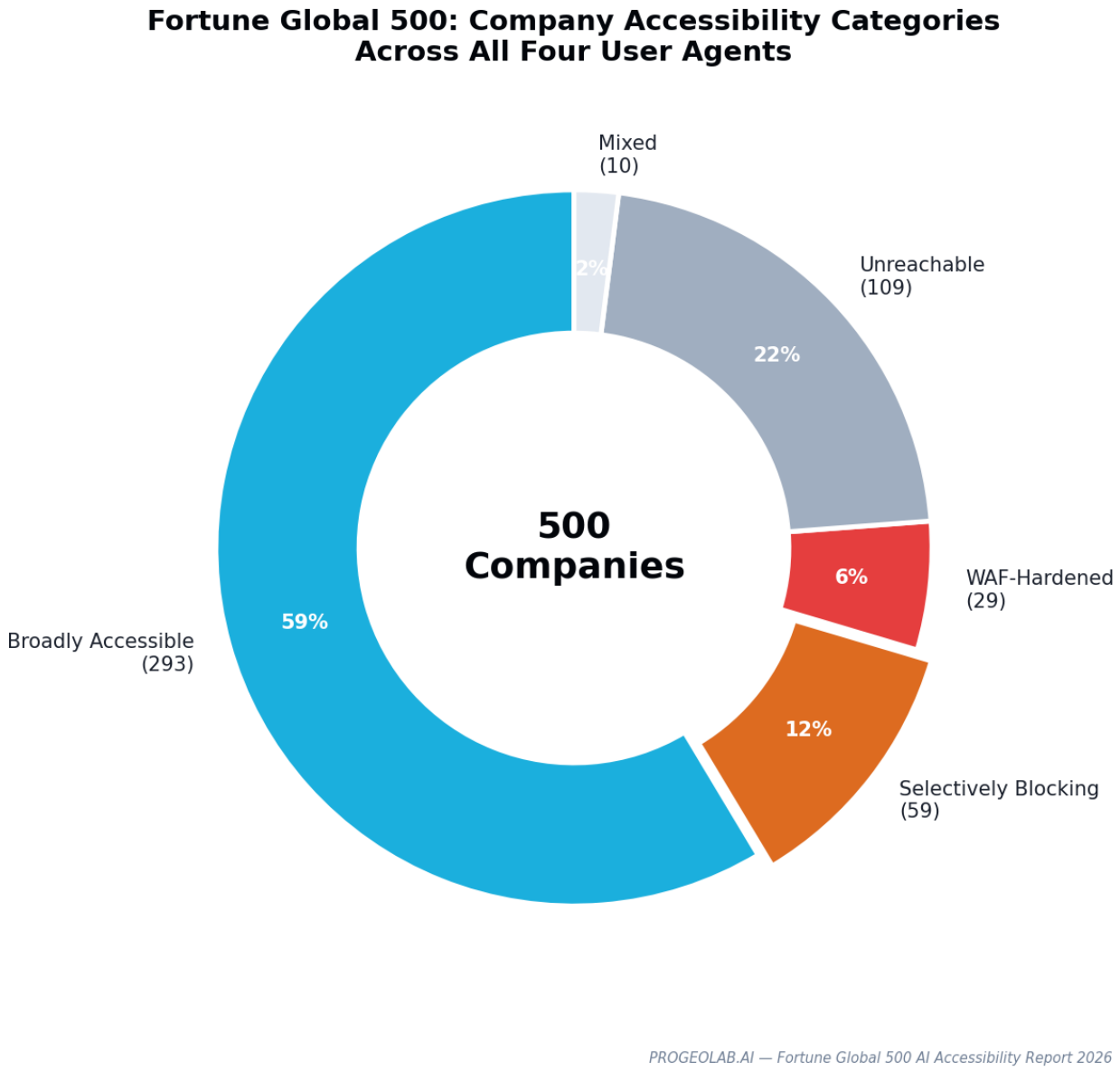

Figure 3.3 · Company buckets

CHAPTER 4

The GEO Visibility Gap — 53 Companies Invisible to AI

The most consequential finding of this study is not about the companies that block all bots, nor the ones behind the Great Firewall. It is about the 53 Fortune Global 500 companies that serve content to web browsers but actively refuse the same content to ChatGPT.

These companies are visible to human visitors. They are visible to Google Search. But when a user asks ChatGPT to look up their latest earnings, their sustainability report, or their product specifications, ChatGPT cannot access the page. The company's own content is replaced by whatever the model recalls from training data — which may be months or years out of date.

This is the GEO Visibility Gap: the measurable difference between what a browser can access and what an AI answer engine can access on the same website.

Defining the Gap

A company falls into the GEO gap when it meets two criteria simultaneously:

1. **Chrome UA returns at least one HTTP 200 response** across 64 HTTP probes — confirming the site serves content to browsers.
2. **ChatGPT-User UA returns zero HTTP 200 responses** across the same 64 probes — confirming the site blocks AI-initiated browsing.

Of 500 companies audited, 53 meet both criteria. They represent $5.7 trillion in combined annual revenue and span 33 industries across 14 countries.

The 53 Companies

Company	Country	Industry	Chrome 2xx	ChatGPT 2xx	Gap
China Telecommunications	China	Telecommunications	64	0	64
Johnson & Johnson	U.S.	Pharmaceuticals	64	0	64
AstraZeneca	Britain	Pharmaceuticals	45	0	45
Alimentation Couche-Tard	Canada	Specialty Retailers	44	0	44
Lockheed Martin	U.S.	Aerospace & Defense	29	0	29
Volvo	Sweden	Motor Vehicles & Parts	18	0	18
Maersk Group	Denmark	Shipping	14	0	14

Company	Country	Industry	Chrome 2xx	ChatGPT 2xx	Gap
3M	U.S.	Chemicals	13	0	13
Oracle	U.S.	Computer Software	13	0	13
Salesforce	U.S.	Computer Software	13	0	13
Chevron	U.S.	Petroleum Refining	12	0	12
IBM	U.S.	Information Technology Services	12	0	12
AT&T	U.S.	Telecommunications	11	0	11
Toyota Motor	Japan	Motor Vehicles & Parts	11	0	11
UniCredit Group	Italy	Banks: Commercial and Savings	11	0	11
BP	Britain	Petroleum Refining	10	0	10
Comcast	U.S.	Telecommunications	10	0	10
ABB	Switzerland	Industrial Machinery	9	0	9
Humana	U.S.	Health Care: Insurance	9	0	9
Paccar	U.S.	Motor Vehicles & Parts	9	0	9
Costco Wholesale	U.S.	General Merchandisers	8	0	8
New York Life Insurance	U.S.	Insurance: Life, Health (Mutual)	8	0	8
Thermo Fisher Scientific	U.S.	Medical Products and Equipment	8	0	8
American Express	U.S.	Diversified Financials	7	0	7
Electricite de France	France	Utilities	7	0	7
Goldman Sachs Group	U.S.	Banks: Commercial and Savings	7	0	7
Munich Re Group	Germany	Insurance: Property and Casualty	7	0	7
Nationwide	U.S.	Insurance: Property and Casualty	7	0	7

Company	Country	Industry	Chrome 2xx	ChatGPT 2xx	Gap
US Foods Holding	U.S.	Wholesalers: Food and Grocery	7	0	7
Cisco Systems	U.S.	Network and Other Comm. Equipment	6	0	6
ELO Group	France	Food & Drug Stores	6	0	6
Lenovo Group	China	Computers, Office Equipment	6	0	6
Metro	Germany	Wholesalers: Food and Grocery	4	0	4
Philip Morris International	U.S.	Tobacco	4	0	4
AIA Group	China	Insurance: Life, Health (stock)	3	0	3
China Mobile Communications	China	Telecommunications	2	0	2
MS&AD Insurance Group Holdings	Japan	Insurance: Property and Casualty	2	0	2
Schneider Electric	France	Electronics, Electrical Equip.	2	0	2
Air France-KLM Group	France	Airlines	1	0	1
BHP Group	Australia	Mining, Crude-Oil Production	1	0	1
Bank of Montreal	Canada	Banks: Commercial and Savings	1	0	1
CK Hutchison Holdings	China	Specialty Retailers	1	0	1
Cardinal Health	U.S.	Wholesalers: Health Care	1	0	1
Daiwa House Industry	Japan	Engineering & Construction	1	0	1
Edeka Zentrale	Germany	Wholesalers: Food and Grocery	1	0	1
Iberdrola	Spain	Utilities	1	0	1
Intel	U.S.	Semiconductors	1	0	1
J. Sainsbury	Britain	Food & Drug Stores	1	0	1
Massachusetts Mutual Life Insurance	U.S.	Insurance: Life, Health (Mutual)	1	0	1

Company	Country	Industry	Chrome 2xx	ChatGPT 2xx	Gap
Panasonic Holdings	Japan	Electronics, Electrical Equip.	1	0	1
Saudi Aramco	Saudi Arabia	Mining, Crude-Oil Production	1	0	1
Tesco	Britain	Food & Drug Stores	1	0	1
Unilever	Britain	Household and Personal Products	1	0	1

Industry Concentration

The GEO gap is not evenly distributed. Certain industries are disproportionately affected.

Telecommunications leads with 4 companies out of 15 in the dataset (26.7% of the sector) blocking ChatGPT-User while serving Chrome. This includes AT&T, Comcast, China Telecommunications, and China Mobile — companies whose customer-facing content (plans, coverage maps, support documentation) is precisely the type of information users ask AI systems about.

Wholesalers and food retailers show a cluster of 6 companies (US Foods, ELO Group, Edeka, Metro, J. Sainsbury, Tesco) — an industry where AI-driven product and supplier research is growing rapidly.

