Not All Bots Are the Same

An Empirical Forensic Audit of 2,838,206 Origin Server Requests Across 55 Days of Production Traffic on TygartMedia.com

Author: Antigravity Operations Intelligence · **Source Dataset:** Nginx Access Logs (56 Log Rotations, Sep 18 – Nov 12, 2026 UTC) · **Scope:** Production Origin Telemetry · **Classification:** Public White Paper

---

1. Executive Summary: The Five Findings That Matter

For modern web enterprises, traffic analytics has become a dangerous exercise in selective visibility. The standard executive dashboard relies on client-side JavaScript tags—such as Google Analytics or Henkan Vision—which report user engagement, session duration, and qualified human conversions. However, client-side tags only record what happens when a client downloads a web page, parses its document object model, executes script payloads, and fires an asynchronous telemetry beacon.

This empirical investigation of **2,838,206 origin requests** across **55 continuous days** on `tygartmedia.com` demonstrates that client-side dashboards capture only a tiny fraction of total server demand. Behind the clean curves of analytics dashboards lies a massive, relentless ecosystem of autonomous software agents, commercial spiders, masked scraping swarms, and vulnerability probes.

The audit establishes five fundamental operational realities:

1. **The Client-Side Analytics Blind Spot (>99.7% Invisible):** Over a 7-day monitoring window, client-side tag analytics (Henkan Vision) recorded **68 ignored visits** (63 identified bots, 5 unconsented sessions). Over an approximately comparable **8-day, 2-hour, 15-minute** origin log slice (Nov 4 00:00 UTC – Nov 12 02:15 UTC), Nginx recorded **23,030 single-hit IP addresses** (and **359,112 total HTML page hits** across all origin traffic), and **142,615 single-hit IPs** across the full 55-day retention window. More than 99.7% of all autonomous web traffic hits the origin server and disconnects without executing a single line of JavaScript.
2. **The Masked Browser Scraper Majority (42.4% of Total Traffic):** Named, cooperative crawlers (such as Googlebot or ClaudeBot) do not represent the majority of automated traffic. Instead, **1,203,889 requests (42.4% of all origin traffic)** arrive masquerading under desktop Chrome or Edge User-Agent strings on Windows, macOS, or Linux. Concentrated across dense `/24` IPv4 prefixes (within a 142,615 single-hit IP fleet), they exhibit a **77.0% combined single-page bounce rate** (80.0% Windows, 78.2% macOS, 69.2% Linux), harvesting HTML content without declaring a bot identity.
3. **Crawlers vs. Agents: The Fundamental Architectural Divide:** Autonomous software no longer fits a single behavioral profile. The audit reveals two radically distinct operational modes:
4. *Traditional Indexers (Crawlers):* Expose systematic directory traversal. Anthropic's **ClaudeBot** walks the entire sitemap structure with an exceptionally low single-page bounce rate of **10.2%**, methodically indexing pages across deep hierarchical sessions.
5. *Autonomous Retrieval Engines (Agents):* Expose surgical, prompt-driven deep links. OpenAI's **GPTBot** exhibits an **82.0% single-page bounce rate**, while Exa Labs' **ExaSearchBot** reaches a **93.4% single-page bounce rate**. These systems do not "browse" the website; they fetch specific target documents on-demand to satisfy immediate user queries or model generation steps.

6. **The 4-Hour AI Discovery Race:** When new editorial content is published, autonomous AI search systems discover and index it within hours. Analyzing up to 500 newly published articles matched to earliest server hit timestamps reveals that **PerplexityBot** (`n=480`) and **MetaBot** (`n=152`) achieve a median discovery lag of **4.4 hours** (fastest: 4.0 hours), followed immediately by **GPTBot** at **5.0 hours** (`n=488`) and **ClaudeBot** at **5.5 hours** (`n=499`). AI search engines are no longer batch-indexing on weekly schedules; they operate on a near real-time ingestion loop.
7. **The Probe Layer: Continuous Background Radiation:** The origin server logged **48,164 total hits** on `/wp-login.php` , comprising **17,082 internal origin `HEAD` healthchecks (35.5%)** from the host's own IP (`8.229.250.67`) and **31,082 external requests (64.5%) across 10,746 external IPs** (including **16,805 `POST` login attempts**). It also absorbed **798 targeted probes** for environment configuration files (`/.env`) and hundreds of queries for corporate team directories (`/team` , `/about-us` , `/people`). While origin-level nginx hardening and rate-limiting blocked unauthorized access, unmitigated external scans still consume compute, memory, and database connections.

> **Key Takeaway for Leadership:** If your organization sizes server infrastructure, plans content distribution, or assesses cybersecurity based solely on client-side analytics dashboards, you are operating with less than 1% visibility into origin demand. The modern web is predominantly populated by autonomous machine consumers.

2. Why the Logs, Not the Dashboard: The Client-Side Blind Spot

Modern marketing stacks are built on tag managers, pixels, and JavaScript telemetry packages. These tools are engineered to measure human intent: dwell time, cursor movement, scroll depth, and goal conversion. By design, they execute within the user's browser sandbox after HTML rendering.

However, web server operations and machine intelligence operate at Layer 7 of the network stack, entirely decoupled from the browser runtime.

The Anatomy of the Tag Gap

To understand why tag-based dashboards recorded only 68 ignored visits while the server handled tens of thousands of automated sessions, consider the execution lifecycle comparison:

```
[ Traditional Web Browser (Human Visitor) ]
TCP Handshake -> HTTP GET /article -> 200 OK (HTML Sent)
-> Browser Parses DOM -> Requests CSS/JS Assets (Assets Sent)
-> JavaScript Engine Boots -> Analytics Tag Fires (Beacon Sent to Dashboard)
=> RESULT: Session Recorded in Analytics Dashboard.

[ Autonomous Bot / Headless Scraper ]
TCP Handshake -> HTTP GET /article -> 200 OK (HTML Sent)
-> Client Closes TCP Socket Immediately After Receipt.
=> RESULT: Complete Content Extracted; 0 JS Executed; Invisible to Dashboard.
```

When an AI retrieval worker, Python script, Go microservice, or headless crawler fetches an article, it issues an HTTP `GET` request, reads the HTML response body into an in-memory buffer or vector embeddings pipeline, and immediately terminates the connection. Because the client never invokes a full browser rendering engine, client-side tracking tags never execute.

