Report

Q2 2025 Fastly Threat Insights Report

Crawled, Scraped, Strained: Insights on AI Bot Behavior

fastly

Table of Contents

The Fastly Threat Insights Report highlights security trends, attack vectors, and threat activity across the application security landscape. Drawing from trillions of requests across our global customer base, this report offers a real-time view into what's materially impacting security teams in the context of larger trends.

In this edition, we exclusively focus on AI bots, a rapidly growing category of automated traffic that is reshaping how content is accessed, scraped, and potentially monetized across the web. As organizations navigate the growing impact of AI-driven automation, understanding the behavior, scale and risks of these bots is essential to maintaining visibility, controlling content access, and preserving competitive advantage.

This quarter's insights are derived from traffic analyzed across Fastly's Next-Gen WAF (NGWAF) and Bot Management products between April 16 and July 15, 2025. These solutions collectively protect over 130,000 applications and APIs* and inspect more than 6.5 trillion requests per month**. Fastly's broad visibility spanning edge and cloud-native architectures, combined with our presence across a wide range of industries, including leading e-commerce, streaming, media and entertainment, financial services, and technology organizations, provides us with a unique and comprehensive view of the global web application threat landscape.

As of April 2025
***Trailing 6-month average as of April 2025*

Key Takeaways

1. AI bots can place significant strain on unprotected web infrastructure, with peak traffic reaching up to 39,000 requests per minute.
2. Commerce, media & entertainment, and high-tech sectors face the highest levels of scraping for training AI models.
3. AI crawlers account for 80% of AI bot traffic, with roughly half of that attributed to Meta alone. In contrast, AI fetchers make up the remaining 20% of the traffic.
4. ChatGPT generates the most real-time traffic to websites, with 98% of fetcher bot requests attributable to OpenAI's bots.
5. Most AI models are trained predominantly on content originating from North America, shaping their alignment according to the region's cultural and geopolitical perspectives.

Definitions

Name	Definition
Bot Traffic	Any non-human internet traffic that can be beneficial (e.g., search engine crawlers) or malicious (e.g., credential stuffing).
AI Crawler Bots	Crawlers used for training AIs and LLMs. These bots are generally used for building AI models or indexes.
AI Fetcher Bots	Fetcher used for AIs and LLMs for enriching results in response to a user query.
Autonomous Systems (AS)	A collection of one or more IP prefixes (networks) managed by a single organization or entity.
Large Language Models (LLMs)	LLMs are advanced AI systems trained on massive text datasets to understand and generate human-like language, enabling tasks such as answering questions, summarizing content and writing text.
AI Training	Training is when an LLM learns language patterns and relationships by processing large amounts of primarily textual data.
AI Inference	Inference is when a trained LLM generates responses or predictions based on new input, using what it learned during training.
Prompting	Prompting is the act of providing inputs such as a question, instructions, or other data to a language model to elicit its response.

Methodology and Limitations

Some AI bots provide clear verification mechanisms, such as publishing IP ranges or supporting reverse DNS checks, that make it straightforward to confirm the authenticity of their traffic. In these cases, impostor traffic is easily distinguished and has been excluded from the dataset used in our analysis.

However, not all AI bots offer this level of transparency. For bots lacking public verification data, we rely on alternative heuristics such as the originating network, behavioral patterns, and other identifiable traffic signatures to assess authenticity. While these methods are generally effective, they do not offer absolute certainty. We've observed a significant amount of traffic identifying as unverifiable AI bots originating from numerous ASs. This pattern suggests that threat actors may be exploiting the unverifiable nature of these bots to impersonate them and evade scrutiny.

As a result, this report may include traffic from bots that could not be fully verified. We have worked to ensure the accuracy of the classification and the integrity of the data presented. We encourage AI bot operators to adopt transparent verification practices to improve attribution.

We have also chosen not to reclassify general-purpose search engine crawlers as AI bots. This is a deliberate decision to avoid misleading conclusions, particularly since blocking such crawlers could unintentionally impact a website's visibility in the associated search engines. Our analysis includes only bots that are clearly and exclusively intended for AI-related use cases, such as model training or inference-time content retrieval. This means some search engine crawlers using their data for LLM training will not factor into our analysis.

It is important to note that some AI models are trained entirely on publicly available datasets such as Common Crawl and Wikipedia. These models may not perform any active web crawling themselves, or may only engage in limited, targeted crawling aligned with a specific purpose or domain. Since this report focuses exclusively on observable AI bot activity across websites, models built without significant crawling behavior are naturally not represented in our analysis.

To protect the privacy of our customers and avoid unintentionally singling out individual websites, certain details have been intentionally excluded or aggregated in the visualizations and analysis presented in this report. Fastly serves a significant portion of the web, giving it a global perspective on AI bot activity. This report focuses on overarching patterns in bot behavior, rather than on any individual site or service.

Findings and Insights

Automation tools commonly known as bots drive a large share of internet traffic. Fastly uses network signals, behavioral analysis, and advanced challenges to distinguish human users from bots.

About 87% of bot traffic is malicious, enabling account takeovers, ad fraud, carding, and more. But there are also legitimate uses of bots, such as search engine crawlers or uptime monitoring tools. The challenge for website owners often is in letting the good bots in while keeping the bad ones out.