The software sector contributes Oracle and Salesforce — a notable finding given that Salesforce maintains one of the most comprehensive llms.txt files in the Fortune 500 (206 lines, 205 links), yet blocks the ChatGPT-User agent that would actually read it. This contradiction is examined further in Chapter 9.

Geographic Distribution

The United States accounts for 23 of the 53 companies in the gap (43%), followed by China and Britain (5 each), Japan and France (4 each), and Germany (3). The US concentration is expected given that 139 of the 500 companies are US-headquartered, but the 23/139 rate (16.5%) is higher than the overall gap rate (53/500 = 10.6%), suggesting US companies are more likely to implement AI-specific blocking than the global average.

What This Means in Practice

When a ChatGPT user asks a question that requires current information from one of these 53 companies, the model cannot retrieve it. The practical consequences differ by company type:

For Johnson & Johnson (Chrome=64, ChatGPT=0): A patient asking ChatGPT about a J&J medication receives information from the model's training data rather than the current prescribing information on jnj.com. If a drug label was updated, a recall was issued, or dosing guidance changed, the AI answer may be incorrect.

For Lockheed Martin (Chrome=29, ChatGPT=0): A defense analyst asking ChatGPT to summarize Lockheed's latest quarterly results receives stale data. The company's investor relations content, press releases, and contract announcements — all served to Chrome — are invisible to the AI.

For Toyota Motor (Chrome=11, ChatGPT=0): A car buyer asking ChatGPT to compare Toyota models gets information from training data rather than Toyota's current model lineup, pricing, and specifications.

In each case, the company's own authoritative content is replaced by whatever the model can reconstruct from its training corpus — content that may be months old, sourced from third parties, or simply wrong.

The Gap Is Intentional

This finding should not be confused with general bot blocking. The 53 companies in the GEO gap have demonstrated, through their differential treatment of Chrome and ChatGPT-User, that they can distinguish between browser and AI traffic and have chosen to block the latter.

This is a deliberate policy decision. Whether driven by legal caution (AI training concerns), competitive strategy (preventing AI systems from surfacing content to competitors), or simply default WAF configuration that was never reviewed for AI-era implications, the effect is the same: these companies are invisible in the fastest-growing information retrieval channel.

Source data: *Cross-database comparison of* `2026-04-18_chrome_ua` *and* `2026-04-18_chatgpt_user_ua` *runs. HTTP probes only (64 per company), excluding DNS, CT log, and TLS certificate probes. A company is classified in the GEO gap when Chrome 2xx > 0 AND ChatGPT-User 2xx = 0. Revenue data from us500.com Fortune Global 500 dataset (fiscal year 2023/2024). Company-level probe results available in the appendix data tables.*