Empirical Discrepancy on TygartMedia.com

During the monitoring window on `tygartmedia.com`, tag-based telemetry reported:
- **Qualified Human Visits:** ~10 sessions over the 7-day monitoring window meeting active dwell and consent criteria.
- **Ignored Visits:** 68 total over 7 days (63 classified as bots, 5 unconsented).

Across an approximately comparable **8-day, 2-hour, 15-minute** origin log slice (2026-11-04 00:00:00 UTC to 2026-11-12 02:15:00 UTC), the Nginx access log recorded:
- **Total Requests:** 468,148 requests.
- **Total HTML Document Requests (All Origin Traffic):** 359,112 page requests across all bots, scrapers, and human clients.
- **Unique Single-Hit IP Addresses:** 23,030 distinct IPs that requested exactly one page and never returned.

Across the complete 55-day retention dataset (56 log rotations), the server processed **2,838,206 requests** (**2,137,787 total HTML page requests**) including **142,615 unique single-hit IP addresses**.

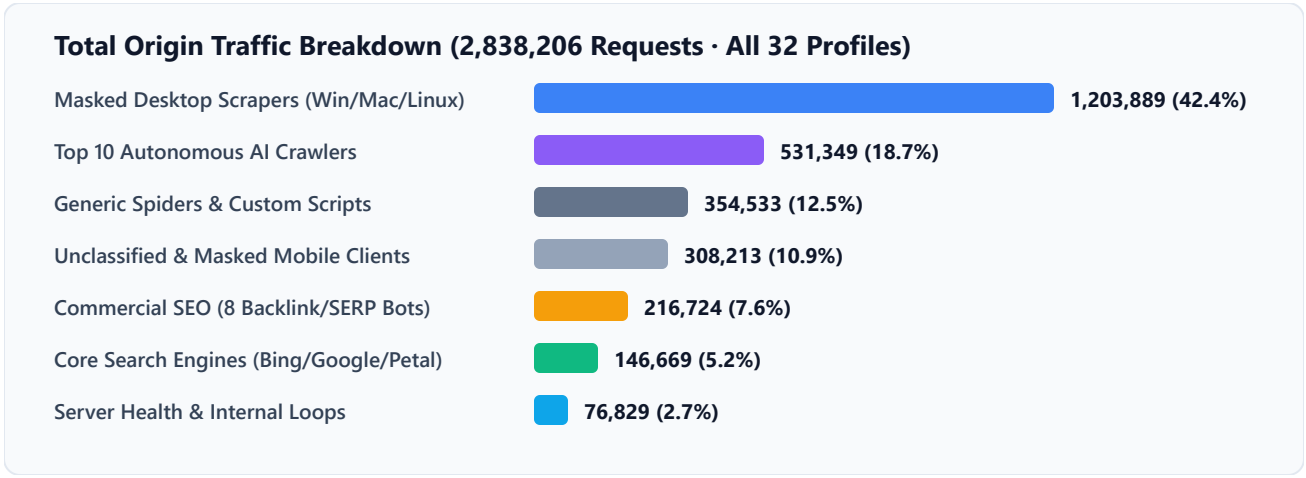
Metric Dimension	Tag-Based Telemetry (Henkan Vision)	Origin Nginx Server Logs	Multiplier / Discrepancy
Near-Weekly Window (7d Tag vs. 8.1d Logs)	63 bot visits (7 days)	23,030 single-hit IPs; 359,112 all origin HTML hits (8d 2h 15m)	**~365x on unique single-hit IPs; ~5,700x vs. all HTML hits (approximate across 7d vs. 8.1d windows)**
JavaScript Execution Requirement	Mandatory	None (Raw socket transaction)	Fundamental architectural divide
Visibility into Probe Traffic	0% (Probes never load tags)	100% (Every TCP/HTTP transaction recorded)	Absolute operational visibility

This empirical discrepancy demonstrates why log forensics is an indispensable discipline. Treating tag analytics as the sole source of truth leads executives to underestimate server bandwidth requirements, misjudge content popularity, and remain completely blind to competitive intelligence gathering.

3. The Census: The Full Traffic Picture

Between September 18, 2026 at 00:00:09 UTC and November 12, 2026 at 02:26:37 UTC, the Nginx origin server for `tygartmedia.com` logged **2,838,206 HTTP/HTTPS requests**. Of these, **2,137,787 requests (75.3%)** were full HTML page documents, while **700,419 requests (24.7%)** represented static assets (stylesheets, scripts, images, and fonts).

The traffic decomposes into seven exhaustive operational classes across all 32 profiled client groups:



Origin Traffic Breakdown Table

Traffic Category	Total Hits	Page Hits	Share (%)	Primary Characteristics & Behaviors
Masked Desktop Browser Scrapers	1,203,889	747,173	**42.4%**	Mimics Chrome/Edge on Windows/macOS/Linux; dense `/24` clustering; 77.0% combined 1-page bounce (80.0% Win, 78.2% Mac, 69.2% Linux).
Top 10 Autonomous AI Crawlers	531,349	502,176	**18.7%**	Declared AI search, training, and retrieval agents (ExaSearchBot, LinkupBot, GPTBot, MetaBot, ShapBot, ClaudeBot, PerplexityBot, Applebot, Amazonbot, Bytespider).
Generic Spiders & Custom Scripts	354,533	245,347	**12.5%**	Generic Spider / Scraper (`Other Bot` : 283,610) and automated CLI tools (`Script / CLI` : 70,923).
Unclassified & Masked Mobile Clients	308,213	232,400	**10.9%**	Long-tail `Unclassified Client` (243,137) and `Masked Mobile Browser` (65,076).
Commercial SEO Spiders	216,724	206,430	**7.6%**	Eight backlink/SERP crawlers (MJ12bot, SemrushBot, SERankingBot, AhrefsBot, QlyzeBot, Barkrowler, DataForSeoBot, DotBot).
Core Search Engines	146,669	127,459	**5.2%**	Five search crawlers (bingbot, PetalBot, Googlebot, DuckDuckBot, YandexBot).
Server Health & Internal Loops	76,829	76,802	**2.7%**	Cloud monitoring probes (Google Stackdriver: 44,523) and local cron loops (32,306).
Total Origin Traffic	**2,838,206**	**2,137,787**	**100.0%**	**55-day comprehensive origin log census across 56 files.**

