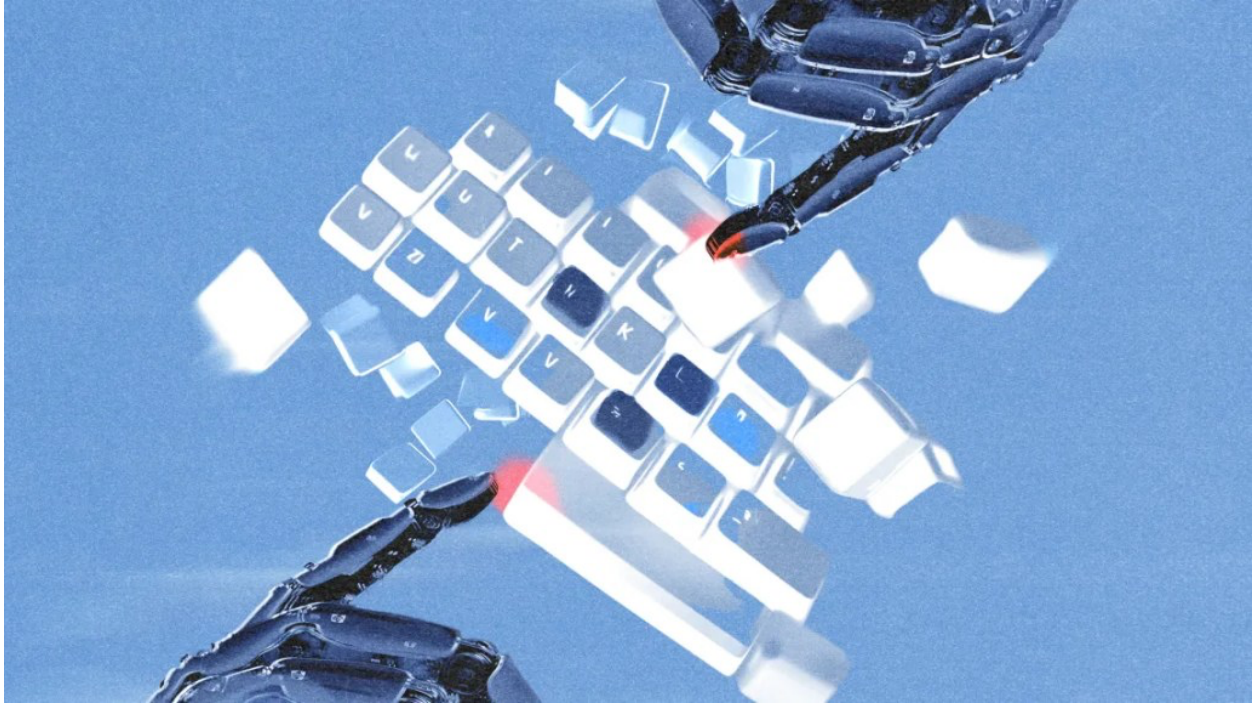


## In Graphic Detail: New data shows publishers face growing AI bot, third-party scraper activity

By [Sara Guaglione](#) • April 13, 2026



A third-party scraper economy is emerging beneath the big AI companies, making it harder than ever for publishers to know who is taking their content, let alone stop them.

“What’s changed is the separation of roles. The entity extracting the data is often not the one using or monetizing it. That abstraction layer is what’s enabling third-party resale markets to scale,” Brent Maynard, senior director of security technology and strategy at content delivery network company Akamai, said. “We’re hearing this directly from publishers. One large publisher described it as: We’re not being scraped by one company anymore, we’re being harvested by an ecosystem.”

Digiday has compiled four graphs revealing how AI web scraping is evolving, and what it means for publishers:

## The rise of third-party scrapers selling publisher content

### QUICK REFERENCE — ALL 21 VENDORS

| Vendor        | CEO & HQ                                       | Funding                 | Top Named Customers                                     | Primary Product                   | Entry Price             | Legal                    |
|---------------|------------------------------------------------|-------------------------|---------------------------------------------------------|-----------------------------------|-------------------------|--------------------------|
| Firecrawl     | Caleb Peffer · San Francisco                   | \$16.2M                 | OpenAI, Apple, Shopify, PwC, Alibaba, Carva             | Website → LLM markdown            | Free · \$16–\$333/mo    | —                        |
| Exa           | Will Bryk · San Francisco                      | \$111M                  | Notion, Cursor, AWS, Databricks, Groq, HubSpot          | AI-native search API              | \$5/1K searches         | —                        |
| Tavily        | Rotem Weiss · New York                         | \$25M                   | Cohere, IBM, BCG, MongoDB, Groq, JetBrains, AWS         | Real-time agent search            | Free · \$30–\$500/mo    | Acquired Nebius          |
| Parallel AI   | Parag Agrawal (ex-Twitter CEO) · San Francisco | \$130M                  | Harvey, Modal, Opendoor, Gumloop, Lindy                 | Full pipeline → 1 API call        | Per-query · custom      | —                        |
| You.com       | Richard Socher (ex-Salesforce CSO) · Palo Alto | ~\$200M · \$1.5B val.   | OpenAI, Amazon, Salesforce, Databricks, The Telegraph   | Web search API suite              | Tiered · enterprise     | —                        |
| Brave Search  | Brendan Eich (JS inventor) · San Francisco     | —                       | No named customers disclosed                            | Independent search index API      | ~\$3/1K queries         | —                        |
| Bright Data   | Or Lenchner · Netanya, Israel                  | \$300M ARR              | 14/20 top LLM labs · Fortune 500                        | Proxy + scraper + datasets        | \$499+/mo · enterprise  | Meta & X suits dismissed |
| Jina AI       | Han Xiao (now at Elastic) · Berlin             | Acquired by Elastic     | Major LLM companies, banks, AI teams                    | URL → LLM text + multimodal       | Free tier · PAYG        | Elastic acq. Oct 2025    |
| SerpApi       | Julien Khaleghy · Austin, TX                   | —                       | Perplexity (court filing)                               | 40+ engine SERP API               | \$50–\$250/mo+          | Reddit & Google lawsuits |
| Diffbot       | Mike Tung · Menlo Park, CA                     | —                       | Dow Jones, Andreessen Horowitz, FINRA, Factset, Cisco   | Knowledge graph + GraphRAG        | \$299–\$799/mo+         | —                        |
| Serper.dev    | Not disclosed · Not disclosed                  | —                       | No named customers                                      | Fast cheap Google SERP API        | \$0.30/1K queries       | —                        |
| Apify         | Jan Čurn (PhD AI) · Prague                     | —                       | Enterprise, universities, LLM developers                | 22,764+ Actor marketplace         | \$49–\$499/mo           | —                        |
| Oxylabs       | Julius Černiauskas · Vilnius                   | —                       | Stanford, Bellingcat, ICIJ, European Commission, Forbes | Enterprise proxy infra            | \$8–15/GB               | Reddit suit Oct 2025     |
| Zyte          | Shane Evans · Cork, Ireland                    | —                       | Sayari, Muckrack, MikMak, InMoment                      | Managed news + AI data feeds      | Managed custom          | —                        |
| Decodo        | Tesonet group · Vilnius (f. Smartproxy)        | —                       | No named customers disclosed                            | 115M+ IP proxy network            | \$8.50/GB · \$75+ min   | —                        |
| ScrapeGraphAI | Marco Perini · Italy                           | —                       | LangChain, CrewAI, Llamaindex ecosystems                | AI-native agent scraper           | Free · credits · custom | —                        |
| Scrapingdog   | Not disclosed · India                          | —                       | No named customers disclosed                            | ChatGPT + Google AI scraper APIs  | \$30–\$200/mo           | —                        |
| Perplexity    | Aravind Srinivas · San Francisco               | \$21B val. · \$200M ARR | Jeff Bezos, NVIDIA, Microsoft Azure, Amazon, Snap       | Real-time AI search API           | Dow Jones, NYT, Reddit  | —                        |
| ScraperAPI    | Daniel Ni · Las Vegas, NV · 2018               | ~\$5M ARR               | Volvo, Bose, HelloFresh, Deloitte, Sony, Nielsen        | Proxy + scraper + structured data | \$49–\$249/mo           | —                        |
| Scrape.do     | Not disclosed · Ankara, Turkey · 2020          | \$1.3M ARR              | No named customers                                      | Proxy API + scraping toolkit      | \$19–\$199/mo           | —                        |
| Grepser       | Amit Chaudhary · Kathmandu / NYC · 2012        | \$13.9M ARR             | Fortune 500 (unnamed)                                   | Managed data-as-a-service         | Enterprise custom       | — Page 6                 |