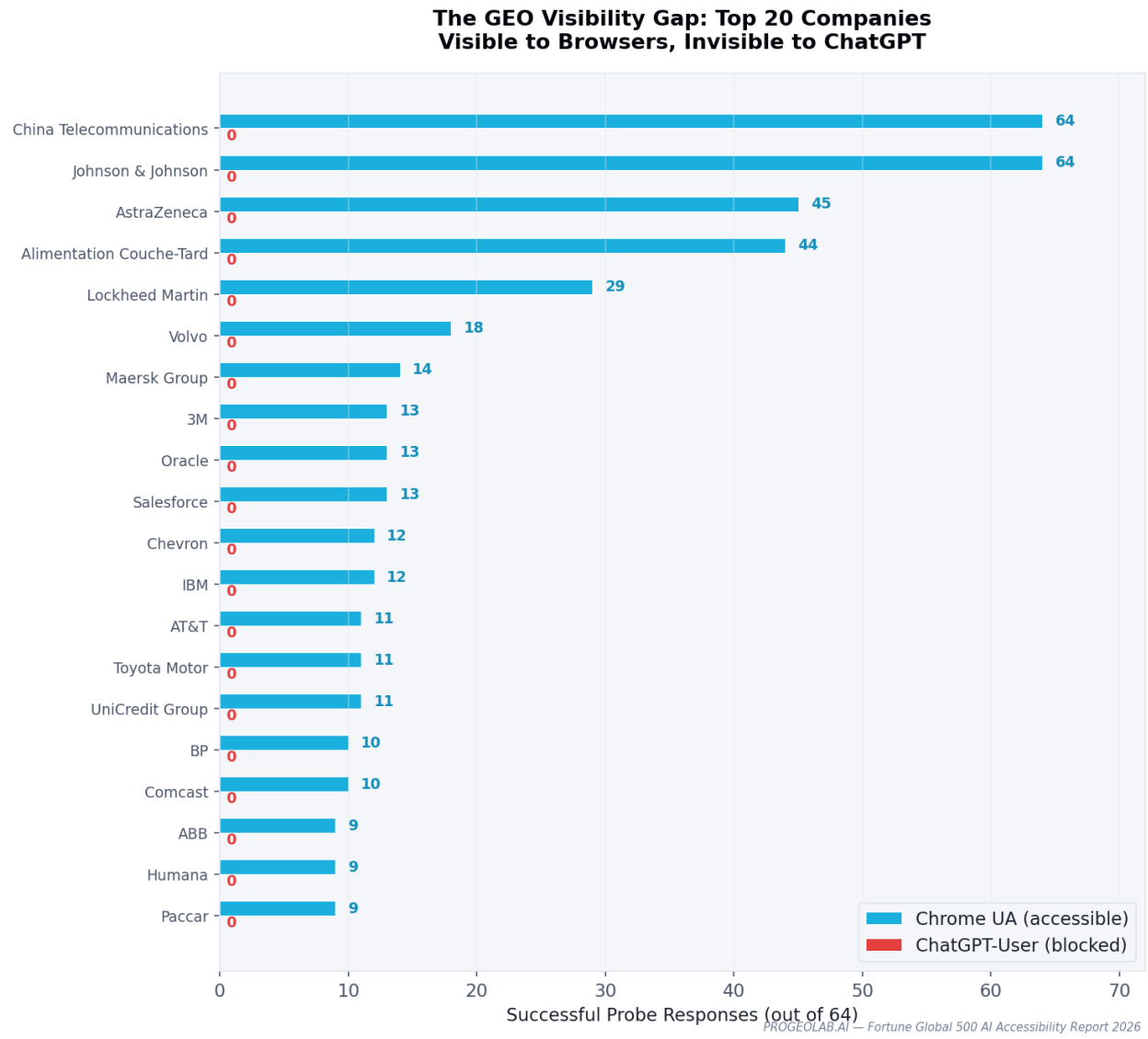

Figure 4.1 · Gap top20

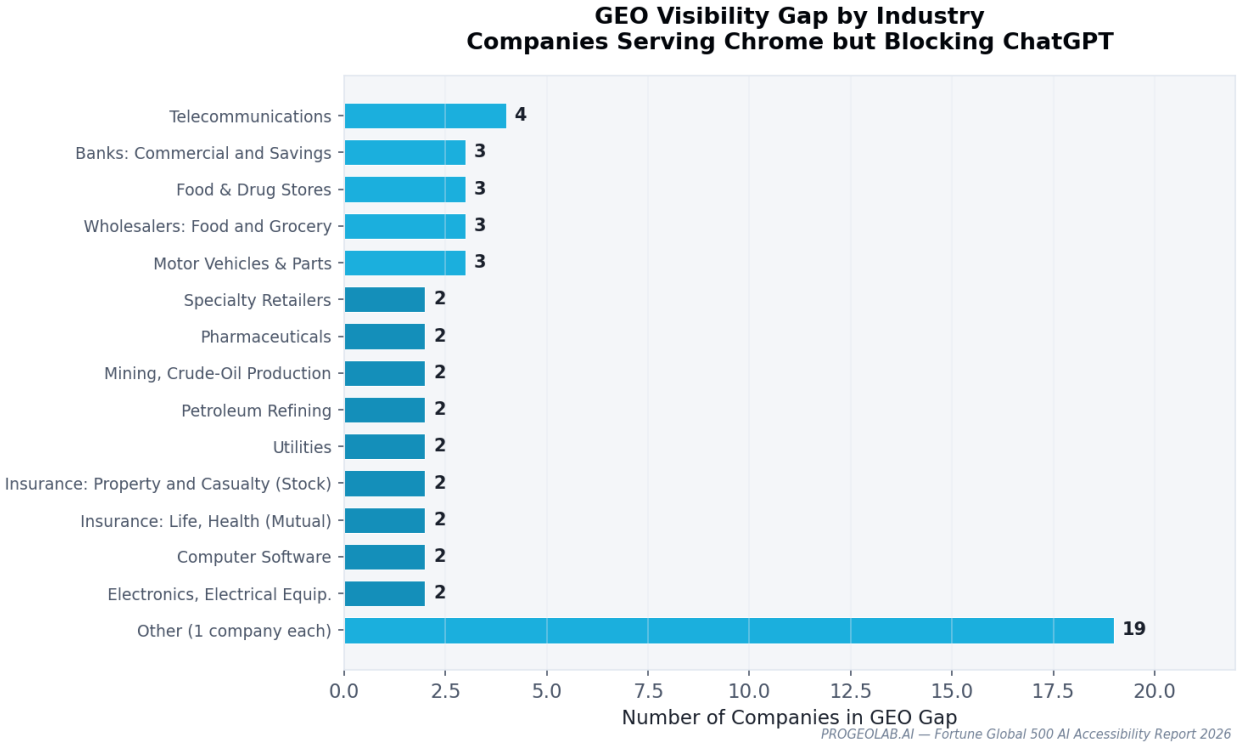

Figure 4.2 · Gap by industry

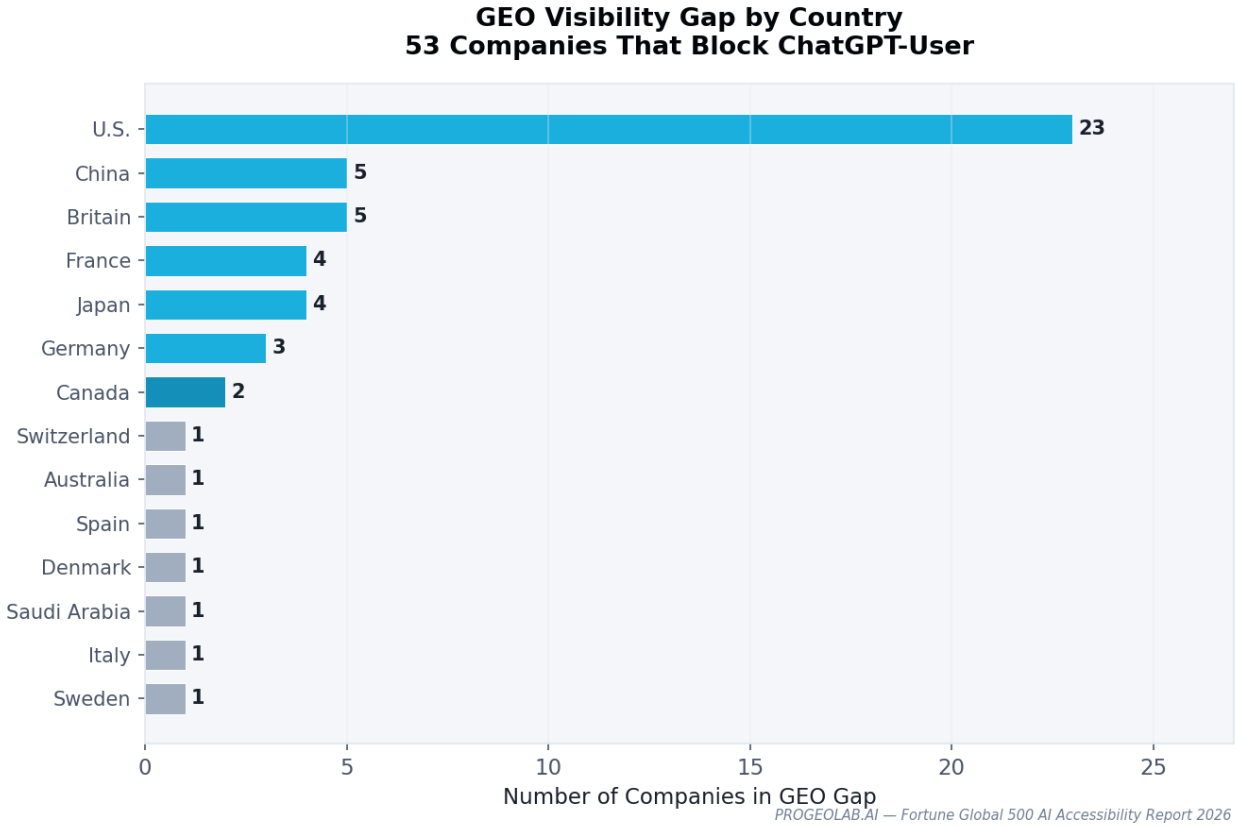

Figure 4.3 · Gap by country

CHAPTER 5

Three Layers of AI Bot Blocking

The 500 companies that fail to respond to at least one user agent do not all fail for the same reason. Analysis of the four-way comparison data reveals three distinct blocking mechanisms operating at different layers of the network stack, each with different implications for AI visibility.

Layer 1: User-Agent String Detection (53 companies)

The most deliberate form of blocking. These 53 companies (detailed in Chapter 4) inspect the User-Agent header and treat AI-identified requests differently from browser requests. The same IP address, the same TLS handshake, the same HTTP request — only the UA string differs — and the server returns 200 to Chrome but 403 to ChatGPT-User.

This is a policy decision, not a technical limitation. The WAF or application layer is configured with rules that match known AI bot identifiers (`ChatGPT-User`, `GPTBot`, etc.) and reject them. This layer is the easiest to implement and the easiest to reverse.

Johnson & Johnson exemplifies this pattern: 64 out of 64 probes succeed with Chrome, zero succeed with ChatGPT-User. The WAF is explicitly configured to pass browser traffic while rejecting AI-identified traffic on every path.

Layer 2: IP Reputation and Datacenter Detection (24 companies)

A deeper form of blocking that operates below the UA string. These 24 companies return 403 (or equivalent blocks) to all four user agents, including Chrome. Changing the User-Agent header has no effect because the WAF is making its decision based on the source IP address.

Enterprise WAFs maintain databases of known datacenter IP ranges (AWS, Azure, GCP, Hetzner, OVH, and others). Requests originating from these ranges are flagged as automated regardless of what UA string they present. Our probes ran from a datacenter environment, triggering this classification.

The 24 IP-blocked companies include:

Company	Country	Research	Googlebot	Chrome	ChatGPT
Tesla	U.S.	0	0	0	0
Home Depot	U.S.	0	0	0	0
Allianz	Germany	0	0	0	0
Taiwan Semiconductor Manufacturing	Taiwan	0	0	0	0
Lufthansa Group	Germany	0	0	0	0
Marathon Petroleum	U.S.	0	0	0	0
Brookfield	Canada	0	0	0	0

Company	Country	Research	Googlebot	Chrome	ChatGPT
Publix Super Markets	U.S.	0	0	0	0
Vale	Brazil	0	0	0	0
SNCF Group	France	0	0	0	0

These sites would likely be accessible from a residential IP address. The blocking is not about who is asking — it is about where the request originates. This has direct implications for AI crawler infrastructure: OpenAI runs ChatGPT-User from Azure datacenter IPs, Anthropic runs ClaudeBot from AWS. Both face the same IP-reputation challenge.

Layer 3: TLS Fingerprinting (15 companies)

The most technically sophisticated blocking layer. These 15 companies terminate the connection during or immediately after the TLS handshake, before the HTTP request (including the User-Agent header) is even transmitted.

The evidence is in the error signatures. Our probe tool (httpx, a Python HTTP client) generates a TLS ClientHello message with a JA3 fingerprint that differs from any real browser. Modern WAFs — particularly Cloudflare Bot Management, Akamai Bot Manager, and F5 Shape Security — inspect the TLS fingerprint and compare it against known browser fingerprint databases. When the JA3 hash does not match Chrome, Firefox, Safari, or another recognized browser, the server resets the connection.

In the Chrome UA run, 666 probe responses across 15 companies returned HTTP/2 StreamReset errors (error code 2). This error occurs when the server has already established the TLS connection and begun HTTP/2 framing, then abruptly terminates. The timing — after TLS but before any meaningful HTTP exchange — indicates fingerprint-based rejection.

No UA string change can bypass this layer. The blocking occurs before the server reads the UA header. Only a real browser (or a tool that replicates a browser's TLS fingerprint, such as a headless Chromium instance) would pass.

The Three Layers in Combination

Most blocked companies exhibit only one layer of blocking. But the layers are not mutually exclusive — some companies deploy multiple detection mechanisms:

Blocking Pattern	Companies	Example
Layer 1 only (UA string)	53	Johnson & Johnson: Chrome ✓, ChatGPT ✗
Layer 2 only (IP reputation)	24	Tesla: all UAs blocked equally
Layer 3 only (TLS fingerprint)	~15	Sites with StreamReset errors even for Chrome UA
Layer 1 + Layer 2	~6	Sites that block AI UAs AND block datacenter IPs for browsers

Blocking Pattern	Companies	Example
All three layers	~3	Home Depot, Allianz: block everything from datacenter + fingerprint

Implications

For enterprises evaluating their AI visibility posture: Layer 1 blocking (UA-based) is the only layer under direct content policy control. If your organization blocks ChatGPT-User in your WAF rules, that is a reversible policy decision. Layers 2 and 3 are infrastructure decisions typically managed by security teams who may not consider AI visibility implications.

For AI companies building crawlers: Layers 2 and 3 represent a structural challenge. Running crawlers from datacenter IPs triggers Layer 2 blocking. Using standard HTTP libraries triggers Layer 3 blocking. The only technical solutions — residential proxy networks or headless browsers — raise their own ethical and cost questions.

For the GEO industry: The three-layer model explains why simple "check your robots.txt" advice is insufficient. A company can have a perfectly permissive robots.txt, explicitly allow GPTBot and ChatGPT-User, and still be invisible to AI systems because their WAF blocks at the IP or TLS layer before the HTTP request (including robots.txt) is ever evaluated.