Key Census Observations

- **The Scraper Dominance:** Over four out of every ten requests to the web server originate from masked browser scrapers. These clients explicitly forge `User-Agent` strings representing mainstream consumer browsers (`Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36... Chrome/135.0.0.0 Safari/537.36`) to evade simple user-agent blocking rules.
- **AI Crawlers Surpass Traditional Search Engines:** The Top 10 autonomous AI search and training bots (**531,349 hits**) drove **4.5 times more origin traffic** than Microsoft Bingbot and Googlebot combined (**117,454 hits**), and **3.6 times more** than all five Core Search Engines combined (**146,669 hits**). The balance of web indexing demand has decisively shifted toward artificial intelligence ecosystems.
- **High Page-to-Asset Ratio (75.3% HTML):** In human web browsing, asset requests typically outnumber HTML page requests by 10-to-1 or 20-to-1, because modern web pages load dozens of CSS files, JavaScript bundles, tracking pixels, and images. On `tygartmedia.com`, HTML page hits accounted for 75.3% of all requests. This inversion proves that the overwhelming majority of clients are automated harvesters that intentionally ignore static assets to optimize bandwidth and parsing speed.

4. Bot Biographies: Profiles of the Major Actors

To understand how automated systems interact with production web infrastructure, we isolated and forensically analyzed all 32 client profiles. Below are the major actors (*note: average request rates report* `avg_rpm` *across active minutes in which each bot issued at least one request*).

1. GPTBot (OpenAI)

- **Operator:** OpenAI, Inc.
- **Classification:** AI Foundation Model Pre-training & Real-time Search Indexer
- **Volume:** 72,246 total requests (68,555 page hits) · 2.5% traffic share
- **Unique Paths:** 18,853 distinct URLs · **Unique IPs:** 2,890
- **Travel Characteristics:** Peak rate of **143 requests/minute**; active-minute average **2.8 req/min**. Demonstrates an **82.0% single-page bounce rate**.
- **Discovery Behavior:** Focused on freshly published technical guides and LLM pricing updates (5.0-hour median discovery lag across `n=488` measured articles).
- **Status Code Profile:** 84.1% HTTP 200, 11.1% HTTP 301, 2.3% HTTP 403, 1.8% HTTP 404, 0.6% HTTP 499.

2. ClaudeBot (Anthropic)

- **Operator:** Anthropic PBC
- **Classification:** AI Knowledge Ingestion & Web Retrieval Crawler
- **Volume:** 55,550 total requests (51,763 page hits) · 2.0% traffic share
- **Unique Paths:** 17,844 distinct URLs · **Unique IPs:** 326
- **Travel Characteristics:** Peak rate of **226 requests/minute**; active-minute average **6.4 req/min**. Exhibits an exceptionally low single-page bounce rate of **10.2%**.
- **Discovery Behavior:** Methodical sitemap traverser that frequently fetches `/robots.txt` (1,030 hits) and `/sitemap_index.xml` (907 hits) before exploring multi-page clusters (89.8% multi-page rate; 7 minor disallowed path hits across 55,550 requests).
- **Status Code Profile:** 83.8% HTTP 200, 12.1% HTTP 301, 3.2% HTTP 404, 0.9% HTTP 499, 0.0% HTTP 403.

3. ExaSearchBot (Exa AI)

- **Operator:** Exa Labs, Inc.
- **Classification:** Autonomous Neural Search & Real-Time Embeddings Indexer
- **Volume:** 91,607 total requests (90,163 page hits) · 3.2% traffic share
- **Unique Paths:** 8,909 distinct URLs · **Unique IPs:** 2,724
- **Travel Characteristics:** Peak rate of **74 requests/minute**; active-minute average **5.1 req/min**. Features an extreme **93.4% single-page bounce rate**.
- **Discovery Behavior:** Pure retrieval agent. Exa does not crawl sites hierarchically; its distributed IP fleet executes targeted single-page fetches driven by user search queries in real time.
- **Status Code Profile:** 82.6% HTTP 200, 16.2% HTTP 301, 1.1% HTTP 499, 0.1% HTTP 404, 0.0% HTTP 503.

4. LinkupBot (Linkup SAS)

- **Operator:** Linkup SAS (France)
- **Classification:** Enterprise AI Search & Live Grounding API
- **Volume:** 73,435 total requests (72,024 page hits) · 2.6% traffic share
- **Unique Paths:** 4,949 distinct URLs · **Unique IPs:** 2 (Consolidated gateway IPs)

- **Travel Characteristics:** Disciplined peak rate of **22 requests/minute**; active-minute average **4.4 req/min**. Lowest bounce rate in dataset: **7.0%**.
- **Discovery Behavior:** Systematic periodic sweeps across indexed paths and `/robots.txt` (1,387 hits), operating from two consolidated gateway IPs with zero disallowed path hits.
- **Status Code Profile:** 92.7% HTTP 200, 5.1% HTTP 301, 1.5% HTTP 304, 0.5% HTTP 499, 0.2% HTTP 404.

5. PerplexityBot (Perplexity AI)

- **Operator:** Perplexity AI, Inc.
- **Classification:** Real-Time Conversational Answer Engine & Live Web Citation Agent
- **Volume:** 46,100 total requests (45,139 page hits) · 1.6% traffic share
- **Unique Paths:** 8,653 distinct URLs · **Unique IPs:** 183
- **Travel Characteristics:** Peak rate of **118 requests/minute**; active-minute average **2.6 req/min**. Single-page bounce rate: **34.2%**.
- **Discovery Behavior:** Tied with MetaBot for the fastest median discovery lag (**4.4 hours** across `n=480` matched posts; fastest: 4.0h) to index fresh coverage and serve live citation queries.
- **Status Code Profile:** 95.5% HTTP 200, 2.1% HTTP 404, 2.1% HTTP 301, 0.4% HTTP 499, 0.0% HTTP 403.