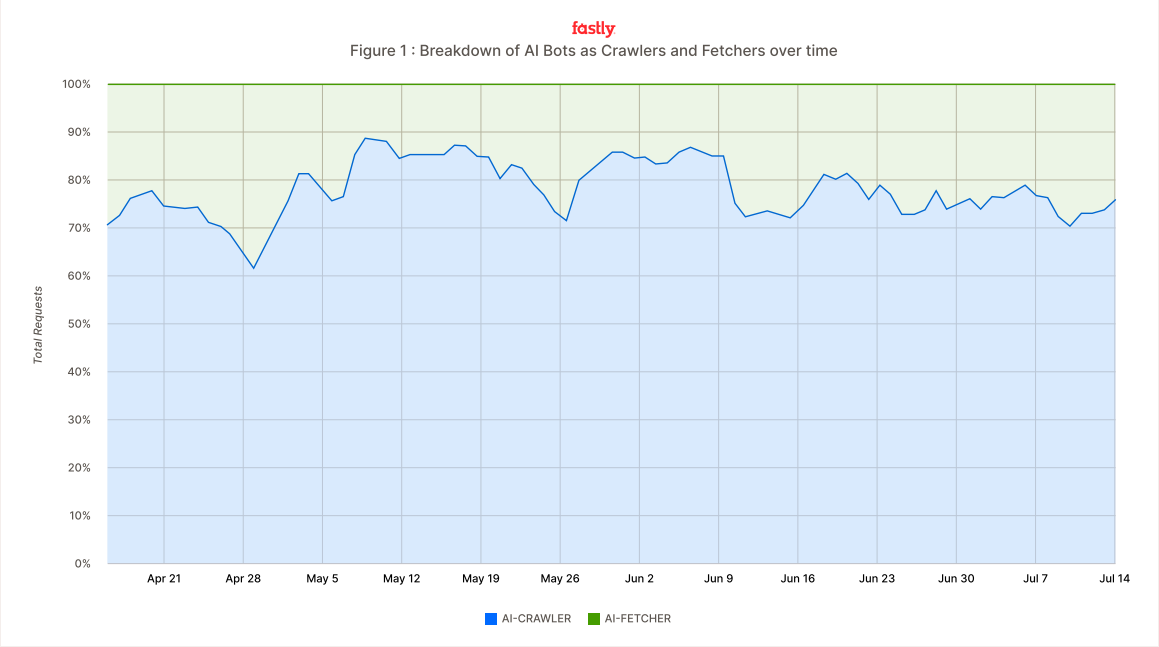
Adding to this complexity, a new class of bots has recently emerged in the form of AI bots, which crawl websites either to train large language models (LLMs) or fetch content to enrich model responses with grounding at inference time. Whether these bots are seen as a benefit or a risk depends on the site owner's priorities.

Understanding AI Bot Behavior

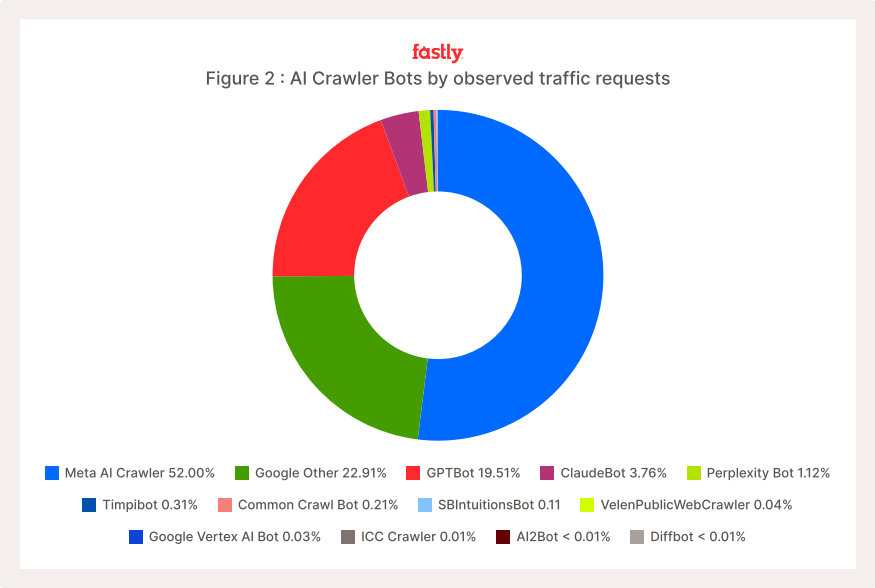
In this report, we categorize AI bots into two types based on their behavior and intended use case: Crawlers and Fetchers.

Crawler bots operate similarly to search engine crawlers - they systematically scan websites to collect content for building searchable indexes or training language models. This process is a precondition to the model's "training" phase.

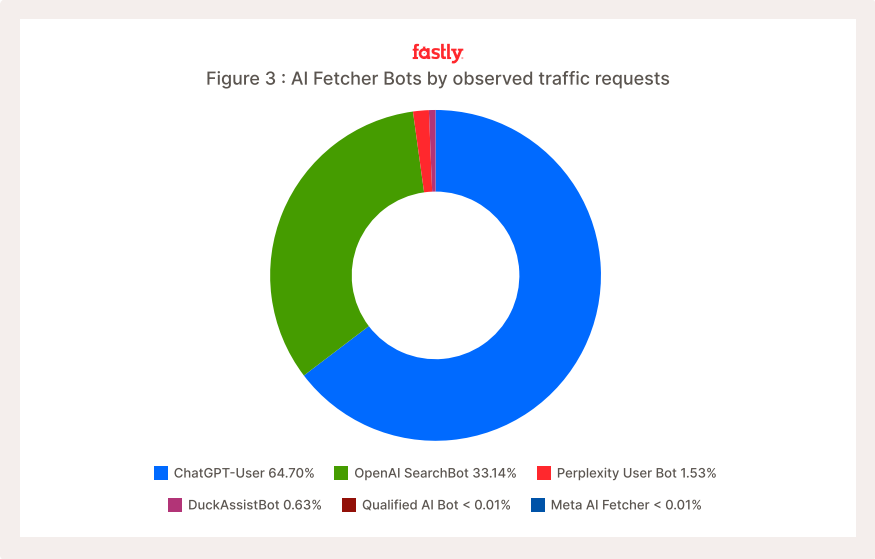
Fetcher bots, on the other hand, access website content in response to user actions. For example, when a user requests up-to-date information on a specific topic, a fetcher bot retrieves the relevant page in real time. They are also used to help surface website links that match a user's search query, directing them to the most relevant content. Crawler bots contribute nearly 80% of the total AI bot request volume, with fetcher bots making up the remaining 20% (Figure 1).



Crawlers typically scrape publicly accessible, authoritative content such as news sites, educational resources, government pages, technical documentation, and open datasets. Meta, Google, and OpenAI account for a combined 95% of all AI crawler request volume, with Meta leading at 52%, followed by Google at 23%, and OpenAI at 20% (Figure 2).



Fetcher bots play a key role in grounding LLM-generated responses by retrieving reference links on the spot, allowing the model to verify or support its outputs with authoritative, real-time sources. This overall process where the model generates a response to a user query, often using real-time external data retrieved by fetcher bots, is commonly referred to as the model's "inference" phase.