Source data: *Layer 1 identified by cross-database comparison (Chrome 2xx > 0 AND ChatGPT-User 2xx = 0). Layer 2 identified by companies with 403 responses across ALL four user agents, confirmed via raw response body WAF signature detection (Chapter 7). Layer 3 identified by* `StreamReset` *error patterns in the Chrome UA run (*`2026-04-18_chrome_ua` *database,* `error_detail LIKE '%StreamReset%'`*). TLS fingerprinting attribution based on error timing analysis (post-TLS, pre-HTTP patterns).*

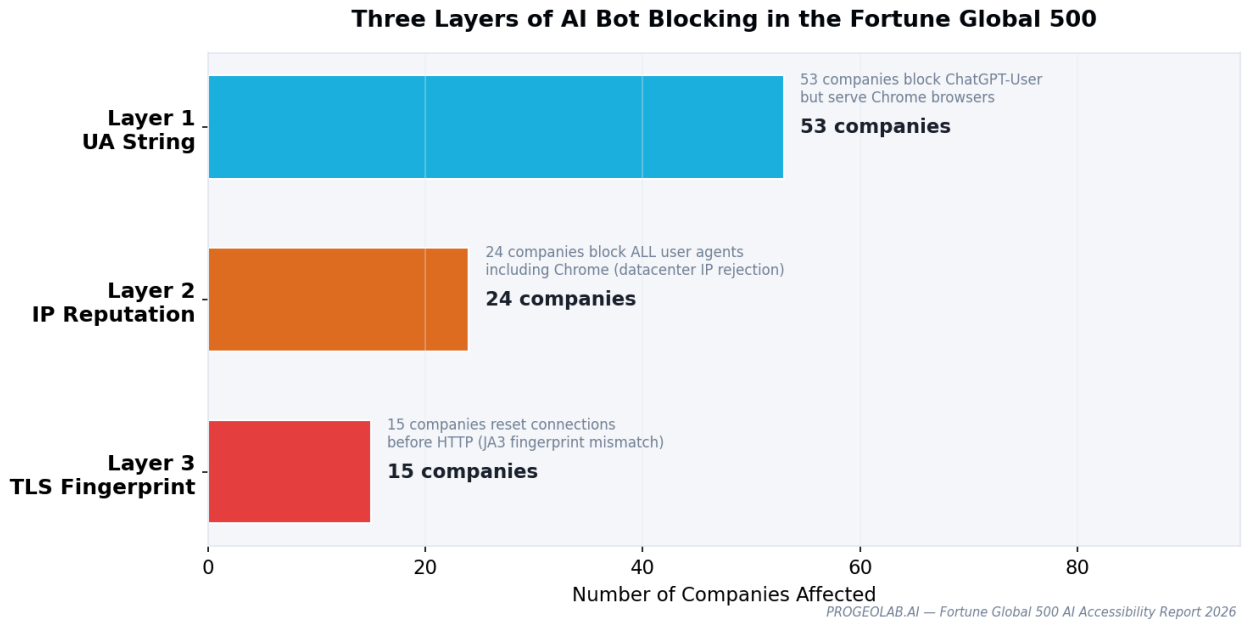

Figure 5.1 · Three layers

CHAPTER 6

The Googlebot Impersonation Backfire

One of the most counterintuitive findings of this study emerged from the Googlebot UA run: presenting as Google's search crawler produced worse results than using an honest, unknown research bot.

The Googlebot UA reached 307 companies, compared to 325 for the research UA — a net loss of 18 companies. This deficit is not explained by random variation. It is caused by a specific security mechanism: reverse DNS verification of Googlebot identity.

How Googlebot Verification Works

Google documents that legitimate Googlebot requests originate from IP addresses that resolve via reverse DNS to `*.googlebot.com` or `*.google.com` hostnames. Google's own documentation recommends this verification method.

Enterprise WAFs implement this check. When a request arrives with `User-Agent: Googlebot/2.1`, the WAF performs a reverse DNS lookup on the source IP. If the result is not a Google-owned hostname, the request is flagged as impersonation and blocked — often more aggressively than an unknown bot would be.

Our probe requests originated from a datacenter IP that does not resolve to a Google hostname. Every Googlebot-identified request was therefore detectable as an impersonator.

The 32 Companies That Detect Impersonation

Thirty-two companies that responded to our research UA (with at least 1 successful probe) returned zero successful responses to the Googlebot UA. These companies have WAF configurations that specifically penalize Googlebot impersonation:

Company	Country	Research UA (2xx)	Googlebot UA (2xx)	Chrome (2xx)
Shandong Hi-Speed Group	China	61	0	0
Maersk Group	Denmark	14	0	14
Salesforce	U.S.	13	0	13
IBM	U.S.	12	0	12
AT&T	U.S.	11	0	11
Intel	U.S.	10	0	1
Samsung Electronics	South Korea	10	0	10
BMW Group	Germany	9	0	0

Company	Country	Research UA (2xx)	Googlebot UA (2xx)	Chrome (2xx)
Ford Motor	U.S.	9	0	0
Humana	U.S.	9	0	9
Royal Ahold Delhaize	Netherlands	9	0	9
Saudi Aramco	Saudi Arabia	9	0	1
Schneider Electric	France	9	0	2
Volvo	Sweden	9	0	18
American Express	U.S.	8	0	7
Cisco Systems	U.S.	8	0	6
Dow	U.S.	8	0	0
General Dynamics	U.S.	8	0	8
ING Group	Netherlands	8	0	0
New York Life Insurance	U.S.	8	0	8

The pattern is consistent: these companies allow an unknown bot (research UA) to access some content, but when the same request claims to be Googlebot from a non-Google IP, it is rejected entirely. The impersonation attempt triggers stricter enforcement than honest identification.

The Mechanism

The Googlebot impersonation backfire operates through a specific decision chain in the WAF:

1. **Request arrives** with `User-Agent: Googlebot/2.1`
2. **WAF checks** the source IP against Google's published IP ranges (or performs reverse DNS)
3. **Mismatch detected** — the IP is not in Google's CIDR blocks
4. **Impersonation flag raised** — the request is classified as malicious rather than merely automated
5. **Enhanced blocking applied** — stricter than the default bot policy, because impersonation suggests adversarial intent

This is a rational security response. An unknown bot might be harmless research or a competitor's crawler. A fake Googlebot is actively lying about its identity, which correlates with credential stuffing, content scraping, and other adversarial activities.

Implications for AI Crawlers

This finding has a direct implication for AI companies: the identity verification gap between AI crawlers and traditional search crawlers is structural.

Google, Bing, and Yandex have established verification mechanisms (reverse DNS, published IP ranges) that WAFs can check. GPTBot, ClaudeBot, and PerplexityBot publish their IP ranges but lack the reverse DNS infrastructure that has been the standard verification method for over a decade.

Until AI crawlers implement equivalent identity verification — and until WAF vendors integrate those verification lookups into their products — AI-identified requests will remain harder to verify and easier to block than search engine requests from established providers.

The Googlebot impersonation finding also explains a pattern observed in the data: companies that block AI crawlers at Layer 1 (UA-based blocking) are disproportionately the same companies with sophisticated Googlebot verification. They have invested in bot management infrastructure and have made deliberate decisions about which automated access to permit.

> ***Source data:*** *The 32 companies identified by comparing* `2026-04-16_baseline` *(research UA) against* `2026-04-17_googlebot_ua` *runs. A company is classified as impersonation-detecting when research UA 2xx > 0 AND Googlebot UA 2xx = 0. Google's Googlebot verification documentation: https://developers.google.com/search/docs/crawling-indexing/verifying-googlebot. Reverse DNS verification is the recommended method.*

CHAPTER 7

The WAF Landscape — F5, Cloudflare, Akamai, and AI

Web Application Firewalls are the infrastructure layer that mediates — and often blocks — AI crawler access to enterprise websites. Raw response body analysis across the Chrome UA run reveals the WAF vendor distribution among Fortune Global 500 companies for the first time.

WAF Vendor Distribution

By analyzing HTTP response body signatures (challenge pages, error page templates, JavaScript challenge patterns, and server headers), we identified the WAF vendor for 300 of the 388 companies that returned any HTTP response in the Chrome UA run.

WAF Vendor	Sites Detected	Market Share (of detected)
F5 BIG-IP	232	53.6%
Cloudflare	64	14.8%
Akamai	59	13.6%
Generic (unidentified)	47	10.9%
Imperva	23	5.3%
PerimeterX	8	1.8%
AWS WAF	8	1.8%
Sucuri	1	0.2%

F5 BIG-IP: The Enterprise Default

F5 BIG-IP dominates the Fortune 500 WAF landscape at 232 detected deployments — more than Cloudflare, Akamai, and Imperva combined. This finding contrasts with the public narrative around WAFs, which tends to focus on Cloudflare due to its visibility in the developer and startup ecosystem.

F5's dominance in this dataset reflects its position in traditional enterprise IT. Fortune 500 companies — banks, insurers, energy companies, manufacturers — have long-standing relationships with F5 for load balancing and application delivery. WAF functionality is an add-on to existing infrastructure rather than a standalone purchase.

For AI visibility, F5 BIG-IP's behavior varies by configuration. Some F5 deployments are transparent (pass all traffic), while others implement aggressive bot management through F5 Shape Security. The range of outcomes within the F5-detected cohort is wider than for any other vendor.

Cloudflare: The Bot Management Leader

Cloudflare appears on 64 Fortune 500 sites and presents the most consistent AI-blocking behavior. Cloudflare's Bot Management product uses a combination of IP reputation, TLS fingerprinting (JA3/JA4), and JavaScript challenge to classify requests.

Cloudflare's challenge pages are distinctive in response bodies — they include the "Just a moment..." interstitial, `cf-ray` headers, and `challenge-platform` JavaScript. These markers made Cloudflare the most reliably detectable WAF in our analysis.

For AI crawlers, Cloudflare presents a particular challenge: its TLS fingerprinting is among the most aggressive in the industry. Even a Chrome User-Agent string is insufficient if the TLS ClientHello does not match a known browser fingerprint. This explains some of the Layer 3 (TLS fingerprint) blocking identified in Chapter 5.

Akamai: Enterprise CDN with Bot Management

Akamai appears on 59 sites, primarily large consumer brands, retailers, and financial institutions. Akamai's Bot Manager product integrates with their CDN infrastructure and uses behavioral analysis alongside traditional signals.

Akamai's blocking pages typically include `AkamaiGHost` server headers or `Reference #` error codes in the response body. Their bot detection includes client-side behavioral analysis that cannot be bypassed by any non-browser HTTP client.

Imperva (Incapsula)

Imperva appears on 23 sites and is identifiable by `_imp_apg_r_` cookies and characteristic challenge page patterns. Imperva's Advanced Bot Protection uses device fingerprinting and behavioral analysis.