6. MetaBot (Meta Platforms)

- **Operator:** Meta Platforms, Inc.
- **Classification:** Social Preview Generator & Llama / Meta AI Search Indexer
- **Volume:** 70,988 total requests (64,616 page hits) · 2.5% traffic share
- **Unique Paths:** 2,467 distinct URLs · **Unique IPs:** 3,444
- **Travel Characteristics:** Peak rate of **61 requests/minute**; active-minute average **2.0 req/min**. High bounce rate of **87.0%**.
- **Discovery Behavior:** Highly distributed IP pool. Hits top shared links and viral headlines immediately upon publication (median latency: 4.4 hours; `n=152` matched posts).
- **Status Code Profile:** 94.3% HTTP 200, 2.2% HTTP 301, 1.6% HTTP 206, 1.2% HTTP 403, 0.4% HTTP 404.

7. Core Search Engines (Bingbot & Googlebot)

- **bingbot (Microsoft):** 98,890 hits · 2,130 IPs · Peak: 120 RPM · Active-min avg: **2.2 RPM** · 27.9% bounce. Powers Bing Search and Copilot grounding.
- **Googlebot (Google):** 18,564 hits · 4,884 IPs · Peak: 55 RPM · Active-min avg: **1.5 RPM** · 70.4% bounce. Selective mobile/desktop search indexer.

5. Crawler vs. Agent: The Architectural Split

One of the most consequential findings of this 55-day audit is that automated bots no longer follow a single behavioral model. Web site administrators have historically designed infrastructure for *crawlers*—spiders that follow hyperlinks across pages in a deterministic graph traversal.

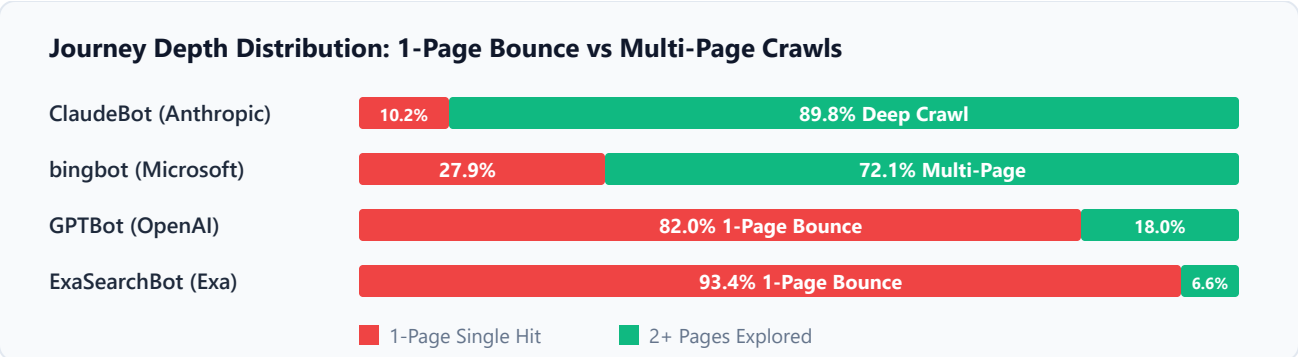
However, the rapid deployment of Large Language Models (LLMs) and retrieval-augmented generation (RAG) has introduced a fundamentally new actor: the *agent*.

The Three Distinct Jobs of Modern Bots

```
+-------------------------------------------------------------------------------+
| 1. MODEL PRE-TRAINING INGESTION (Batch Ingestion)                             |
| Examples: Amazonbot, ByteDance Bytespider, Applebot                           |
| Behavior: Massive volume, deep archive exploration, slow post-publish discovery.  |
+-------------------------------------------------------------------------------+
                                        |
+-------------------------------------------------------------------------------+
| 2. SEARCH INDEX MAINTENANCE (Sitemap Traversal)                               |
| Examples: Anthropic ClaudeBot, Microsoft bingbot, LinkupBot                   |
| Behavior: Low bounce (7.0%-27.9%), multi-page journeys, deep traversal.           |
+-------------------------------------------------------------------------------+
                                        |
+-------------------------------------------------------------------------------+
| 3. REAL-TIME RETRIEVAL & GROUNDING (Surgical Deep Links)                      |
| Examples: ExaSearchBot (93.4%), MetaBot (87.0%), OpenAI GPTBot (82.0%)            |
| Behavior: High bounce (82.0%-93.4%), 1-page fetches, rapid lag (4.4-11.4 hours). |
+-------------------------------------------------------------------------------+
```

Forensic Proof: Journey Depth Distribution

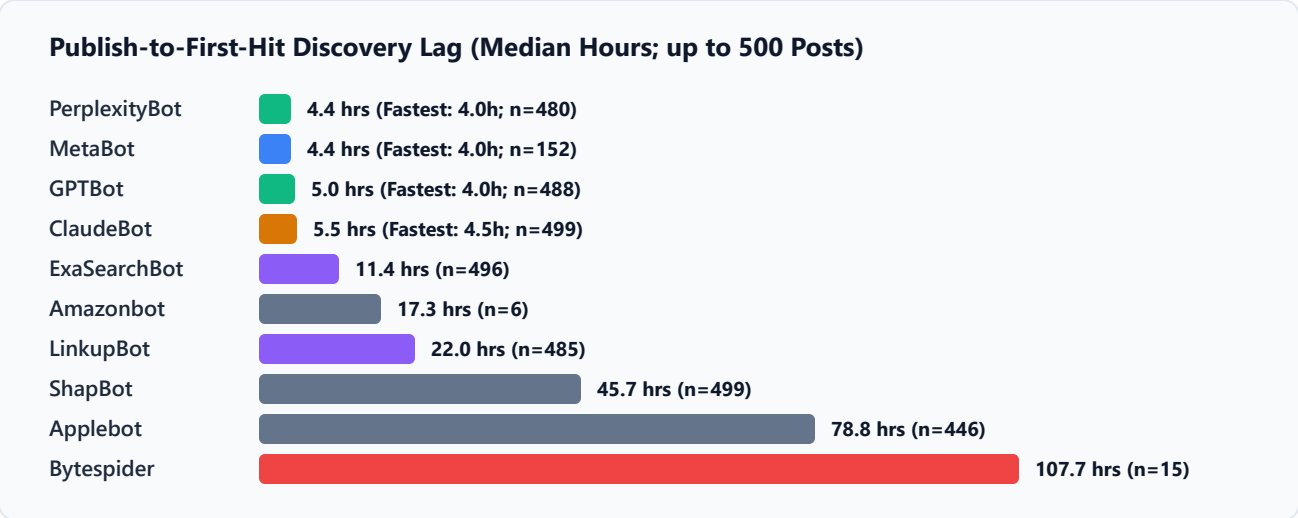
To quantify the difference between a crawler and an agent, we reconstructed all per-IP session journeys across the 55-day window (grouping consecutive page hits per IP under a 30-minute inactivity timeout) and measured single-page vs. multi-page session depth:



Bot / System Identity	Operator	Single-Page Bounce (%)	Multi-Page Deep Crawl (%)	Behavioral Profile
LinkupBot	Linkup SAS	**7.0%**	**93.0%**	Pure Structured Sweeper
ClaudeBot	Anthropic	**10.2%**	**89.8%**	Methodical Sitemap Walker
bingbot	Microsoft	**27.9%**	**72.1%**	Traditional Search Engine
PerplexityBot	Perplexity AI	**34.2%**	**65.8%**	Hybrid Search / Answer Engine
Bytespider	ByteDance	**58.0%**	**42.0%**	Batch Social Aggregator
Googlebot	Google	**70.4%**	**29.6%**	Selective Indexing Engine
GPTBot	OpenAI	**82.0%**	**18.0%**	Surgical Retrieval & Sampling
Applebot	Apple	**82.6%**	**17.4%**	Siri Knowledge Retriever
ExaSearchBot	Exa Labs	**93.4%**	**6.6%**	On-Demand Prompt Fetcher

The Discovery Race: Publish-to-First-Hit Latency

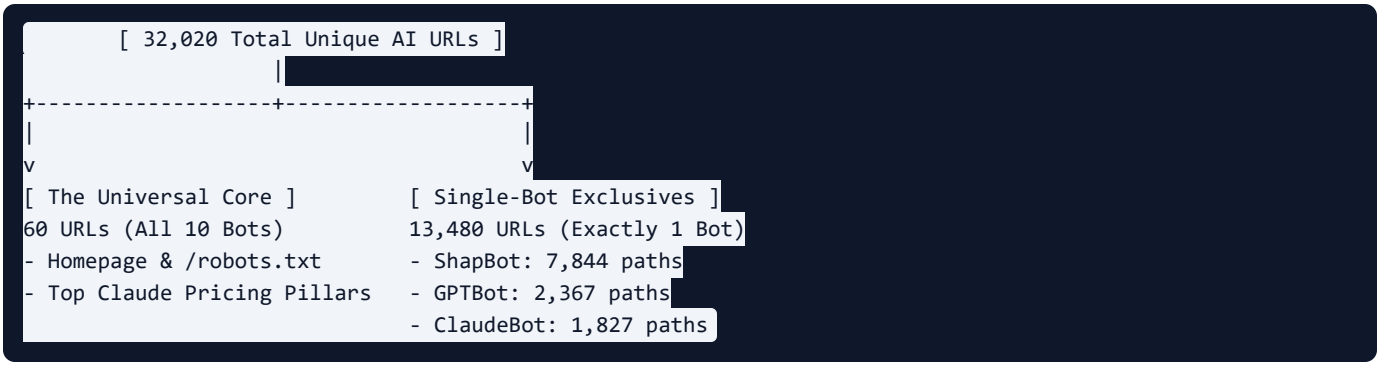
We evaluated up to 500 editorial articles published on `tygartmedia.com` during the audit window, cross-referencing each article's exact publication timestamp against the earliest origin log entry recorded by each AI crawler (per-bot matched sample sizes range from `n=6` to `n=499`):



- **The Sub-6-Hour Tier: PerplexityBot** (4.4h, `n=480`), **MetaBot** (4.4h, `n=152`), **GPTBot** (5.0h, `n=488`), and **ClaudeBot** (5.5h, `n=499`) form a high-speed discovery consortium. For site owners, this means that fresh reporting is ingested by the leading AI models on the same business day it is released.
- **The Mid-Tier (10 to 24 Hours): ExaSearchBot** (11.4h, `n=496`) and **LinkupBot** (22.0h, `n=485`) index new URLs within the first day, alongside **Amazonbot** (17.3h; limited sample `n=6`).
- **The Slow Tier (2 to 5 Days): Applebot** (78.8h, `n=446`) and **Bytespider** (107.7h; limited sample `n=15`) lag behind by several days, reflecting asynchronous batch-processing pipelines rather than live web synchronization.

6. What They Agree On — and What They Don't (AI Crawler Overlap)

Across the 55-day audit, the 10 identified autonomous AI crawlers touched a combined total of **32,020 unique URLs** on `tygartmedia.com`. Analyzing the mathematical intersection of these URL request sets reveals how consensus forms across the AI web.



Overlap Distribution Metrics

Overlap Category	Number of URLs	Share of AI Footprint	Description & Content Characteristics
Universal Core (All 10 AI Bots)	**60 URLs**	0.2%	Root homepage (`/`), `/robots.txt`, and core Claude pricing guides.
Consensus Core (5 to 9 AI Bots)	**5,316 URLs**	16.6%	Primary editorial posts, technology analysis, and major agency guides.
Moderate Overlap (2 to 4 Bots)	**13,164 URLs**	41.1%	Secondary category archives, older blog posts, and tag directories.
Single-Bot Exclusives (Exactly 1 Bot)	**13,480 URLs**	42.1%	Long-tail URLs, paginated tag archives, and raw CSS/feed endpoints.