With *ChatGPT-User* and *OAI-SearchBot* accounting for nearly 98% of all fetcher requests, it's clear that OpenAI, driven largely by its ChatGPT product, has the most significant impact on websites in terms of fetcher traffic (Figure 3). While the volume of fetcher requests from Perplexity is much lower at 1.53%, we observe an upward trend of its impact over time (Figure 4).

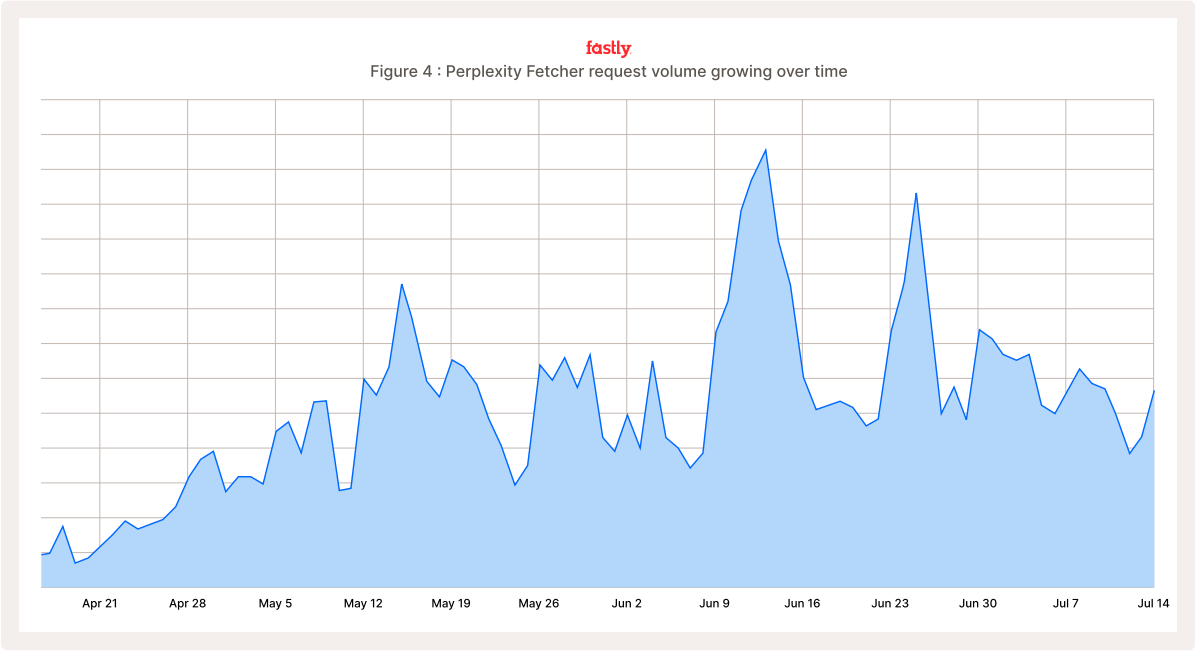


Figure 4 : Perplexity Fetcher request volume growing over time

Crawler Trends

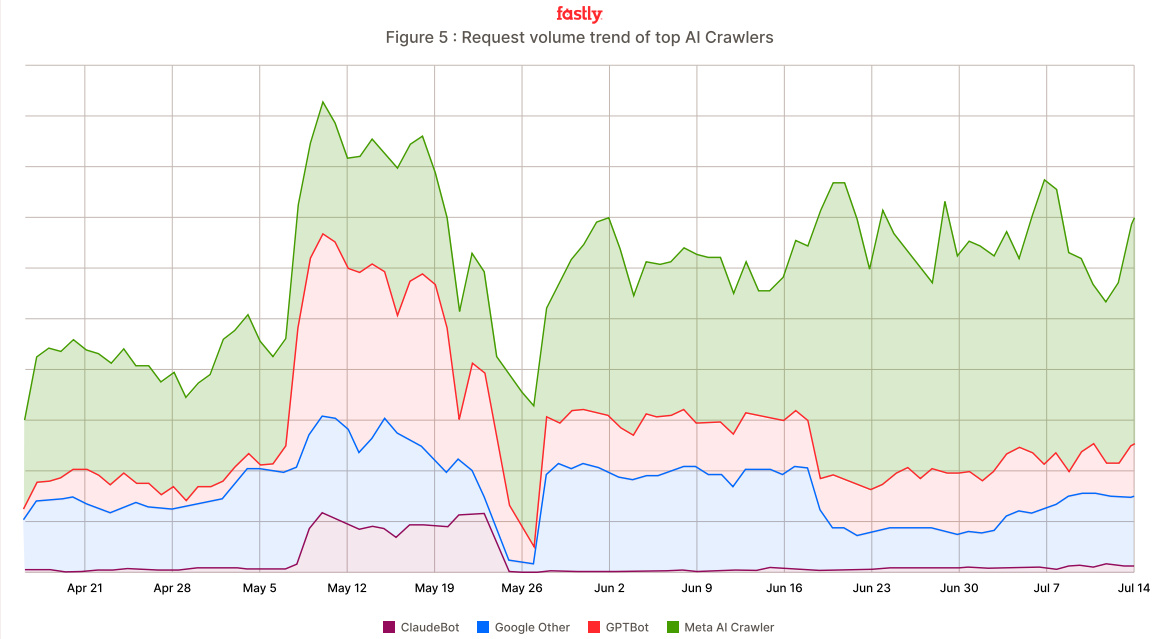


Figure 5 : Request volume trend of top AI Crawlers

Examining the request volume trend of the top AI Crawlers, we observe Meta's crawler having an upward trend (Figure 5). Most crawlers exhibit irregular patterns of activity, with extended periods of moderate or low request volumes interspersed with sustained increases lasting days or even weeks, often around 2–3 times the baseline rate. The only consistent exception is Common Crawl's *CCBot*, which follows a more regular and predictable crawling schedule.