What This Means for AI Visibility Strategy

The WAF vendor matters because each vendor's bot management product treats AI crawlers differently:

Configuration exposure: F5 deployments have the widest range of AI accessibility outcomes because F5 gives administrators the most granular control. A company using F5 with a permissive bot policy may be fully accessible, while another F5 customer with Shape Security enabled blocks everything. The variability is a configuration choice, not a vendor limitation.

Cloudflare's structural challenge: Cloudflare's TLS fingerprinting creates a baseline blocking layer that no User-Agent change can bypass. Companies on Cloudflare who want to allow AI crawler access must explicitly configure Bot Management exceptions — a step that requires awareness that the issue exists.

The verification gap: None of the major WAF vendors currently offer native verification for AI crawler identity (equivalent to Googlebot reverse DNS verification). This means WAF administrators who want to allow GPTBot but block scrapers have no automated way to verify that a request claiming to be GPTBot actually originates from OpenAI. Until WAF vendors add AI crawler verification features, the all-or-nothing dynamic will persist.

> ***Source data:*** *WAF vendor attribution performed on raw HTTP response bodies from the* `2026-04-18_chrome_ua` *run. Detection based on signature matching in response HTML, JavaScript challenge patterns, HTTP headers, and cookie names. Detection signatures documented*

in `extract_raw_data.py` (WAF_SIGNATURES dictionary). 300 of 388 responding companies had detectable WAF signatures. The remaining 88 either use WAFs with unrecognized signatures or do not deploy a WAF on their public-facing corporate website.

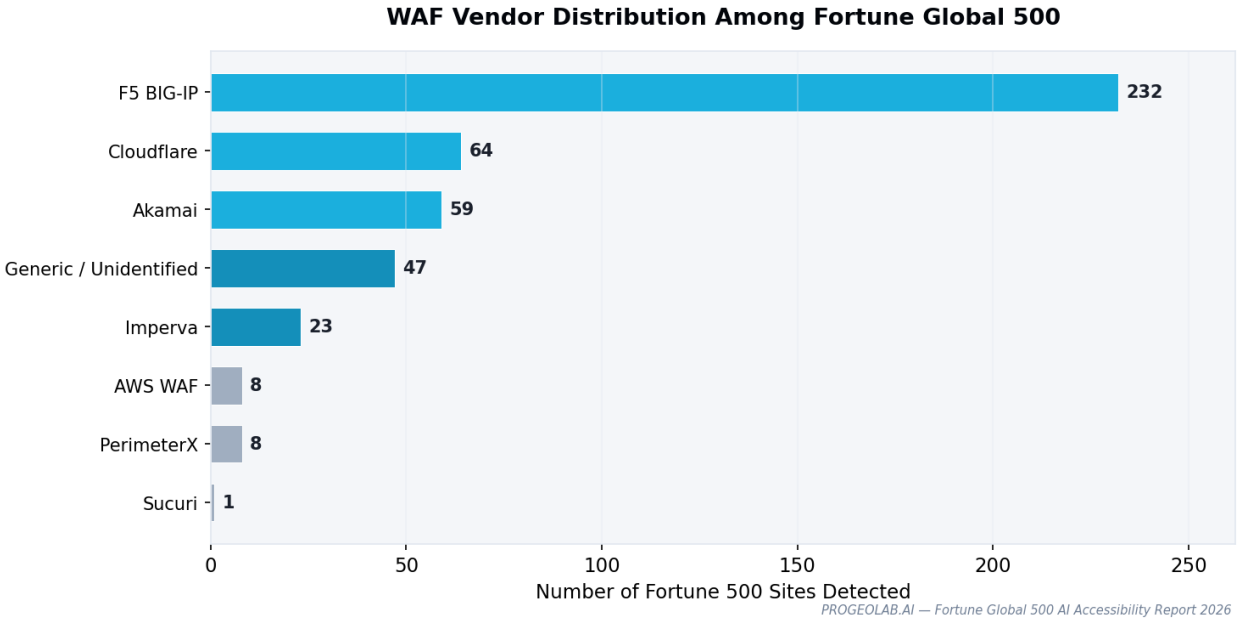

Figure 7.1 · Waf vendors

CHAPTER 8

The Soft-404 Problem — Why Status Codes Lie

A significant methodological finding of this study challenges the accuracy of any AI visibility audit that relies solely on HTTP status codes: 160 of the 500 companies in our dataset serve catch-all pages that return HTTP 200 for any URL path, including paths that definitively do not exist.

This means that naive measurements of llms.txt adoption, agents.json deployment, or MCP server presence are inflated by approximately 30%. The true adoption rate of these emerging standards is significantly lower than what status-code-only scans report.

What Is a Soft 404?

A soft 404 occurs when a web server returns an HTTP 200 (OK) status code for a URL that does not contain the requested content. Instead of a proper 404 response, the server sends its default page — typically the homepage, a branded error page, or a generic template — with a success status code.

For traditional SEO, soft 404s are a known problem that search engines have learned to detect through content analysis. For AI visibility auditing, the problem is more acute because the files being probed (llms.txt, agents.json, mcp.json) are new enough that automated content validation has not yet been developed.

How We Detected Soft 404s

Our raw body analysis compared the response content across multiple probe paths that should return different content — or not exist at all. Specifically, we fingerprinted the response bodies for these paths per company:

- `/.well-known/agents.json`
- `/.well-known/agents.txt`
- `/.well-known/mcp.json`
- `/.well-known/mcp-server`
- `/llms.txt`
- `/llms-full.txt`
- `/changelog`

When three or more of these different-purpose paths returned identical response body hashes, we classified the site as a soft-404 site. The reasoning: it is not plausible that a company's agents.json, MCP server configuration, and llms.txt all contain exactly the same content.

The Scale of the Problem

Of 388 companies that returned any HTTP response in the Chrome UA run:

- **160 companies (41%)** were classified as soft-404 sites
- These sites returned HTTP 200 on every probed path, including paths like `/.well-known/model-context-protocol.json` that no Fortune 500 company has actually implemented

Impact on Adoption Metrics

The soft-404 problem directly inflates every probe's hit rate:

Metric	Status-Code Count	After Soft-404 Filtering	Inflation
llms.txt present	357 sites (71.4%)	**14 sites (2.8%)**	**25× overstated**
agents.json present	~154 sites	**~0 sites**	entirely false
mcp.json present	~135 sites	**~0 sites**	entirely false
security.txt present	~143 sites	**55 sites (11%)**	~2.6× overstated

The llms.txt case is the most striking: 353 companies appeared to have the file based on HTTP 200 responses. Raw body analysis revealed that 339 of those were HTML error pages, generic landing pages, or homepage redirects served with a 200 status. Only 14 files contained actual llms.txt-format content (Markdown with headers and links).

For agents.json and MCP endpoints, the soft-404 inflation is even more severe — we found zero real implementations among Fortune 500 companies, yet status codes alone would suggest over 100.

Examples of Soft-404 Content

Examining the raw response bodies reveals what these soft-404 pages actually contain:

Abbott Laboratories — `/llms.txt` returns HTTP 200 with the full HTML of the Abbott 404 error page, including the title "Page Not Found | Abbott" and a 200 status code. The HTML contains navigation, footer, and marketing content — none of which is llms.txt content.

Deutsche Bank — `/.well-known/agents.json` returns HTTP 200 with the Deutsche Bank homepage HTML. The same HTML is returned for every unknown path.

Starbucks — All 7 agent/MCP paths return identical 200 responses containing the Starbucks mobile app redirect page.

Why This Happens

Soft 404s at this scale are typically caused by:

1. **Single-page application (SPA) routing:** React/Angular/Vue applications serve the same `index.html` for all paths and handle routing client-side. The server has no knowledge of which paths exist.
1. **Default error page configuration:** Some web servers and CDNs are configured to return a branded error page with a 200 status instead of 404, often to avoid showing "ugly" error pages.
1. **Catch-all redirects:** Load balancers configured to route all unmatched paths to the homepage or a default landing page.

Implications for the GEO Industry

Any tool or research claiming to measure llms.txt adoption, AI agent readiness, or MCP deployment must validate response bodies, not just status codes. A status-code-only scan will report adoption rates 10–17

times higher than reality.

This is not a theoretical concern. Several widely cited industry reports on llms.txt adoption rely on status code scans. Our finding suggests their numbers are significantly overstated.

The PROGEOLAB audit methodology addresses this through two mechanisms: automated body fingerprinting (hash comparison across multiple paths to detect catch-all servers) and content validation (parsing the response body for expected format markers — Markdown headers and links for llms.txt, JSON structure for agents.json).