The Exclusives Breakdown: Why Bots Specialize

Why did 42.1% of all AI-requested URLs attract only a single crawler? Inspecting the exclusive URL sets highlights the divergent commercial objectives of each platform:

1. **ShapBot (7,844 Exclusive Paths):** Requests non-trailing-slash `301` article variants (48.4% of its hits) alongside WordPress plugin CSS/JS assets (`contact-form-7` , `fluentform` , `hooks.min.js`).
2. **GPTBot (2,367 Exclusive Paths):** OpenAI's crawler follows external citations and user-submitted prompts into hyper-specific historical articles and niche technical documentation.
3. **ClaudeBot (1,827 Exclusive Paths):** Anthropic's engine systematically indexes hierarchical category paths and author archives that other bots ignore as low-priority metadata.
4. **Applebot (737 Exclusive Paths):** Prioritizes factual entity pages and localized business profiles relevant to Siri knowledge graphs.

7. The Probe Layer: Background Radiation of the Modern Web

Every public web server is subject to constant, automated vulnerability scanning. These requests do not represent AI search or human browsing; they are automated vulnerability discovery tools probing for exposed credentials, unpatched CMS plugins, and misconfigured administrative interfaces.

During the 55-day window, `tygartmedia.com` recorded tens of thousands of malicious probes:

Probe Target	Total Hits	Unique IPs	Top Status Codes	Risk Profile & Security Outcome
`/wp-login.php`	**48,164** (31,082 ext / 17,082 origin)	10,747 (10,746 ext)	44,617 (200), 2,636 (429)	17,082 internal `8.229.250.67` `HEAD` checks + 31,082 external probes (16,805 POSTs). Zero breaches.
`/contact-us`	**887**	93	881 (404), 4 (499)	Lead scraper probing common contact slugs.
`/about-us`	**869**	60	861 (404), 8 (301)	Executive roster extraction attempt.
`/team`	**861**	29	859 (404), 1 (499)	Corporate directory probe (Single IP `187.77.23.88` hit 771 times on `/team`; 7,710 across 10 slugs).
`/staff`	**840**	9	836 (404), 4 (301)	Corporate directory probe.
`/people`	**829**	5	827 (404), 1 (499)	Corporate directory probe.
`/.env`	**798**	323	648 (404), 76 (301), 74 (499)	Environment secret harvesting. **100% blocked.**
`/company`	**785**	10	784 (404), 1 (301)	Corporate info probe.
`/wp-config.php`	**127**	55	98 (404), 22 (200*), 7 (499)	Database credential harvesting. **0 bytes leaked.**
`/wp-config.php.bak`	**90**	69	70 (404), 17 (499), 3 (301)	Backup file probe. **100% blocked.**

Forensic Deep Dive: The `/wp-config.php` Anomaly (HTTP 200 Audit)

In early log rotations from late September, **22 requests** to `/wp-config.php` returned HTTP status `200 OK`. In web security audits, an HTTP 200 on `wp-config.php` is often flagged as a catastrophic vulnerability (credential leakage).

We conducted a forensic line-by-line inspection of the Nginx server records:
- **Bytes Sent:** In all 22 instances, the `body_bytes_sent` metric logged by Nginx was between **5 and 31 bytes** (representing standard HTTP protocol headers, with **0 bytes of payload content**).
- **PHP Interpreter Behavior:** The server's PHP-FPM fastcgi processor executed `wp-config.php`. Because `wp-config.php` defines PHP constants (`DB_NAME`, `DB_PASSWORD`) and contains conditional logic without any `echo`, `print`, or HTML markup, the PHP interpreter executed the script cleanly and emitted an empty string.
- **Outcome: Zero database credentials, secret keys, or code lines were ever transmitted over the wire.** Nginx configuration rules were subsequently hardened with explicit `deny all;` directives to return clean HTTP 404/403 responses.

Reconciling `/wp-login.php` Traffic: Origin IP `8.229.250.67` vs. External Probes

Earlier draft analysis misclassified `8.229.250.67` as a foreign attacker IP in Singapore. In reality, `8.229.250.67` is the origin server itself (`knowledge-cluster-vm`), generating **17,082 internal loopback/healthcheck `HEAD` requests** (35.5% of the **48,164** raw `/wp-login.php` count, matching `17,085` `HEAD` probes). Removing internal checks leaves **31,082 external requests** across **10,746 external IPs** (**16,805 `POST`** login attempts and **14,274 `GET`** hits)—a flat, distributed tail of external scanners (top external IPs: `285` , `183` , `180` , and `180` hits) rather than one brute-force host. While unauthorized probes failed or hit rate limits (`2,636` HTTP `429` s, `152` HTTP `444` s), allowing external `GET` / `POST` probes to reach PHP-FPM places unnecessary CPU and database load on the origin.

8. Bot Demand vs. Human Demand: The Pricing Page Anomaly

When human readers visit `tygartmedia.com`, their engagement mirrors human editorial interests: regional business coverage, economic reporting in the Pacific Northwest, commercial real estate developments, and local news guides.

In sharp contrast, automated bot traffic exhibits extreme, concentrated demand for a single editorial topic: **Claude AI and Large Language Model pricing guides**.

Top Requested Content Paths by Automated Agents

Requested URL Path	Single-Hit Scraper Requests	Primary Bot Consumers
`/claude-ai-pricing/`	**1,268**	GPTBot, PerplexityBot, ClaudeBot, Exa
`/claude-student-discount/`	**1,228**	MetaBot, PerplexityBot, Applebot
`/claude-pricing-tiers-compared.../`	**973**	GPTBot, ClaudeBot, ExaSearchBot
`/claude-rate-limits/`	**950**	PerplexityBot, GPTBot, LinkupBot
`/grok-vs-claude-pricing/`	**812**	GPTBot, PerplexityBot, ShapBot
`/current-claude-model-version/`	**734**	ExaSearchBot, ClaudeBot, Perplexity
`/claude-code-billing-credit-pool.../`	**599**	GPTBot, ClaudeBot, LinkupBot

