

AI Crawlers Are Reportedly Draining Site Resources & Skewing Analytics via @sejournal, @MattGSouthern

Website operators across the web are reporting increased activity from AI web crawlers. This surge raises concerns about site performance, analytics, and server resources.

These bots consume significant bandwidth to collect data for large language models, which could impact performance metrics relevant to search rankings.

Here's what you need to know.

How AI Crawlers May Affect Site Performance

SEO professionals regularly optimize for traditional search engine crawlers, but the growing presence of AI crawlers from companies like OpenAI, Anthropic, and Amazon presents new technical considerations.

Several site operators have reported performance issues and increased server loads directly attributable to AI crawler activity.

"SourceHut continues to face disruptions due to aggressive LLM crawlers," reported the git-hosting service on its status page.

In response, SourceHut has "unilaterally blocked several cloud

providers, including GCP [Google Cloud] and [Microsoft] Azure, for the high volumes of bot traffic originating from their networks.”

Data from cloud hosting service Vercel shows the scale of this traffic: OpenAI’s GPTBot generated 569 million requests in a single month, while Anthropic’s Claude accounted for 370 million.

These AI crawlers represented about 20 percent of Google’s search crawler volume during the same period.

The Potential Impact On Analytics Data

Significant bot traffic can affect analytics data.

According to DoubleVerify, an ad metrics firm, “general invalid traffic – aka GIVT, bots that should not be counted as ad views – rose by 86 percent in the second half of 2024 due to AI crawlers.”

The firm noted that “a record 16 percent of GIVT from known-bot impressions in 2024 were generated by those that are associated with AI scrapers, such as GPTBot, ClaudeBot and AppleBot.”

The Read the Docs project found that blocking AI crawlers decreased their traffic by 75 percent, from 800GB to 200GB daily, saving approximately \$1,500 per month in bandwidth costs.

Identifying AI Crawler Patterns

Understanding AI crawler behavior can help with traffic analysis.

What makes AI crawlers different from traditional bots is

their frequency and depth of access. While search engine crawlers typically follow predictable patterns, AI crawlers exhibit more aggressive behaviors.

Dennis Schubert, who maintains infrastructure for the Diaspora social network, observed that AI crawlers “don’t just crawl a page once and then move on. Oh, no, they come back every 6 hours because lol why not.”

This repeated crawling multiplies the resource consumption, as the same pages are accessed repeatedly without a clear rationale.

Beyond frequency, AI crawlers are more thorough, exploring more content than typical visitors.

Drew DeVault, founder of SourceHut, noted that crawlers access “every page of every git log, and every commit in your repository,” which can be particularly resource-intensive for content-heavy sites.

While the high traffic volume is concerning, identifying and managing these crawlers presents additional challenges.

As crawler technology evolves, traditional blocking methods prove increasingly ineffective.

Software developer Xe Iaso noted, “It’s futile to block AI crawler bots because they lie, change their user agent, use residential IP addresses as proxies, and more.”

Balancing Visibility With Resource Management

Website owners and SEO professionals face a practical consideration: managing resource-intensive crawlers while maintaining visibility for legitimate search engines.

To determine if AI crawlers are significantly impacting your

site:

- Review server logs for unusual traffic patterns, especially from cloud provider IP ranges
- Look for spikes in bandwidth usage that don't correspond with user activity
- Check for high traffic to resource-intensive pages like archives or API endpoints
- Monitor for unusual patterns in your Core Web Vitals metrics

Several options are available for those impacted by excessive AI crawler traffic.

Google introduced a solution called Google-Extended in the robots.txt file. This allows websites to stop having their content used to train Google's Gemini and Vertex AI services while still allowing those sites to show up in search results.

Cloudflare recently announced "AI Labyrinth," explaining, "When we detect unauthorized crawling, rather than blocking the request, we will link to a series of AI-generated pages that are convincing enough to entice a crawler to traverse them."

Looking Ahead

As AI integrates into search and discovery, SEO professionals should manage crawlers carefully.

Here are some practical next steps:

1. Audit server logs to assess AI crawler impact on your specific sites
2. Consider implementing Google-Extended in robots.txt to maintain search visibility while limiting AI training access
3. Adjust analytics filters to separate bot traffic for more accurate reporting

4. For severely affected sites, investigate more advanced mitigation options

Most websites will do fine with standard robots.txt files and monitoring. However, high-traffic sites may benefit from more advanced solutions.

Featured Image: Lightspring/Shutterstock

<https://www.searchenginejournal.com/ai-crawlers-draining-site-resources/543011/>