> ***Source data:*** *Soft-404 classification based on MD5 hash comparison of response bodies from* `extracted_data.json`*. A company is classified as soft-404 when 3 or more probe paths (from the 7-path fingerprint set) return identical body hashes. Raw response bodies from the Chrome UA run (*`f500_audit_chrome_ua/runs/raw/2026-04-18_chrome_ua/`*). llms.txt validation performed by content-type analysis: HTML detection (*`<html`*, *`<!DOCTYPE`*), error page detection (*`404`*, *`not found`*, *`error`*), and format validation (Markdown H1 presence, link syntax).*

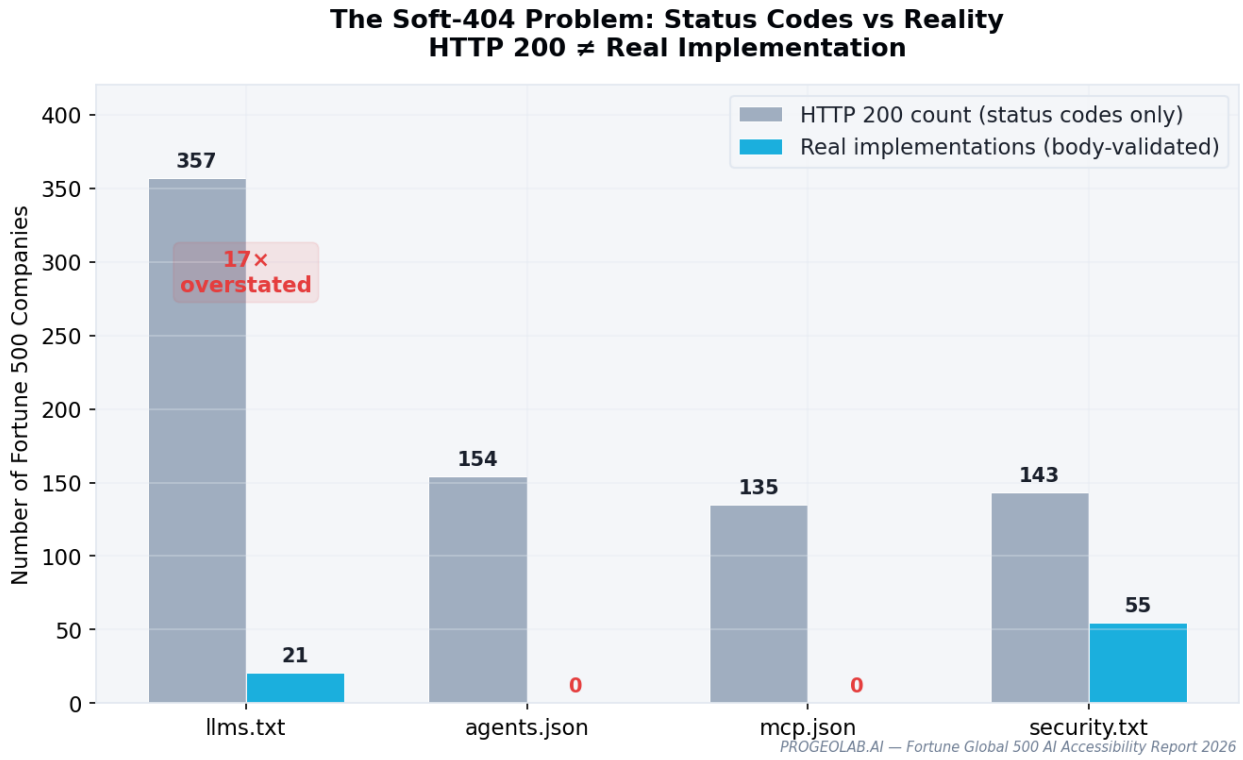

Figure 8.1 · Soft404 inflation

CHAPTER 9

Industry Analysis — Which Sectors Block AI Most

The 53-company GEO visibility gap is not evenly distributed across industries. Certain sectors show structural patterns in how they handle AI crawler access, driven by regulatory environment, competitive dynamics, and security posture.

AI Accessibility by Industry

The following table shows the 15 largest industry sectors in the Fortune Global 500 dataset, ranked by their GEO gap rate — the proportion of companies in each sector that serve Chrome but block ChatGPT-User.

Industry	Companies	Chrome-Accessible	ChatGPT-Accessible	GEO Gap	Gap Rate
Telecommunications	15	14	10	**4**	**26.7%**
Wholesalers: Food and Grocery	5	5	2	**3**	**60.0%**
Food & Drug Stores	18	13	10	**3**	**16.7%**
Motor Vehicles & Parts	37	23	20	**3**	**8.1%**
Banks: Commercial and Savings	57	45	43	**3**	**5.3%**
Pharmaceuticals	14	13	11	**2**	**14.3%**
Computer Software	4	4	2	**2**	**50.0%**
Petroleum Refining	33	23	21	**2**	**6.1%**
Electronics, Electrical Equip.	14	10	8	**2**	**14.3%**
Insurance: Prop. & Casualty	18	14	12	**2**	**11.1%**
Mining, Crude-Oil Production	17	10	8	**2**	**11.8%**
Metals	23	13	13	0	0%
Trading	15	8	8	0	0%
Energy	14	7	7	0	0%

Industry	Companies	Chrome-Accessible	ChatGPT-Accessible	GEO Gap	Gap Rate
Engineering & Construction	17	6	5	1	5.9%

Sector-Specific Patterns

Telecommunications (26.7% gap rate): The telecom sector has the highest concentration of AI-specific blocking among major industries. AT&T, Comcast, China Telecommunications, and China Mobile all block ChatGPT-User while serving Chrome. Telecom companies handle sensitive customer data and operate heavily regulated networks. Their security teams are accustomed to blocking automated access, and AI crawlers are classified alongside other automated threats.

Computer Software (50% gap rate, small sample): Oracle and Salesforce — two of the four software companies in the dataset — block ChatGPT-User. This is particularly notable for Salesforce, which maintains one of the most comprehensive llms.txt files found in this study (206 lines, 205 links to documentation), yet blocks the user agent that would read it. The llms.txt exists to guide AI systems, but the WAF prevents AI systems from reaching it.

Pharmaceuticals (14.3% gap rate): Johnson & Johnson and AstraZeneca represent the pharma gap. Pharmaceutical companies face unique regulatory requirements around content distribution. Drug information, clinical trial data, and prescribing information are subject to FDA and EMA regulations that may create caution about uncontrolled AI redistribution of medical content.

Wholesalers: Food and Grocery (60% gap rate, small sample): Three of five food wholesalers block ChatGPT-User (US Foods, Edeka, Metro). This high rate in a traditionally low-tech sector suggests default WAF configurations rather than deliberate AI policy decisions.

Banks (5.3% gap rate): Despite being heavily regulated and security-conscious, banks show a relatively low AI-blocking rate. Of 57 banks, only 3 (Goldman Sachs, UniCredit, Bank of Montreal) fall in the GEO gap. This may reflect the financial sector's earlier adoption of digital transformation and API-driven architectures, which require more nuanced bot management than binary block/allow rules.

Metals and Trading (0% gap rate): These sectors show no AI-specific blocking. Companies that are accessible to Chrome are also accessible to ChatGPT-User. This likely reflects simpler web infrastructure with fewer WAF configurations.

The China Factor

Chinese companies present a distinct pattern that is not primarily about AI-specific blocking. Of 128 Chinese companies in the dataset, 78 are unreachable from our datacenter location (connection errors across all four user agents). This is consistent with Great Firewall infrastructure that blocks or severely degrades international traffic to Chinese corporate websites.

Among the 50 Chinese companies that are reachable, only 5 fall in the GEO gap (China Telecommunications, China Mobile, Lenovo, AIA Group, CK Hutchison). The remaining 45 reachable Chinese sites treat Chrome and ChatGPT-User equivalently — they either serve both or block both.

The US Dominance

US companies account for 23 of the 53 companies in the GEO gap — 43% of the gap, from 28% of the dataset. The US gap rate (23/139 = 16.5%) is materially higher than the global average (53/500 = 10.6%).

This concentration reflects two factors: US companies are more likely to use sophisticated WAF products with AI-specific blocking capabilities, and the US regulatory environment (particularly around AI training data) has created heightened awareness of AI crawler access among legal and compliance teams.

Industries Absent from the Gap

Several industries show zero GEO gap despite having multiple accessible companies: Metals (13 chrome-accessible, 13 chatgpt-accessible), Trading (8/8), and Energy (7/7). These sectors tend to have simpler corporate websites, fewer WAF customizations, and less concern about AI-specific content access.

Source data: *Industry classifications from the us500.com Fortune Global 500 dataset. GEO gap counts from cross-database comparison of Chrome UA and ChatGPT-User UA runs. Gap rate = (companies with Chrome 2xx > 0 AND ChatGPT 2xx = 0) / (total companies in industry). Industries with fewer than 4 companies excluded from gap rate analysis to avoid small-sample distortion.*