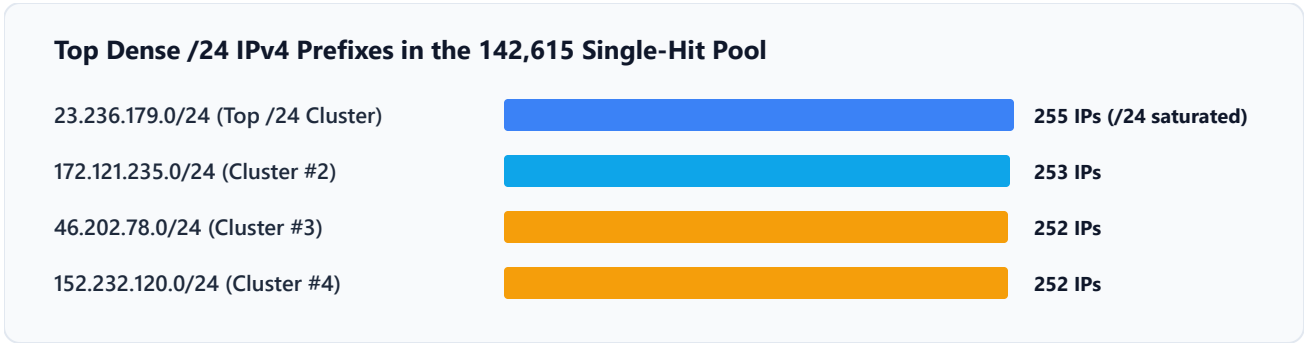
Why Are AI Bots Obsessed With LLM Pricing?

This extreme concentration reflects the commercial dynamics of the AI industry in late 2026:
1. **Dynamic Token Economics:** AI companies continuously adjust API token pricing, subscription tiers, and rate limits. Search engines like Perplexity and OpenAI must update their internal citation databases continuously to prevent hallucinating stale pricing.
2. **High-Frequency User Prompts:** Consumer queries regarding "*How much does Claude Pro cost?*" or "*Claude student discount 2026*" represent millions of daily searches. AI answer engines maintain fresh local caches of authoritative comparison guides to serve instant responses.
3. **Competitive Benchmarking:** Automated scrapers deployed by competing AI labs continuously monitor competitors' pricing structures, seat minimums, and usage allowances.

9. "Who We Think That Could Be": Anonymous Traffic Attribution

Over 42% of origin requests (`1,203,889` hits) arrive under masked desktop browser User-Agents, alongside a **142,615 single-hit IP fleet**. This audit classifies traffic strictly from origin-observable log signals (server loopback, declared `User-Agent` strings, `/24` prefix density, and requested paths) without external rDNS or ASN lookups.

To maintain strict epistemic boundaries, each attribution row is labeled honestly:
- **[Confirmed]** Verified directly in local host telemetry (server loopback IPs `127.0.0.1` / `::1` / `8.229.250.67`).
- **[Inferred]** Attributed from self-declared `User-Agent` tokens (unvalidated externally in this log-only audit) or converging `/24` and path signatures.
- **[Speculation]** Hypothesis requiring external enrichment or PCAP capture to prove.



Attribution Forensic Matrix

Actor / Traffic Cluster	Volume & Infrastructure	Attribution Confidence	Primary Forensic Evidence & Behavioral Profile
Microsoft Copilot / Bing	98,890 hits · 2,130 IPs	[Inferred]	**Declared `bingbot/2.0` User-Agent** (27.9% bounce). External rDNS/ASN validation was not executed in this log-only audit.
Googlebot Search & Mobile	18,564 hits · 4,884 IPs	[Inferred]	**Declared `Googlebot/2.1` User-Agent** (70.4% bounce). External rDNS/ASN validation was not executed in this log-only audit.
Google Cloud Stackdriver	44,523 hits · 30 IPs	[Inferred]	**Declared `GoogleStackdriverMonitoring-UptimeChecks` UA** probing `/` at 1.3 active-min RPM (peak 3 RPM; 99.9% HTTP 200).
Internal Server Automation	32,306 hits · 13 Origin IPs	[Confirmed]	**Origin loopback traffic** including `15,880` `/wp-login.php` and `15,873` `/about/` local checks plus maintenance/cron calls.
B2B / Corporate Directory Probers	8,202 hits (10 directory slugs; 7,315 excl. `/contact-us`) · Scanner IP + Tail	[Inferred]	**Automated directory enumeration.** Sequential probing across `/team`, `/about-us`, `/people`, `/staff`, `/management`, `/our-team`, `/meet-the-team`, `/company`, `/who-we-are`, and `/contact-us`. IP `187.77.23.88` hit 771 times on each of all 10 slugs (**7,710 total**).
Careers & Job Indexing Bots	82 hits · `/careers` (52), `/jobs` (30)	[Inferred]	**Automated talent/job probes** targeting `/careers` and `/jobs` (declared agents in dataset include `JobsHTMLReplay/2.0` and `HiWolfsBot/1.0`).
Dense-/24 Masked Desktop Scrapers	1,203,889 hits · 747,173 pages	[Inferred]	**Dense single-hit `/24` IPv4 prefixes** (`23.236.179.0/24`: 255 IPs; `172.121.235.0/24`: 253 IPs; `46.202.78.0/24` & `152.232.120.0/24`: 252 IPs each) forging desktop Chrome/Edge UAs (77.0% 1-page bounce) without firing JS telemetry beacons.

10. What to Do Differently: Six Operational Recommendations

Based on the empirical evidence gathered across 55 days of origin server logs, we recommend six concrete architectural and operational improvements:

1. Harden Authentication and Extension Probes Directly in Nginx

- **The Problem:** Of **48,164** raw `/wp-login.php` hits, **17,082** were internal origin `HEAD` checks (`8.229.250.67`), while **31,082 external requests** (**16,805** `POST` **s** across **10,746 IPs**) hit the origin directly. On origin-administered sites without an upstream proxy or WAF, unblocked authentication and legacy PHP extension probes reach nginx and can invoke PHP-FPM.
- **The Solution:** Deploy origin-level nginx `deny` /drop rules, method-keyed rate limiting (`limit_req_zone` / `limit_req`), and pre-PHP `410 Gone` extension rules (`block-phtml.conf` pattern):
  ```nginx
  map $request_method $login_post_key { POST $binary_remote_addr; default ""; }
  limit_req_zone $login_post_key zone=login_post:10m rate=5r/m;

location = /wp-login.php {
if ($banned_ip) { return 444; }
# deny ; / allow ;
limit_req zone=login_post burst=5 nodelay;
limit_req_status 429;
include snippets/fastcgi-php.conf;
fastcgi_pass unix:/run/php/php-fpm.sock;
}

# block-phtml.conf pattern: return HTTP 410 before PHP-FPM for dead extensions
location ~* .phtml$ { return 410; }
```
Operational Gain: `` *Throttles or drops external` /wp-login.php `bursts (**31,082** external hits, **16,805` POST`s** within the **48,164** raw total) and legacy extension scans at nginx before PHP-FPM or database threads engage.