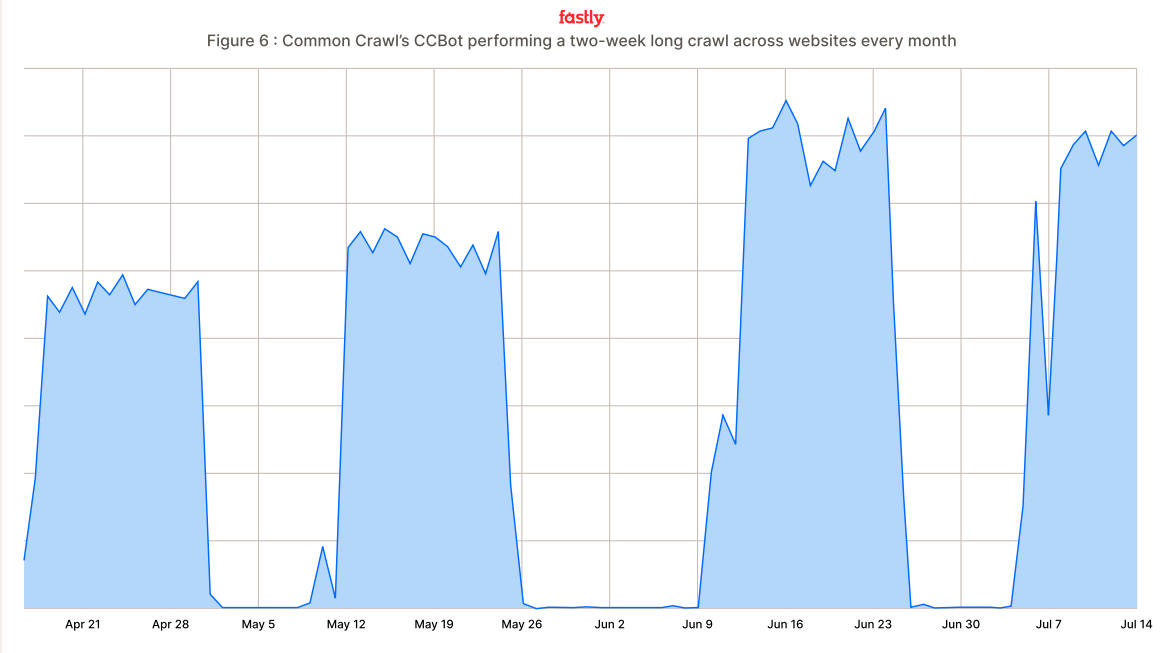


Figure 6 : Common Crawl's CCBot performing a two-week long crawl across websites every month

Common Crawl is a non-profit organization that regularly crawls the web using its crawler, *CCBot*, and provides free, open-access datasets of raw web content. Many AI models are trained on this data or its derivatives such as the C4 dataset, often in combination with their own crawled sources, making *CCBot* one of the most influential crawlers in the AI ecosystem. We observed *CCBot* conducting broad, two-week long crawls across a wide range of websites each month, with the volume of crawl requests steadily increasing over time (Figure 6).

Impact on Web Infrastructure

Some AI bots, if not carefully engineered, can inadvertently impose an unsustainable load on webservers, leading to performance degradation, service disruption, and increased operational costs.

1 K
RPM
Peak crawler rate on website

39 K
RPM
Peak fetcher rate on website

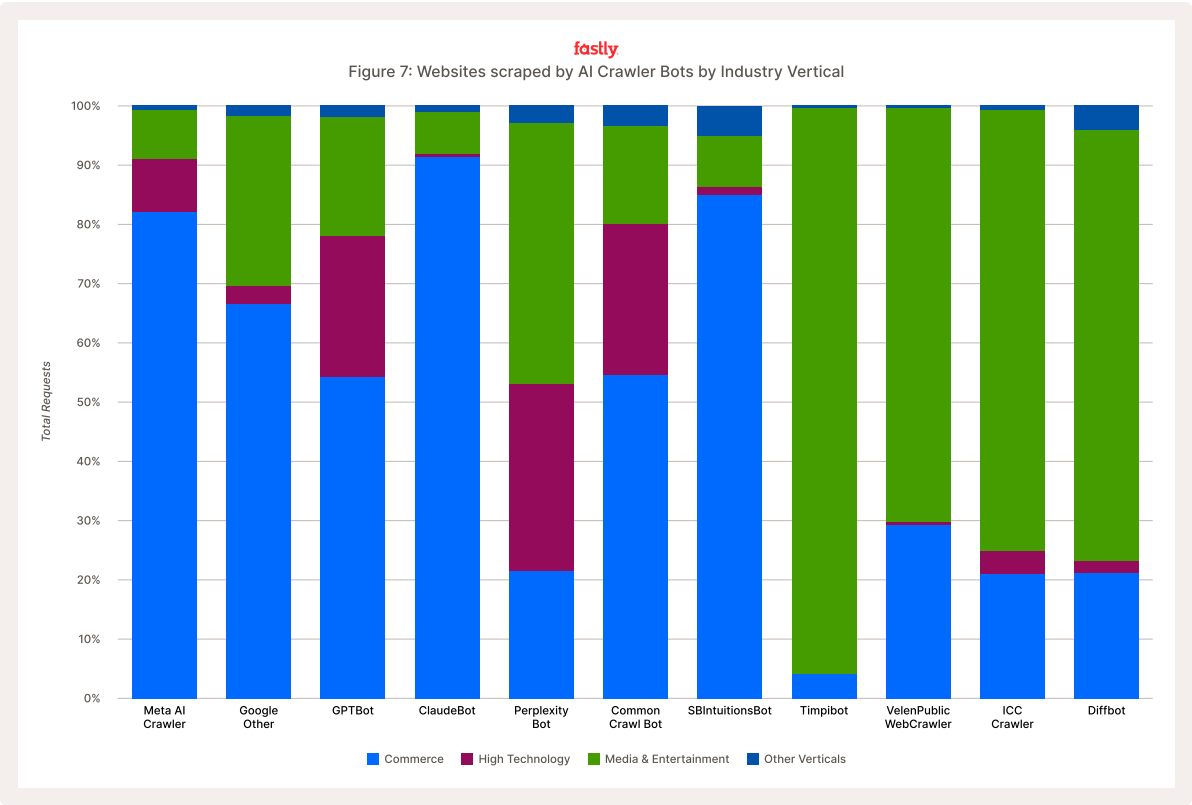
A concerning trend is the spike in traffic from large-scale AI bots. In one case, a single crawler reached a peak of around 1,000 requests per minute. While this isn't an exceptionally high volume, it can still place a significant burden on websites that rely on database queries or provide web interfaces for browsing Git repositories, such as Gitea. For such systems, even short bursts of activity can cause slowdowns, timeouts, or disruptions without effective bot controls or scaling measures.