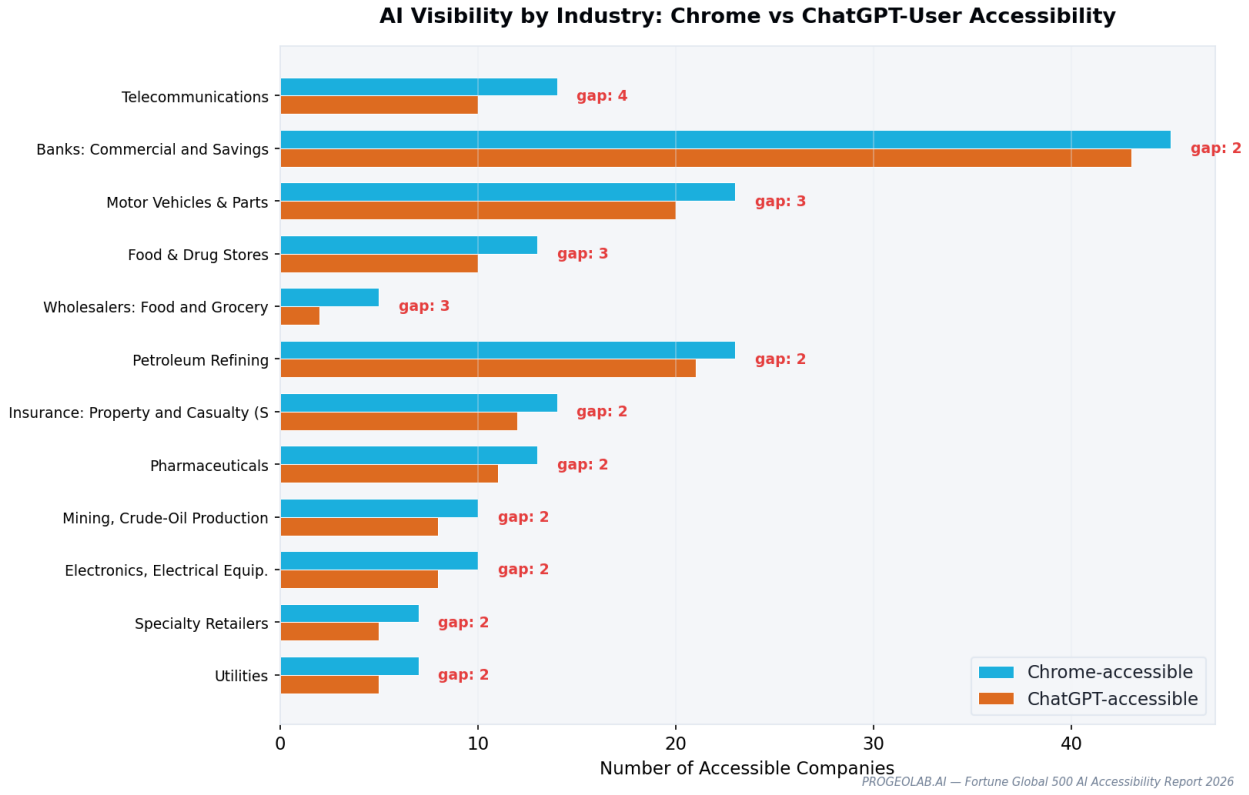

Figure 9.1 · Industry gap

CHAPTER 10

Country and Regional Analysis

The Fortune Global 500 spans 35 countries. AI crawler accessibility varies dramatically by geography, driven by three distinct factors: national internet infrastructure (the Great Firewall), WAF adoption patterns, and regulatory environment.

Accessibility by Country

Country	Companies	Chrome-Accessible	ChatGPT-Accessible	Unreachable	GEO Gap
U.S.	139	117 (84%)	94 (68%)	17	23
China	128	48 (38%)	44 (34%)	78	5
Japan	40	25 (63%)	21 (53%)	14	4
Germany	29	23 (79%)	20 (69%)	4	3
France	24	22 (92%)	18 (75%)	2	4
Britain	17	17 (100%)	12 (71%)	0	5
South Korea	15	12 (80%)	12 (80%)	3	0
Canada	14	11 (79%)	9 (64%)	3	2
Switzerland	11	11 (100%)	10 (91%)	0	1
Netherlands	11	9 (82%)	9 (82%)	1	0
Spain	9	8 (89%)	7 (78%)	1	1
India	9	6 (67%)	6 (67%)	3	0
Brazil	9	6 (67%)	6 (67%)	3	0

The Great Firewall Effect

China's 128 Fortune Global 500 companies represent the single largest national cohort, yet only 48 (38%) are accessible even to a Chrome browser from our datacenter location. Seventy-eight companies (61%) returned connection errors across all four user agents.

This is not AI-specific blocking. The Great Firewall of China restricts inbound international traffic to many Chinese corporate domains. State-owned enterprises — which dominate the Chinese Fortune 500 presence — host their websites on domestic infrastructure that may not be configured for or optimized for international access.

Among the 48 Chinese companies that are reachable, AI accessibility is relatively high: 44 of 48 (92%) also respond to ChatGPT-User. This suggests that once traffic passes through the network layer, Chinese corporate websites do not typically implement AI-specific blocking at the application layer.

The United States: WAF Capital

The United States shows the largest absolute GEO gap: 23 companies accessible to Chrome but blocking ChatGPT-User. The US gap rate (23 out of 117 chrome-accessible = 19.7%) is the highest among countries with more than 10 companies in the dataset.

This concentration reflects the US enterprise security market. US companies are the primary customers of Cloudflare, F5, Akamai, and Imperva. They have the most sophisticated WAF deployments and the most granular bot management configurations. US legal awareness of AI training data issues (prompted by copyright lawsuits and regulatory attention) also drives more aggressive AI-specific blocking policies.

Seventeen US companies are unreachable from our datacenter IP — not due to geographic blocking but due to aggressive datacenter IP rejection (Layer 2 blocking). Tesla, Home Depot, and Publix are examples of US companies that block all traffic from recognized datacenter IP ranges.

Britain: All Accessible, But 5 Block AI

Britain presents a clean case study: all 17 British Fortune Global 500 companies are accessible to Chrome (100%), but 5 block ChatGPT-User. These are AstraZeneca (pharma), BP (energy), J. Sainsbury (retail), Tesco (retail), and Unilever (consumer goods).

The British gap is concentrated in consumer-facing industries where web scraping has been a longstanding concern. UK grocery retailers (Sainsbury's, Tesco) have historically aggressive bot management to protect pricing data.

Japan: The Timeout Problem

Japan has 40 companies in the dataset, of which 14 are unreachable (35%). Unlike China's GFW-driven inaccessibility, Japan's unreachable sites are characterized by timeout errors rather than connection refusals. This suggests network latency issues, restrictive firewall rules for international traffic, or hosting on domestic infrastructure optimized for Japanese ISP routing.

Among the 25 accessible Japanese companies, 4 fall in the GEO gap (Toyota, Daiwa House, MS&AD Insurance, Panasonic). Japanese automotive and electronics companies show a pattern of cautious AI crawler handling.

France and Germany: European Patterns

France (24 companies, 4 in gap) and Germany (29 companies, 3 in gap) show moderate AI-blocking rates. The French gap includes Air France-KLM, ELO Group, Electricite de France, and Schneider Electric. The German gap includes Edeka Zentrale, Metro, and Munich Re.

European companies are subject to GDPR, which has created organizational sensitivity to automated data processing. However, the GDPR connection to AI crawler blocking is indirect — GDPR applies to personal data processing, not to web crawling of public corporate content. The European gap appears to be driven more by WAF configuration defaults (particularly F5 BIG-IP, which is heavily deployed in European enterprises) than by regulatory mandate.

South Korea, Switzerland, Netherlands: Low or Zero Gap

South Korea (15 companies, 0 gap), Switzerland (11 companies, 1 gap), and the Netherlands (11 companies, 0 gap) show minimal AI-specific blocking. These countries have well-developed digital infrastructure and competitive technology sectors that may be more open to AI-driven content access.

Regional Summary

Region	Companies	Chrome Rate	ChatGPT Rate	GEO Gap Rate
North America (US + Canada)	153	83.7%	67.3%	16.3%
Europe (all EU/UK/CH)	102	86.3%	76.5%	11.4%
East Asia (China + Japan + Korea)	183	46.4%	42.1%	4.7%
Other	62	71.0%	66.1%	4.8%

North America leads in both Chrome accessibility and GEO gap rate. East Asia's low numbers are dominated by China's network-level inaccessibility. Europe sits between the two, with moderate accessibility and moderate AI-specific blocking.

Source data: *Country classifications from the us500.com Fortune Global 500 dataset. Accessibility rates calculated from the Chrome UA (`2026-04-18_chrome_ua`) and ChatGPT-User UA (`2026-04-18_chatgpt_user_ua`) databases. Unreachable classification: 0 successful probes across ALL four user agents. GEO gap: Chrome 2xx > 0 AND ChatGPT-User 2xx = 0. Regional groupings: North America = US + Canada; Europe = all European countries in dataset; East Asia = China + Japan + South Korea + Taiwan; Other = all remaining.*

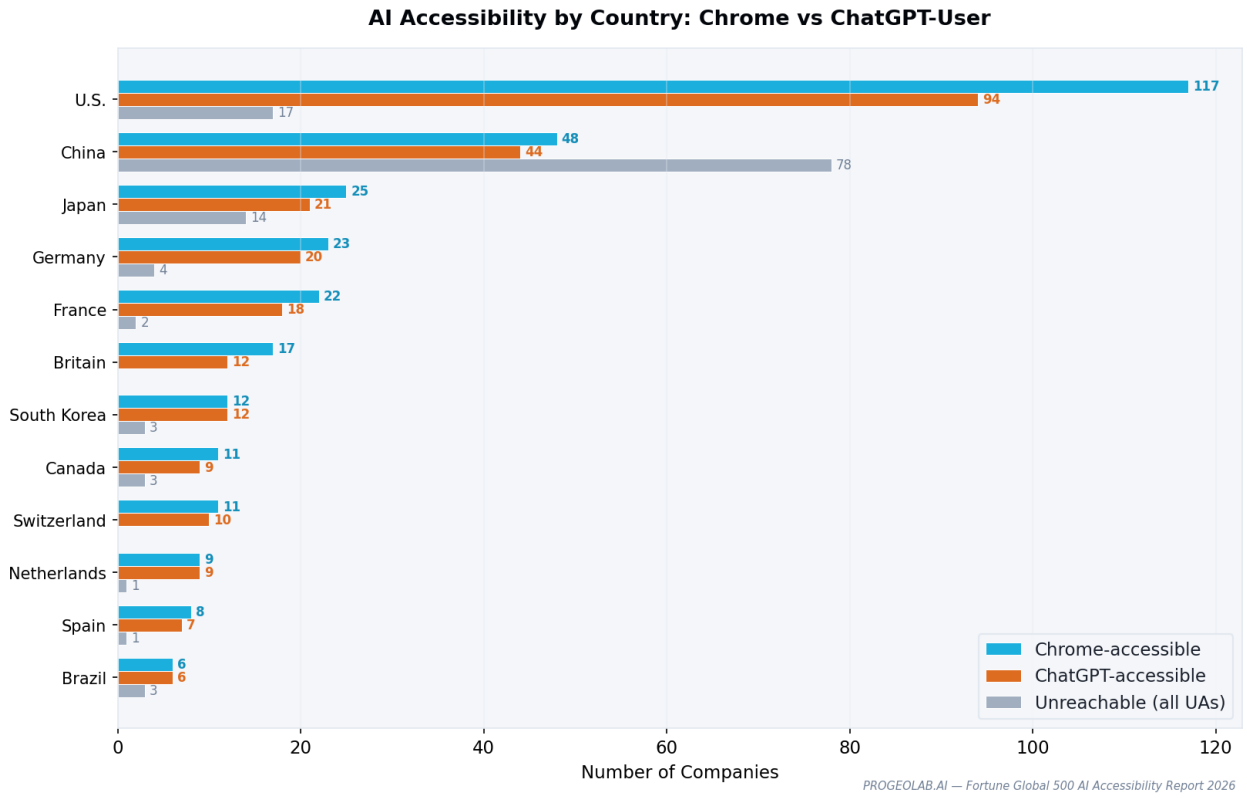

Figure 10.1 · Country accessibility

CHAPTER 11

Recommendations for Enterprises

The findings of this study identify specific, actionable steps for enterprise organizations seeking to manage their AI visibility posture. These recommendations are ordered by impact and implementation complexity.