Source: MSG

Publishers have spent years fighting the big AI companies over scraping. The smaller operators — dozens of third-party vendors crawling the web and selling content to enterprise clients — have largely escaped scrutiny. Media analyst Matthew Scott Goldstein's report on the "scraper economy", presented at a recent meeting he regularly hosts between publishers and tech companies and media analysts, aims to change that.

Publishers make no money from these relationships, according to Goldstein. He said this is a \$1 billion industry, citing Mordor Intelligence data.

"Organized publisher supply into a licensed content marketplace doesn't need to create new demand. Every enterprise in this report has demonstrated willingness to pay. The infrastructure exists. The buyers are named, funded, and growing. The only missing piece is organized publisher supply at scale — and the urgency to move before the compliance window closes," Goldstein wrote in his report, which he shared with Digiday.

The report identified 21 vendors doing this, including Firecrawl, Exa, Tavily, Brave, You.com, Perplexity Sonar and Bright Data. (TollBit also has a running index on third-party scrapers, identifying nearly 40 vendors.)

Over 70 companies were found to be paying for publisher content from these vendors, including BCG, IBM, Cohere, AWS, Salesforce, Apple, Latham & Watkins, Zoom, PwC, Shopify and Alibaba.

## **Publishers are getting hit hard by AI bot activity surge**

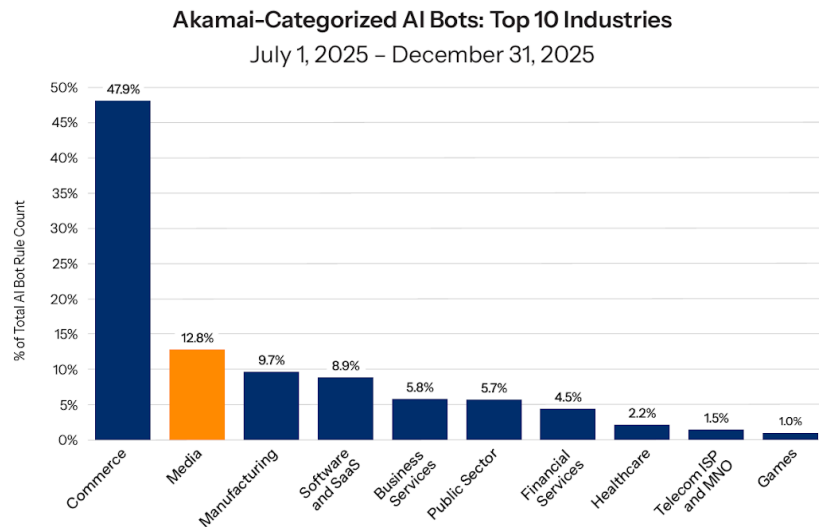


Fig. 1: The media industry (which includes publishing) recorded the second-highest AI bot activity globally in 2H 2025  
**Source: Akamai**

What's exacerbating the problem is the rise in AI bot activity.

Akamai, which handles more than one-third of all global internet traffic, saw a 300 percent surge in AI bot activity in 2025, according to its recent “Protecting Publishing: Navigating the AI Bot Era” report.

While most of this targeted the commerce industry (48 percent), media was in second place (13 percent), the report said. Specifically, publishers represented 40 percent of all media-related AI bot activity.

OpenAI, Meta, and ByteDance were the top three culprits between July and December last year. Of those, OpenAI generated the highest volume of AI bot traffic targeting media companies. Publishers accounted for 40 percent of that traffic, per the report.