Real-time fetching by AI bots presents a greater challenge. In one instance, a fetcher bot made 39,000 requests per minute to a single website at peak load. This traffic volume can overwhelm origin servers, strain server resources, consume bandwidth, and cause expensive DDoS-like effects, even without malicious intent.

Excessive bot traffic can degrade user experience, drive up infrastructure costs, and skew website analytics. These challenges highlight the need for advanced bot management solutions that accurately detect AI bots, distinguish between helpful and harmful traffic, and adapt in real time to prevent overload.

Training Content Insights

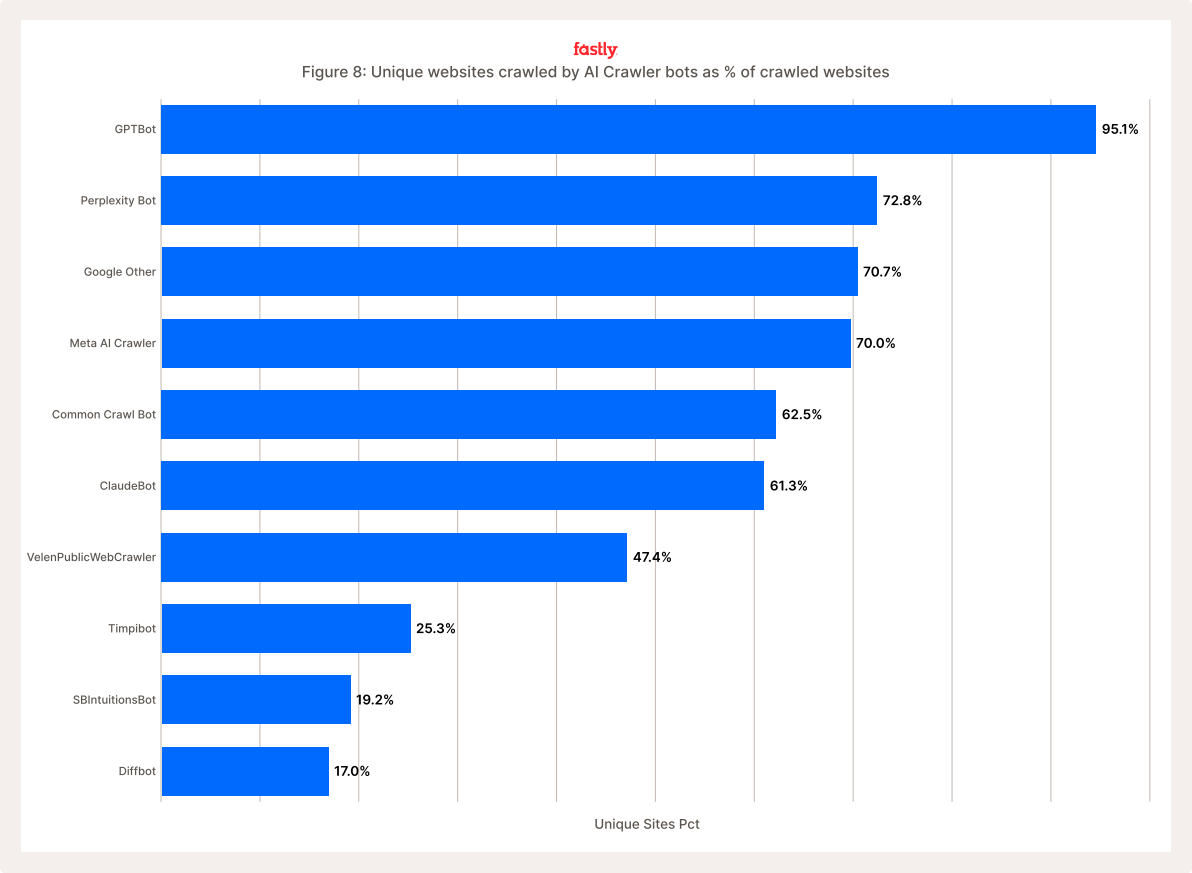
Content of interest



We observe that most AI crawler bots primarily target websites in the Commerce, Media & Entertainment, and High Tech sectors for scraping content (Figure 7). This likely reflects the high value of these domains in terms of fresh, dynamic, and information rich content such as product listings, news articles, reviews, and technical documentation, which are useful for training or grounding language models.

What's notable is that the top four crawlers (Meta, Google, OpenAI and Claude) seem to prefer Commerce websites. Common Crawl's *`CCBot`*, whose open data set is widely used, has a balanced preference for Commerce, Media & Entertainment and High Tech sectors. Its commercial equivalents *`Timpibot`* and *`Diffbot`* seem to have a high preference for Media & Entertainment, perhaps to complement what's available through Common Crawl.