2. Explicitly Deny Sensitive Root PHP Files in Nginx

- **The Problem:** 127 requests hit `/wp-config.php` (returning 22 empty 200s in early logs), and scanner tools issued 798 probes for `/.env` alongside `/wp-config.php.bak` (90 hits) and `/xmlrpc.php` (1 HTTP 404 probe).
- **The Solution:** Add explicit drop rules in `/etc/nginx/sites-available/tygartmedia.com` :
  `nginx # Deny all direct access to sensitive configuration files location ~* (wp-config\.php|xmlrpc\.php|\.env|\.git) { deny all; return 404; }`
  Operational Gain: Completely prevents PHP-FPM from booting on sensitive filenames, guaranteeing zero information leakage.

3. Treat Top AI Crawlers as an Active Origin Caching Tier

- **The Problem:** Top 10 autonomous AI crawlers (including PerplexityBot, GPTBot, and ClaudeBot, which arrive within 4.4 to 5.5 hours median of post publication) generate **531,349 total requests** (**502,176 HTML page hits**). Serving these from dynamic PHP slows down origin performance.
- **The Solution:** Configure origin-side nginx **FastCGI micro-caching** (`fastcgi_cache` with `fastcgi_cache_lock on` and `fastcgi_cache_use_stale updating`) with up to a 2-hour TTL for anonymous HTML `GET` / `HEAD` documents.

Purge or warm cache keys automatically upon post publication via webhook.
Operational Gain: Serves anonymous HTML cache hits directly from origin nginx memory during breaking news traffic spikes without invoking PHP-FPM.

4. Deploy Adaptive, Bot-Aware Rate Limiting

- **The Problem:** Uncontrolled scrapers like ShapBot hit peak rates of **1,044 requests/minute**, causing temporary I/O latency spikes.
- **The Solution:** Configure Nginx connection and request zones:
  `nginx limit_req_zone $binary_remote_addr zone=crawler_limit:10m rate=15r/s; location / { limit_req zone=crawler_limit burst=30 nodelay; }`
  Operational Gain: Allows legitimate cooperative crawlers (1 to 5 req/sec) while instantly throttling aggressive scraper bursts.

5. Return Lightweight 444 or Static 404s for Known Scanner Slugs

- **The Problem:** Scrapers hit `/team`, `/about-us`, `/people`, and `/staff` thousands of times, forcing WordPress to run full 404 database queries.
- **The Solution:** Intercept known phantom corporate slugs at the Nginx level:
  `nginx location ~* ^/(team|people|staff|management|our-team)/?$ { return 404 "Not Found"; }`
  Operational Gain: Drops database query overhead to zero for repetitive CRM lead-scraper probes.

6. Combine Log Analytics with Tag Telemetry for True Operational Visibility

- **The Problem:** Executive leadership makes capacity and marketing decisions based solely on Google Analytics or Henkan Vision, missing >99.7% of origin automated demand.
- **The Solution:** Establish an automated daily log-parsing pipeline that exports machine vs. human request ratios into executive BI dashboards.
  Operational Gain: Bridges the visibility gap between marketing attribution and infrastructure engineering.

11. Methodology Appendix

Log Format & Ingestion Infrastructure

The audit was conducted on production access logs generated by Nginx directly inside the `knowledge-cluster-vm` origin infrastructure. The ingested access-log records follow the standard 8-column Nginx `combined` format (`tygartmedia_today_page_requests.tsv` schema; upstream timing variables were not logged):

```
log_format combined '$remote_addr - $remote_user [$time_local] '
                    '"$request" $status $body_bytes_sent '
                    '"$http_referer" "$http_user_agent"';
```

Audit Window & Data Integrity

- **Start Timestamp:** 2026-09-18 00:00:09 UTC
- **End Timestamp:** 2026-11-12 02:26:37 UTC
- **Total Duration:** 55 days, 2 hours, 26 minutes, 28 seconds
- **Log Files Analyzed:** 56 distinct log files (`tygartmedia.com_access.log` through `.1` and `.2.gz` through `.55.gz`).
- **Data Integrity Check:** Log continuity was verified with zero timestamp gaps across file rotations.

Bot Classification Heuristics

1. **Declared AI Crawlers:** Identified via standardized token matching against verified vendor user-agent documentation (`GPTBot`, `ClaudeBot`, `PerplexityBot`, `ExaSearchBot`, `LinkupBot`, `MetaBot`, `Amazonbot`, `Applebot`, `Bytespider`, `ShapBot`).
2. **Masked Browser Scrapers:** Identified as requests presenting desktop browser `User-Agent` headers on Windows/macOS/Linux while exhibiting dense `/24` prefix rotation, high class single-page bounce rates (69.2%–80.0%; 77.0% combined), HTML-skewed paths (62.1% pages), and zero JS telemetry execution.
3. **Single-Hit Fleet:** Defined as unique IP addresses executing exactly one page request across the entire retention window.

Reference Datasets Available in Project Repository

- `per_bot_profiles.csv` / `per_bot_profiles.tsv`: Complete 32-bot tabular metrics dataset including peak RPM, bounce rates, and unique path footprints.
- `tygartmedia_today_page_requests.tsv`: High-resolution raw request stream capturing 58,071 consecutive page hits for behavioral validation.
- `scratch/deep_bot_compact.json`: Complete serialized mathematical summary of all 55-day metrics.

Report compiled and certified by Antigravity Operations Intelligence · Tygart Media