PRACTITIONER RESEARCH WHITE PAPER

Training Crawlers vs. Live AI Agents: 25:1 Origin Gap

What 24.92x the Traffic Means for robots.txt,
Caching, and Real-Time Redirect Latency

Tygart Media · tygartmedia.com · Anonymized Multi-Property Telemetry Edition

24.92x	279.67 vs 1.15	0.00%	37.56%
Bulk-to-Live Ratio	**Paths/IP**	**Live-Agent 404 Rate**	**Claude-User 301 Tax**
21,932 vs 880	GPTBot vs ChatGPT-User	0/907	Redirect latency

Executive Summary

Treating all AI bots as a single monolithic category obscures two completely opposite server behaviors. Based on a comprehensive origin-log study of 225,276 verified requests across eight WordPress properties, we found that four major bulk indexers (GPTBot, ClaudeBot, Meta-ExternalAgent, Bytespider) generate 21,932 hits—24.92x the volume of five major live answer-time user agents (880 hits).

This staggering 24.92x ratio demonstrates that training and search crawlers place a fundamentally different load on origin infrastructure compared to real-time prompt fetchers. While bulk indexers consume massive bandwidth across thousands of paths in parallel, live citation fetchers arrive surgically to retrieve exactly what a user has prompted them to read.

1. Why Blocking Training Bots and Blocking Live Fetchers Are Opposite Decisions

When publishers configure their firewalls and robots.txt files, they often fail to separate bulk training ingestion from answer-time citation retrieval. Blocking training bots limits the unlicensed scraping of your content corpus, but blocking live fetchers ensures your brand cannot be cited or verified when a user explicitly asks an AI agent about your product.

Bulk indexers sweep broadly with 1.24% 404s and 4.46% 304 caching, imposing high crawl-budget costs. Live answer-time agents arrive with minimal footprint (median 1 path), hit 0.00% 404s, rarely cache, and absorb painful 301 redirect latency that directly hurts the end-user's response time.

KEY TAKEAWAY FOR PUBLISHERS & CDN TEAMS

A one-size-fits-all AI bot blocking strategy is dangerous. Rate-limit or block bulk indexers to protect your crawl budget and IP, but ensure live prompt-time fetchers have a fast, unimpeded path to your server-rendered HTML or Markdown endpoints.

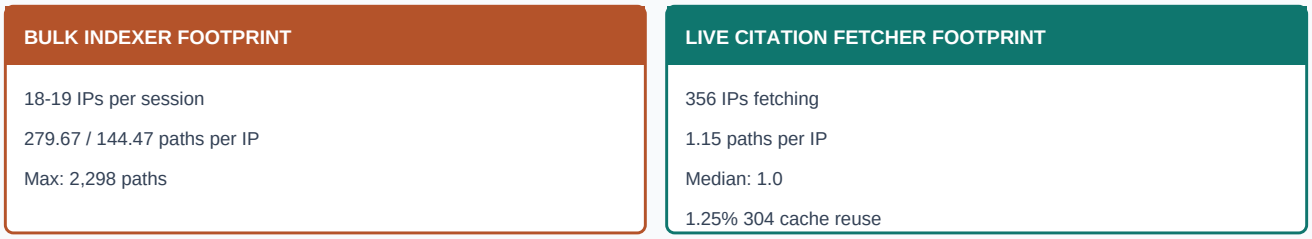
2. Session Breadth & IP Concentration Forensics

Bulk indexers operate via massive coordinated sweeps from a tightly clustered set of IP addresses. For example, GPTBot and ClaudeBot route their massive sweeps through just 18-19 IPs, each generating hundreds or thousands of requests in a single session.

Live agents, on the other hand, exhibit extremely high IP dispersion with almost zero session breadth. A tool like ChatGPT-User will spin up from hundreds of different IPs (356 in our dataset) but only fetch 1.15 paths per IP on average. This means they fetch exactly what the user asks for and nothing else.

- **GPTBot Session Breadth:** Averaged 279.67 unique paths per IP, peaking at a massive 2,298 paths from a single IP address, acting as a classic bulk crawler.
- **ClaudeBot Sweep Pattern:** Showed similar bulk behavior with 144.47 unique paths per IP, maxing out at 1,268 paths in a single coordinated crawl burst.
- **ChatGPT-User Precision:** Distributed across 356 IPs fetching a median of exactly 1.0 path per IP, proving these are purely user-driven real-time fetches.

FIGURE 1: ORIGIN SERVER FOOTPRINT — BULK INDEXERS VS. LIVE CITATION FETCHERS



Why Session Breadth Matters for Crawl Budget

Understanding session breadth is critical for setting WAF rules and rate limits. A rate limit of 50 requests per minute per IP will completely strangle GPTBot's attempt to index your site, while leaving ChatGPT-User and Claude-User entirely unaffected. Conversely, attempting to block live agents by IP reputation is nearly impossible due to their high IP dispersion across datacenter ranges.