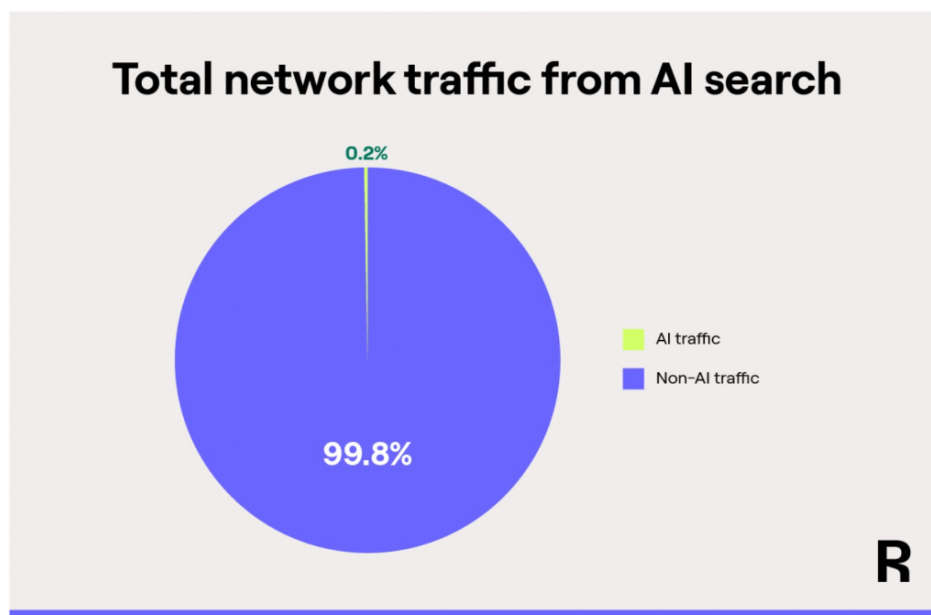
Worse still, the methods are getting more sophisticated. Akamai found that “AI fetchers” — or bots that grab specific web pages in real time to answer specific user requests on AI assistants like chatbots — accounted for 24 percent of AI bot types in the media industry in the second half of last year. Publishers represented 43 percent of that. Meanwhile, AI training crawlers, which scan and collect large

amounts of data from websites to train LLMs, made up 63 percent of the AI bot activity Akamai tracked.

“We’re seeing growth in real-time fetchers that pull content dynamically to answer queries,” Maynard said. “The risk isn’t just that content is taken, it’s that the visit never happens. We’re already seeing publishers connect this to declining referral traffic and changing user behavior... AI bot traffic is persistent and growing, but it doesn’t create value for publishers. It consumes infrastructure and content without contributing to revenue.”

Cybersecurity company Human Security found automated traffic is now growing eight times faster than human traffic, according to its recently published [“2026 State of AI Traffic & Cyberthreat Benchmarks” report](#). AI scraper traffic grew 597 percent from January to December 2025, and AI-driven traffic overall grew 187% in 2025, nearly tripling year over year, per the report.