Content Sources



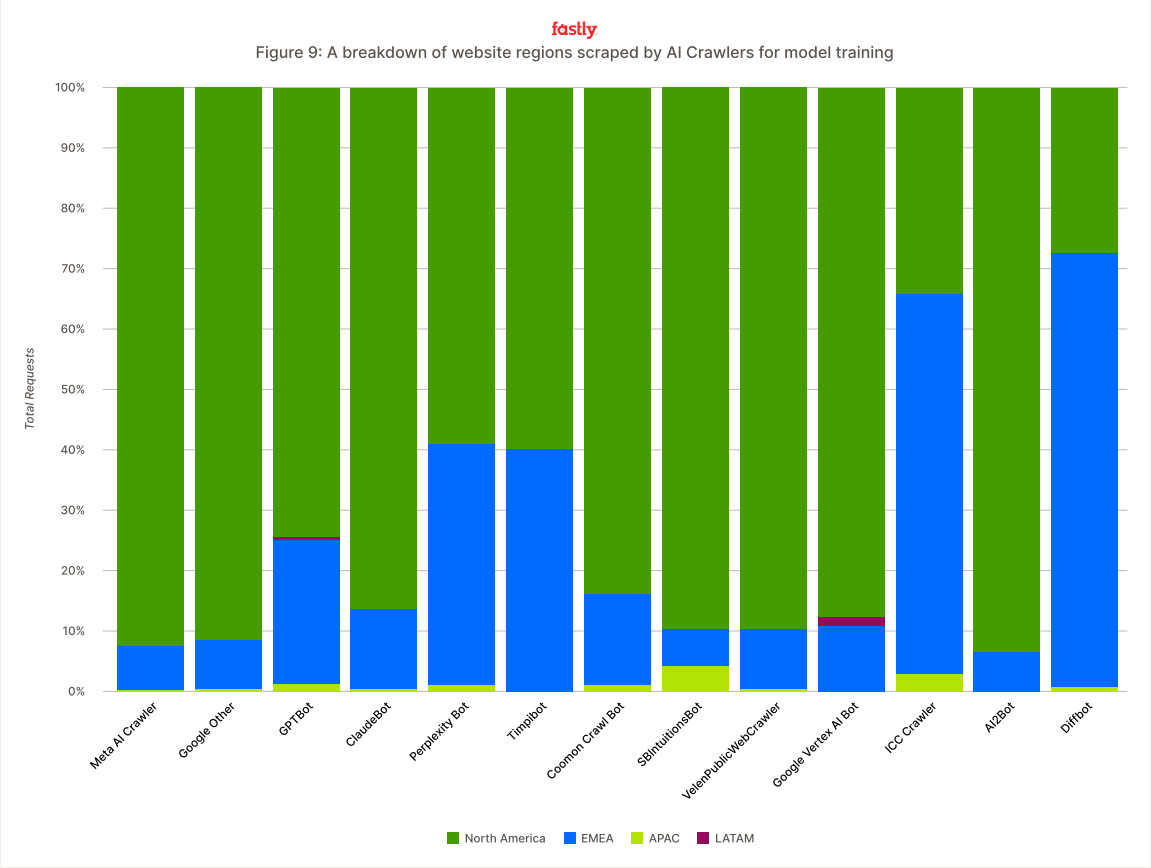
When analyzing AI bot crawling activity across a wide range of websites, OpenAI's *GPTBot* stands out for its breadth of coverage, indexing content from 95% of all websites crawled by AI bots (Figure 8). Interestingly, despite ranking third in total crawl request volume, after Meta and Google, at 20%, *GPTBot* crawls the largest number of unique websites. This wide coverage could give *GPTBot* a notable advantage in accessing diverse and unique content, which is essential for training high-performing language models.

Trailing *GPTBot* in unique site coverage are crawlers from Perplexity, Google, and Meta each reaching around 70%. Notably, Meta's crawler, despite generating the highest volume of requests, targets fewer sites, suggesting a deep indexing of select domains rather than broad, wide-scale coverage.

Commercial bots like *Timpibot* (25%) and *Diffbot* (17%), which specializes in selling crawl data to AI companies, currently access a significantly smaller share of the total set of websites crawled by AI bots. This reflects the varied approaches and scales of different organizations when it comes to content sourcing. In contrast, major players like Meta, OpenAI, and Google conduct broader indexing efforts, benefiting from their larger infrastructure. The differences in scale highlight the diversity of roles and capabilities across the AI crawling landscape.

Our observations also highlight the vital role of open data initiatives like Common Crawl. Unlike commercial crawlers, Common Crawl makes its data freely available to the public, helping create a more inclusive ecosystem for AI research and development. With coverage across 63% of the unique websites crawled by AI bots, substantially higher than most commercial alternatives, it plays a pivotal role in democratizing access to large-scale web data. This open-access model empowers a broader community of researchers and developers to train and improve AI models, fostering more diverse and widespread innovation in the field.

Content Regions



The geographic origin of training data can influence the alignment of AI models on various topics (Figure 9). This alignment is shaped by the inherent viewpoints and perspectives embedded within the data, including data derived from crawled websites. The content on these websites, in turn, is influenced by the cultural and geopolitical context of their respective countries or regions.

A significant observation is the apparent heavy reliance of most AI models on content sourced from North America. This concentration suggests a potential bias towards North American perspectives in their learned understanding. In contrast, specific crawlers such as *`Diffbot`* and *`ICC Crawler`* demonstrate a more diverse approach, standing out for indexing a higher proportion of content from the EMEA region.

The largest share of website content from the APAC region was indexed by crawlers operated by SoftBank, a prominent Japanese technology conglomerate, and NICT (National Institute of Information and Communications Technology), one of Japan's national research institutes. This highlights a regional concentration of data collection efforts, suggesting that AI models trained on APAC-specific data would likely reflect the dominant cultural and informational landscapes of countries within that region, particularly Japan.

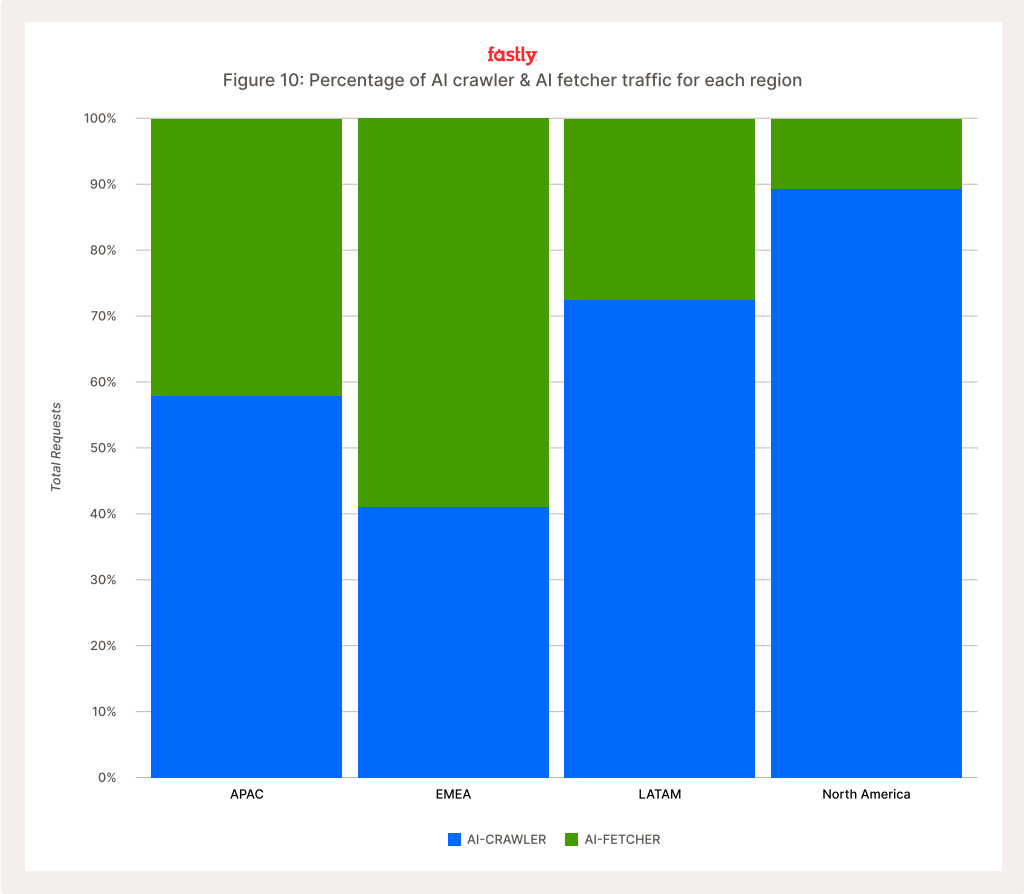
The implications of this data sourcing extend to how AI models interpret and respond to queries, potentially affecting their applicability and relevance across different global contexts.