3. Empirical Footprint & Status Matrix

Verified Family	Mean Paths/IP	Median Paths/IP	200s	301s	304s	404s
GPTBot	279.67	1.0	83.55%	15.46%	0.10%	0.84%
ClaudeBot	144.47	16.0	82.93%	14.21%	1.71%	0.83%
Meta-ExternalAgent	8.92	9.0	82.50%	11.89%	3.47%	2.14%
Bytespider	6.73	2.0	73.60%	3.68%	15.77%	0.96%
ChatGPT-User	1.15	1.0	98.55%	1.21%	0.24%	0.00%
Claude-User	3.82	2.0	59.90%	37.56%	1.52%	0.00%
Claude-Code-CLI	1.41	1.0	80.70%	19.30%	0.00%	0.00%
Google-NotebookLM	6.29	7.0	93.59%	1.92%	4.49%	0.00%
OAI-SearchBot	6.90	1.0	78.24%	3.77%	14.29%	1.73%
PerplexityBot	72.88	5.0	94.72%	3.04%	1.56%	0.68%
Verified Search Crawlers (Class 1 Ref)	4.28	1.0	85.75%	4.17%	8.01%	1.53%

4. The Real-Time 301 Redirect Latency Tax vs. robots.txt Governance

Status code hygiene is the hidden performance killer for AI answer engines. When a bulk crawler hits a 301 redirect, it logs the new canonical URL and schedules a fetch for its next pass. But when a live prompt-time agent hits a 301 redirect, the end user waiting for a chat response absorbs the full latency penalty of the extra HTTP hop, TLS handshake, and round-trip time.

This latency tax is severe. Because live agents cannot rely on caching (evidenced by their near-zero 304 rates), every redirect forces a cold retrieval. Our data shows that Anthropic's tools are paying a massive performance penalty on modern web infrastructure due to trailing-slash and www-canonicalization redirects.

- **Claude-User (37.56% 301s):** More than a third of all live fetches by Claude-User hit a redirect before reaching the payload, severely degrading answer speed.
- **Claude-Code-CLI (19.30% 301s):** Developers using the CLI tool face similar latency, while ChatGPT-User avoids this entirely (98.55% 200s, 0.00% 404s).

5. Selective Routing & Canonical Hygiene Checklist

1. **Differentiated robots.txt/rate-limiting tiers:** Use dedicated User-Agent blocks to aggressively rate-limit GPTBot and ClaudeBot to protect your crawl budget, while explicitly allowing high-speed access for live citation fetchers like ChatGPT-User and Claude-User.
2. **Eliminating 301 hops on citation URLs:** Ensure any URL you feed into an AI prompt, knowledge base, or link schema resolves directly to a 200 OK without trailing-slash or www redirects to prevent answer-time latency taxes.
3. **Serving clean 200 HTML payloads for answer-time agents:** Deliver fast, statically available or server-rendered HTML content, because live agents operate on strict 1-3 second latency budgets and cannot execute client-side JavaScript.
4. **Providing distinct endpoints:** Serve structured Markdown twins or dedicated clean endpoints alongside your HTML for agents that struggle with complex DOM structures, ensuring they retrieve clean facts instantly.
5. **Monitoring crawl footprint:** Audit your access logs weekly to verify that live agents aren't getting caught in indiscriminate WAF rules, IP bans, or rate limits that were originally designed for aggressive bulk indexers.

Primary Verification Sources

- **RFC 9309 (robots.txt):** https://www.rfc-editor.org/rfc/rfc9309.html — Standardizes how user agents process robots.txt files.
- **RFC 9110 (HTTP Semantics):** https://www.rfc-editor.org/rfc/rfc9110.html — Explains the latency and hop impacts of status 301 and 304.
- **OpenAI GPTBot Feed:** https://openai.com/gptbot.json — Official CIDR ranges for the GPTBot bulk indexer.
- **OpenAI ChatGPT-User Feed:** https://openai.com/chatgpt-user.json — Official CIDR ranges for live chat fetches.
- **Anthropic Crawler Feed:** https://claude.com/crawling/bots.json — Specifies IP ranges for ClaudeBot and Claude-User.
- **Google User-Triggered Fetchers:** https://developers.google.com/crawling/docs/crawlers-fetchers/google-user-triggered-fetchers — Describes how user-triggered tools fetch URLs.
- **Tygart Media JS Rendering Study:** https://tygartmedia.com/do-ai-crawlers-render-javascript-server-log-white-paper/ — Preceding study on AI rendering footprints.

ABOUT THIS RESEARCH

Published by Tygart Media (tygartmedia.com) on November 23, 2026. All traffic counts reflect anonymized server access logs across eight independent WordPress properties audited with strict vendor IP/ASN verification. Internal hostnames and full IPs have been omitted to maintain operational security.