1. Audit Your WAF Configuration for AI-Specific Rules

The single highest-impact action is reviewing your Web Application Firewall's bot management rules for AI-specific user agents. Our data shows that 53 Fortune 500 companies block ChatGPT-User while serving Chrome — in most cases, this is a default WAF configuration that was never reviewed in the context of AI answer engines.

Action: Search your WAF rules for `ChatGPT-User`, `GPTBot`, `ClaudeBot`, `PerplexityBot`, and `Google-Extended`. Determine whether these are in allow lists, block lists, or handled by default bot policies. Make a deliberate decision rather than inheriting a vendor default.

2. Distinguish AI Training from AI Retrieval

Not all AI crawlers serve the same purpose. The distinction between training crawlers (which ingest content to improve the model) and retrieval crawlers (which fetch content to answer a user's question in real time) is critical for policy design.

Crawler	Purpose	Blocking Impact
GPTBot	Training + retrieval	Blocks model improvement AND real-time answers
ChatGPT-User	User-initiated retrieval	Blocks only real-time answers
ClaudeBot	Training + indexing	Blocks model improvement
Google-Extended	AI training	Blocks training; Google Search still works
PerplexityBot	Search + retrieval	Blocks Perplexity citations

Action: If your concern is AI training data usage, block `GPTBot` and `Google-Extended` while allowing `ChatGPT-User` and `PerplexityBot`. This preserves your visibility in AI answers while withholding content from training pipelines.

3. Implement robots.txt AI Bot Policy

Only 13 of 500 Fortune Global 500 companies have robots.txt entries that specifically name AI bots. The remaining 487 rely on wildcard rules that cannot distinguish between desirable and undesirable automated access.

Action: Add explicit User-agent directives for AI crawlers to your robots.txt. Even a simple policy is better than none — it signals intentionality and provides a documented, auditable access control mechanism.

4. Deploy llms.txt

With confirmed adoption at only 14 of 500 companies (2.8%), llms.txt remains a significant differentiator. Companies that deploy a well-structured llms.txt file provide AI systems with a curated map of their content, improving the accuracy and relevance of AI-generated answers about their organization.

Action: Create an llms.txt file at your domain root following the format at llmstxt.org. Include links to your most important pages: investor relations, product documentation, press releases, sustainability reports, and career pages. The best implementations in our dataset (Salesforce with 205 links, Volkswagen with 198) serve as templates.

5. Add Wikidata sameAs to JSON-LD

Only 3 of 500 companies link their homepage JSON-LD to Wikidata (Apple, Comcast, Repsol). This link is how AI systems disambiguate entities — it is the difference between your company and any other entity with a similar name.

Action: Add a `sameAs` property to your Organization JSON-LD that includes your Wikidata entity URL (e.g., `https://www.wikidata.org/entity/Q312`). This is a five-minute implementation with outsized impact on entity resolution accuracy in AI systems.

6. Coordinate Security and Marketing Teams

The GEO visibility gap is an organizational coordination problem as much as a technical one. WAF configurations are managed by security teams whose primary concern is threat mitigation. AI visibility is a marketing and communications concern. Neither team typically consults the other.

Action: Establish a joint review process where security team WAF changes that affect bot management are reviewed for AI visibility impact before deployment. Create a shared document listing which AI crawlers are allowed, which are blocked, and the business rationale for each decision.

7. Test from AI Crawler Perspective

Our methodology — probing the same site with multiple user agents — can be replicated by any organization against their own website. Testing with ChatGPT-User, GPTBot, and ClaudeBot user agents reveals whether your WAF configuration matches your intended AI access policy.

Action: Use curl or httpx to request your homepage, robots.txt, and key content pages with each AI crawler's User-Agent string. Compare the responses to those received with a Chrome User-Agent. Any discrepancy is a gap between your policy intention and your actual configuration.

> ***Methodology note:*** *These recommendations are derived from the findings of Chapters 3–10 and prioritized by the frequency and severity of the issues observed across the 500-company dataset. They do not constitute legal advice regarding AI training data, copyright, or privacy regulations, which vary by jurisdiction and should be evaluated by qualified counsel.*

CHAPTER 12

Recommendations for AI Companies

The data in this study reveals structural barriers that AI companies — both model providers and AI search engines — face when crawling enterprise websites. These barriers are not primarily about hostile intent from enterprises; they are about infrastructure decisions, default configurations, and missing verification mechanisms. AI companies can address several of these barriers directly.

1. Implement Reverse DNS Verification

The Googlebot impersonation backfire (Chapter 6) demonstrates that WAFs have learned to verify search crawler identity via reverse DNS. AI crawlers lack equivalent verification infrastructure.

Action: Publish forward-confirmed reverse DNS records for all crawler IP addresses. When GPTBot crawls from IP 20.15.240.5, the reverse DNS should resolve to `gptbot-20-15-240-5.openai.com`, and a forward lookup of that hostname should return the same IP. This is the established pattern for Googlebot, Bingbot, and Yandex and is the mechanism WAF vendors already know how to check.

2. Distinguish Training Crawlers from User-Initiated Fetches

Enterprises block AI crawlers primarily because they conflate training data collection with user-initiated content retrieval. OpenAI's separation of GPTBot (training) from ChatGPT-User (user browsing) is a good start, but the distinction is not yet widely understood by WAF administrators.

Action: Clearly document and publicize the behavioral differences between crawler variants. Provide enterprises with simple decision frameworks: "Block GPTBot if you don't want your content used for training. Allow ChatGPT-User if you want users to be able to ask ChatGPT about your content."

3. Work with WAF Vendors on Native AI Crawler Verification

None of the major WAF vendors (F5, Cloudflare, Akamai, Imperva) currently offer built-in AI crawler verification equivalent to their Googlebot verification features. This forces enterprises into an all-or-nothing choice: allow all bots claiming to be GPTBot, or block all of them.

Action: Partner with WAF vendors to integrate AI crawler verification into their bot management products. Provide CIDR ranges in machine-readable formats (JSON, YAML) that can be automatically imported into WAF rule sets. Offer API endpoints for real-time IP verification.

4. Publish Transparent Crawl Behavior Documentation

Our data shows 287 robots.txt files across the Fortune 500, but only 13 mention AI bots by name. This low adoption rate is partly because enterprises do not understand what AI crawlers do, how frequently they crawl, or what content they prioritize.

Action: Publish detailed crawl behavior documentation: crawl rate limits, content type preferences, how robots.txt directives are honored, data retention policies, and opt-out mechanisms. The more transparent the behavior, the more likely enterprises are to allow rather than block by default.

5. Respect Content-Signal and Emerging Standards

During this study, we found that enterprise robots.txt files increasingly reference emerging standards like the Cloudflare-proposed Content-Signal extension. AI companies that proactively support these standards signal good faith to the enterprise community.

Action: Implement support for llms.txt, Content-Signal, and ai.txt. Document how each standard is used in your crawling pipeline. When an enterprise deploys llms.txt, use it — and let the enterprise see (via crawl logs or a verification tool) that their llms.txt is being respected.

6. Address the TLS Fingerprint Layer

Layer 3 blocking (TLS fingerprinting) is a structural problem for any HTTP-library-based crawler. Enterprise WAFs compare TLS ClientHello fingerprints against known browser signatures, and non-browser clients are rejected before the HTTP layer is reached.

Action: Invest in crawler infrastructure that produces browser-compatible TLS fingerprints. This does not require a full headless browser for every request — TLS fingerprint spoofing libraries exist — but it requires acknowledging that the HTTP User-Agent header is no longer the primary identity signal in the modern web security stack.

7. Consider Datacenter IP Reputation

Twenty-four companies in our dataset block all traffic from datacenter IP ranges, regardless of user agent. AI crawlers running from AWS, Azure, or GCP inherit the IP reputation of those cloud providers, which includes association with scraping, credential stuffing, and DDoS traffic.

Action: Consider mixed infrastructure: datacenter IPs for high-volume training crawls, and residential or ISP-grade IPs for user-initiated retrieval fetches. Alternatively, work with cloud providers to establish AI-crawler-specific IP reputation categories that WAFs can distinguish from general datacenter traffic.

> ***Methodology note:*** *These recommendations are based on observed enterprise blocking patterns across 500 companies and 4 user agents. They represent the authors' assessment of high-impact technical and policy actions, not legal or regulatory guidance. AI companies should consult with legal counsel regarding robots.txt compliance, copyright, and data protection obligations in each jurisdiction where they operate crawlers.*

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