AI Bot Activity Insights

The previous sections of this report represent AI bot traffic from the perspective of the individual bots – what websites, regions, and verticals the individual bots are targeting. In this section, we analyze patterns in AI bot activity from the perspective of a website operator – what types of AI bot traffic a website operator is likely to receive if they are in a specific region or vertical.

Regional differences in AI crawlers & fetchers

Website operators may have servers located in several regions, each of which experience AI bot traffic differently. These regional differences can help website operators understand which resources might experience traffic from different AI bots based on their location (Figure 10).

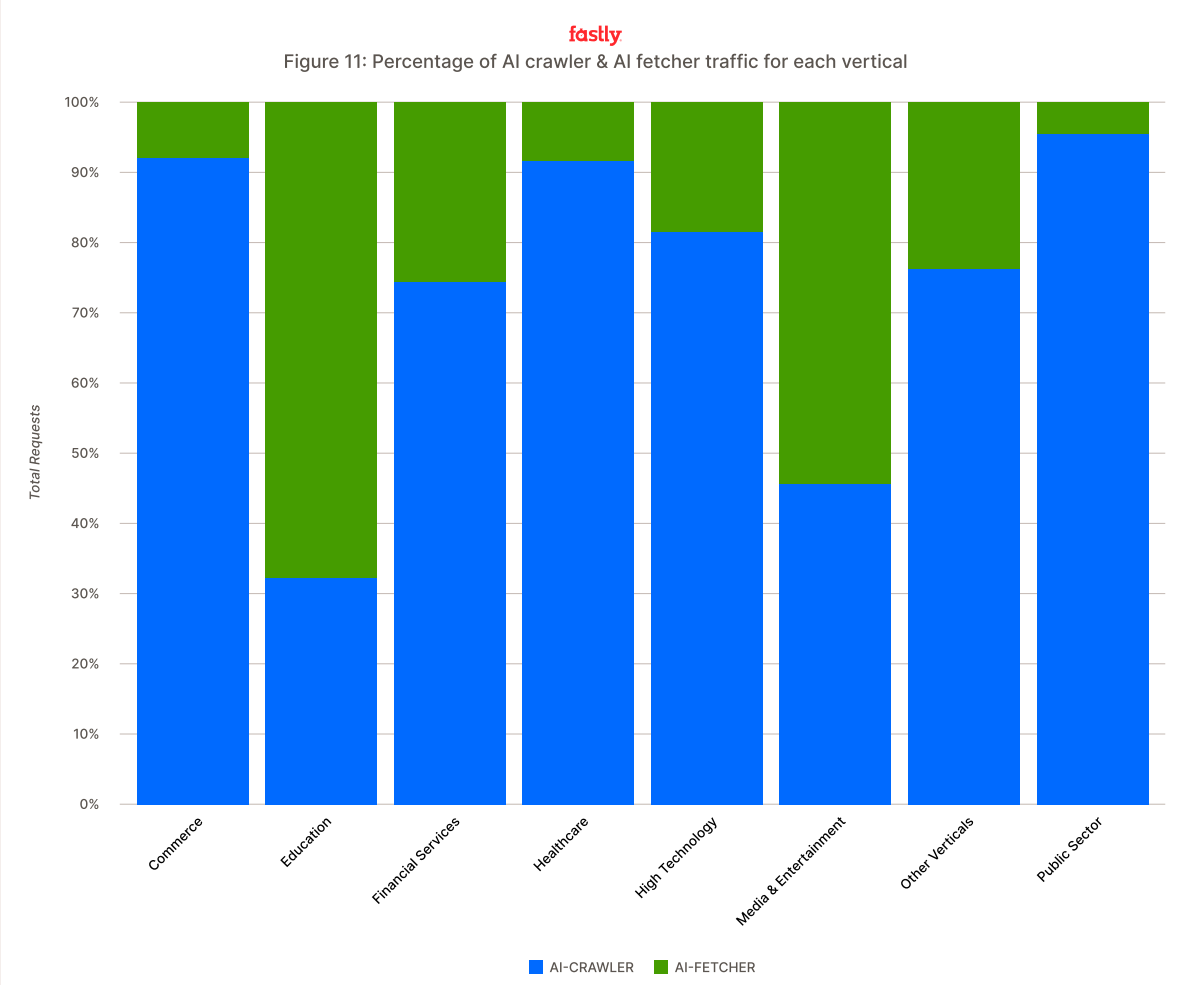


While overall breakdowns of AI crawler (80%) and AI fetcher (20%) traffic heavily skew towards AI crawlers, this does not hold true across various regions. While North America receives a heavy skew towards AI crawler traffic (almost 90%), every other region has a significantly smaller percentage of AI crawler traffic. Latin America (72%) and APAC (58%) still have a majority of their AI bot traffic coming from crawlers, but EMEA has less than half (41%).

With 59% of AI bot traffic coming from fetchers, operators in EMEA may want to focus their efforts towards managing those AI bots first, where operators in LATAM, APAC, and especially North America, will want to focus efforts towards crawlers, which represent a larger majority of their AI bot traffic.