## Plenty of bot activity, but little to show for it

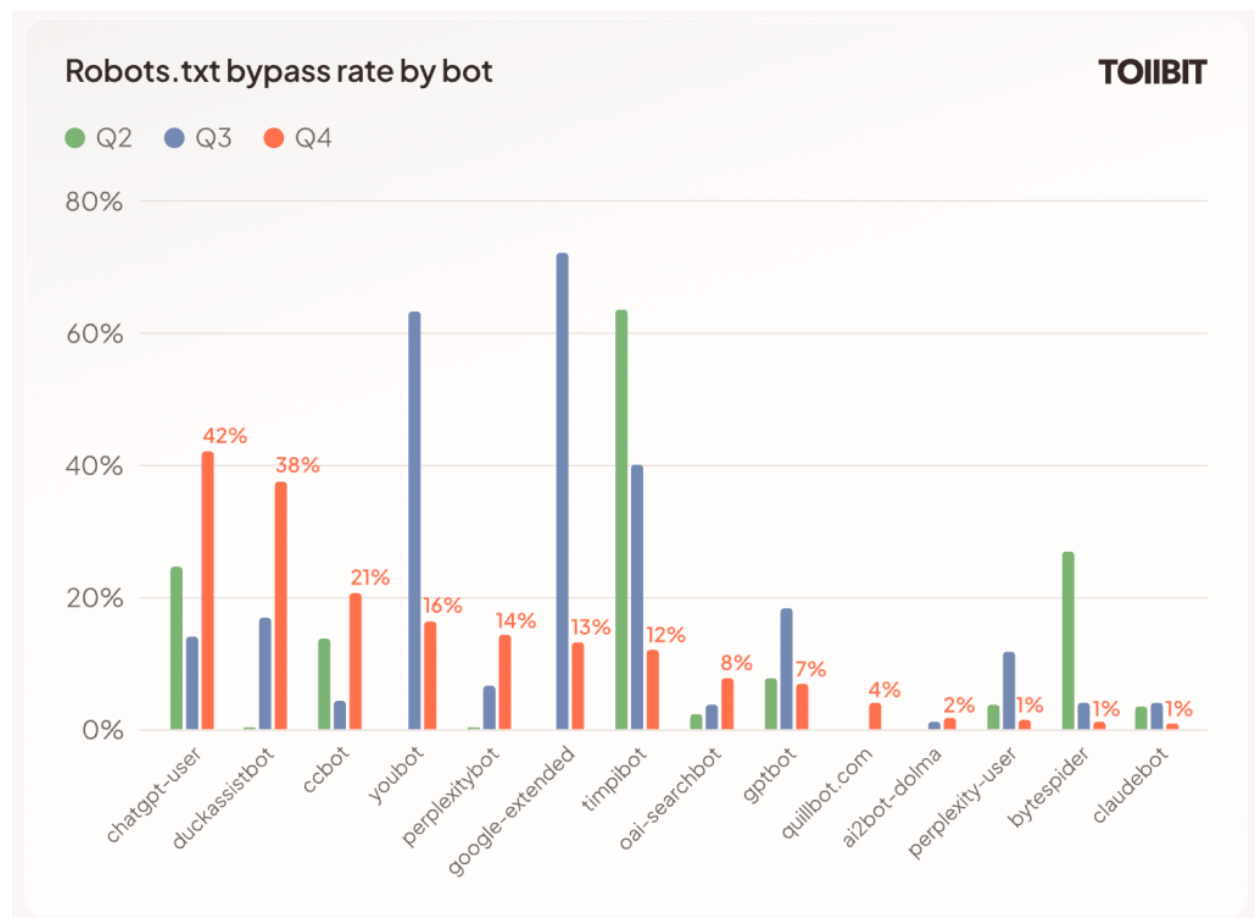


Source: Raptive

Despite the volume of bot activity, publishers are seeing little in return. ChatGPT is driving less than 0.2% of traffic to Raptive’s network of 6,000 independent publishers, according to [a recent Raptive guide](#) on AI bot blocking – an issue [Digiday has previously reported](#) on. In short, the more AI companies take, the less they send back.

“As a publisher, or content creator, you don’t get anything from this,” Paul Bannister, chief strategy officer at Raptive, said. “It’s minuscule. It grew fast last year and then it’s flat lined for a while. They’re not sending more traffic out. And until there’s a better quid pro quo here, whether it’s traffic or money or something else, what’s in it for you? What’s in it for any of us?”.

## Bot blocking still a challenge



The solution to all of this, according to Goldstein, is easy: “Block the shit out of every single one of these companies. Block, block, block, and then block some more.”

The problem is that blocking is harder than it sounds. Most publishers rely on a single cybersecurity tool or bot detection system. AI companies have dozens of scraping tools to choose from, many designed specifically to circumvent bot blocking mechanisms.

Tollbit’s latest “State of the Bots” report underscored the point. It found that about 30 percent of AI bot scrapes violate explicit instructions in robots.txt to “disallow” web crawlers to access their content.

“Robots.txt... is as useful as a chocolate teapot right now,” one publishing exec told Digiday, under the condition of anonymity.

There is some hope. Content delivery networks like Akamai, Fastly and Cloudflare have more sophisticated ways to identify AI bots and block them. That’s helped publishers like People Inc. recently take the route of blocking all bots and only allowing the ones they’ve deemed “permissible” to access their content, Jon Roberts, chief innovation officer at People Inc, said onstage at the Digiday Publishing Summit in Vail, Colorado last month. Mark Howard, Time’s chief operating officer, also said at the same event that his company was planning on implementing a more aggressive bot-blocking approach.

“The organizations that get ahead of this will treat AI access as both a security problem and a business model decision. If your content is fueling AI, you should have a say in how it’s accessed, used and monetized,” Maynard said.

The Atlantic has done similar, allowing only bots from companies that it has a commercial or strategic deal with to access its content.