Industry verticals by bot type

Similar to different regions experiencing AI crawlers and fetchers in different proportions, different industry verticals also have significant differences in their breakdowns of AI bot traffic (Figure 11).



Several verticals have AI bot traffic splits that are similar to the overall breakdown between crawlers and fetchers (80/20), including Financial Services (74%), and High Technology (81%). However, three verticals exceed the overall split with over 90% of their traffic coming from AI crawlers, specifically Commerce (92%), Healthcare (92%), and the Public Sector (96%). Operators in these verticals should focus on managing AI crawlers first, as the vast majority of their AI bot traffic is coming from these types of bots.

On the other hand, two industry verticals have significantly different patterns of AI bot traffic, with more than half of their overall AI bot traffic coming from fetchers. The Education (68% fetcher) and Media & Entertainment verticals (54% fetcher) have to account for traffic from both bot types, and particularly fetchers.

The prevalence of AI fetchers in the Education vertical is likely driven by students and researchers using tools such as ChatGPT for academic questions, which causes the fetcher to send additional traffic. The Media & Entertainment vertical is likely experiencing similar behavior from fetching more recent and frequently changing content from the news and current events. Website owners in these verticals may want to place a heavier emphasis on managing AI fetchers in order to help manage the increased load they experience from them compared to crawlers.

Recommendations and Actionable Advice

While some bots are well-behaved, others engage in abusive scraping. This can result in excessive load on web infrastructure, unauthorized use of website content, and skewed website analytics. We advise website owners to implement a layered strategy to safeguard their sites from abusive AI bots.

Web Standards based directives

Web standards offer several directives that specify which parts of a website are accessible to bots and which are not.

- Websites use *`robots.txt`* to request that bots limit access to specific sections. The following example *`robots.txt`* file asks OpenAI, Google, and Claude not to scrape the website for AI training, while still allowing Googlebot to index the site for search. Conversely, if a website intends for its content to be used for AI training, it can explicitly allow access or simply not list the bot in the *`robots.txt`* file at all.

```
User-agent: GPTBot
Disallow: /

User-agent: Google-Extended
Disallow: /

User-agent: ClaudeBot
Disallow: /

User-agent: googlebot
Allow: /
```

- To prevent bots from storing the page, include an HTTP response header (e.g., `X-Robots-Tag: noindex, nofollow`) or an HTML meta tag (e.g., `<meta name="robots" content="noindex, nofollow" />`).

While many bots comply with these methods, an abusive bot can disregard these instructions and continue to crawl and scrape content.

Implement Technical Controls

Website owners can deploy various controls to mitigate the impact of AI bots on web infrastructure. These controls include rate limiting, IP blocking, and challenges like CAPTCHA and Proof of Work systems (e.g. Anubis) to deny access or downgrade bot activity. However, care must be exercised when employing these techniques to avoid accidentally blocking legitimate users or downgrading their experience.

Advanced Bot Management Solutions

Advanced bot management solutions like Fastly Bot Management give website owners fine-grained control over which AI bots are allowed, how frequently they can access the site, and which content they can reach. These tools dynamically identify the growing number of AI bots in real time and provide comprehensive visibility into bot activity to help monitor and manage access effectively.

Where appropriate, AI bots can be redirected to a content licensing platform, enabling website owners to monetize their content in a secure and controlled way. This approach offers two key benefits: first, it ensures that AI bot traffic is managed within a defined framework, aligning content access with clearly defined usage guidelines; second, it creates new monetization opportunities for content creators.

Fastly's integration with Tollbit, a content licensing and monetization platform for AI usage, illustrates how site owners can license their content for uses like training AI models or generating AI outputs. By adopting this strategy, website owners can turn bot traffic from a resource burden into a revenue stream, an innovative response to the challenges of today's AI-driven creation and consumption of web content.

Good practices for AI bot operators

While the previous sections focused on recommendations for website owners, it's important to also highlight best practices for AI bot operators.

First, bots should always identify themselves transparently using a unique name, and link to a bot page. Although some bots disguise their identity deliberately by using `User-Agent` strings of regular web browsers, this deception often causes even desirable AI bots to be classified as malicious bots and subsequently blocked by website defenses.

To facilitate bot verification, operators should publish their IP address ranges or support

verification methods such as reverse DNS lookups. OpenAI's AI bots are a strong example, as they publicly share their IP ranges, making it easy for website owners to identify and verify them.

AI bots should honor the *robots.txt* protocol and any explicit web standard based opt-outs defined by websites. Respecting these directives helps ensure that bots only access content the website owner permits for use in AI training, builds trust with website owners, and reduces the likelihood of being blocked or misclassified. Google is well known for honoring such directives.

It is also essential for bot operators to implement internal rate limiting and traffic controls to ensure their bots operate within reasonable request limits, preventing excessive load on web infrastructure. This applies not only to AI crawlers but also to AI fetcher bots, which we've observed can generate even higher traffic volumes, particularly during spikes driven by increased AI product usage or newsworthy events.

Some organizations use a single crawler for both search engine indexing and AI training purposes, allowing websites to differentiate between use cases through specific *User–Agent* directives in *robots.txt*. While this approach can reduce redundant crawling and conserve resources, it also introduces limitations.

When the same bot is used for both purposes, website owners lose the ability to apply distinct handling rules, such as redirecting AI crawlers to content licensing platforms or applying different access controls. This blending of functions can create operational and policy challenges for site administrators. To promote clearer separation of intent and better control for website operators, organizations should consider using separate bots for search indexing, AI crawling, and AI fetching.

Finally, publishing crawl schedules when possible helps website owners prepare and manage capacity. Common Crawl's *CCBot* is a notable example, crawling sites on a predictable two-week cycle each month, as discussed earlier in this report.

Safer, Faster, Engaging and AI-Ready

AI bots have emerged as a transformative yet complex force across the web. Their rapid growth has introduced new challenges for infrastructure, security, and content ownership. As AI-driven traffic continues to scale, it is vital for organizations to manage it precisely. Fastly Bot Management provides the visibility and control needed to distinguish between helpful and harmful bot activity in real time. When it comes to bot operators, transparent intent, verifiable identification, adherence to standards, and responsible crawling can help strike a balance between innovation, fair content use, and preserving control for website owners. Ultimately, adapting to this evolving landscape will be key to safeguarding digital assets and unlocking new opportunities.