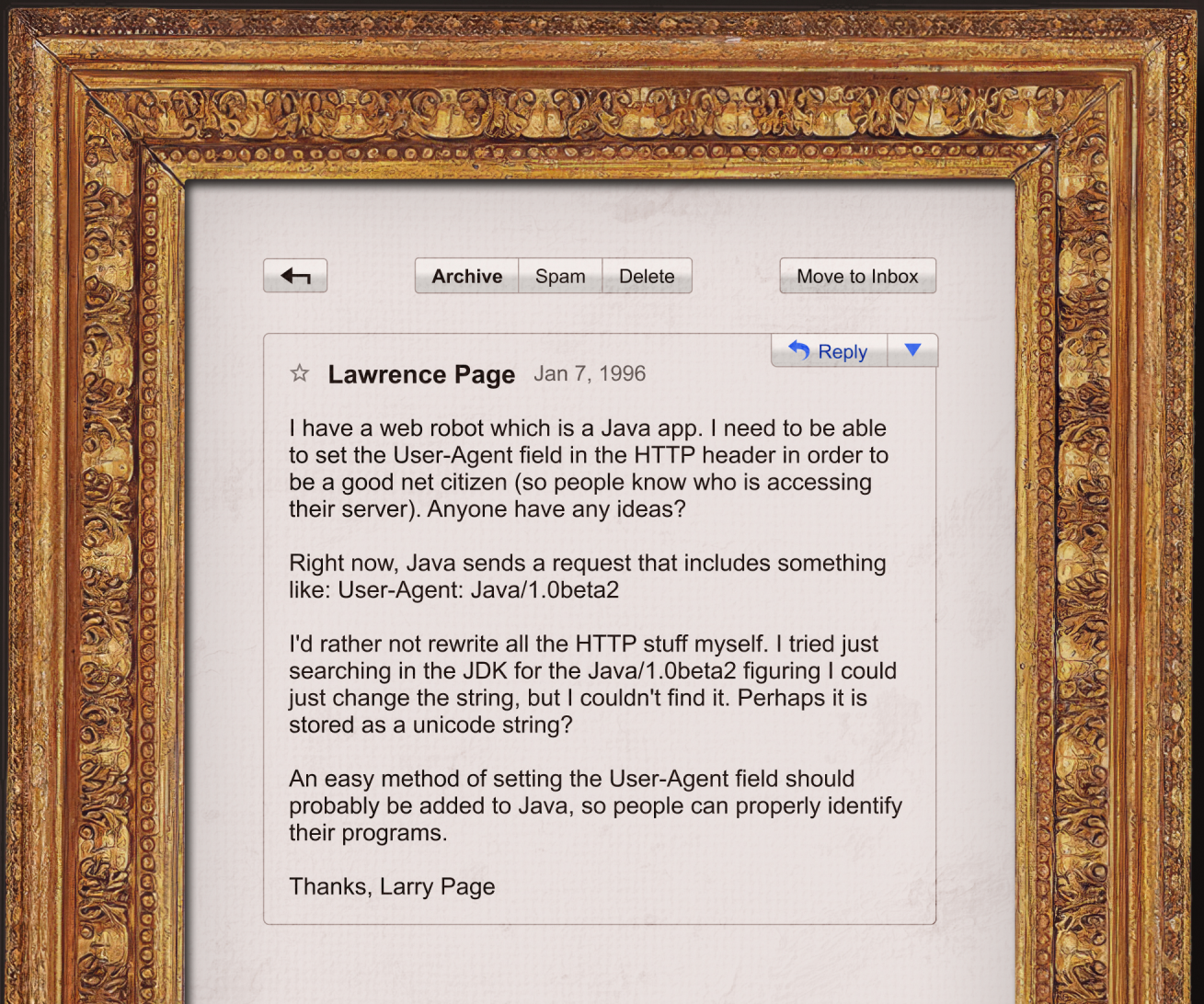
Good Net Citizens *c. Jan 1996*

In a 1996 Java forum post, Google cofounder Larry Page asked how to set the user agent on his web crawler. He argued that programmatic visitors should identify themselves in order to be "good net citizens," allowing websites to know what programs were accessing them. How non-human visitors present themselves has been a consideration since the earliest days of the web. Nearly thirty years later, the same question looms—only now the answer will shape the future of the web.



Executive Summary

- The next wave of AI visitors won't look like bots or agents; they're increasingly looking like humans. The latest AI browsers like Perplexity Comet, and devtools like Firecrawl or Browserless are indistinguishable from humans in site logs: Chrome appears as their user agent and they behave like human users — loading pages, sometimes loading ads, and solving CAPTCHAs. This is why it is mission-critical to push for a future where non-human site traffic is required to self-identify.
- AI bot activity now surpasses crawls from Bingbot: AI bot aggregate traffic has surpassed that of the world's second-largest search engine, Bing — highlighting the growing scale of AI adoption relative to traditional search engines.
- Google's expansion of 'AI Overviews' in October 2024 resulted in 34.8% increase in Googlebot crawls. At the same time, the number of Google crawls needed to get a human visitor got worse, increasing by 24.4%. The overall increase in AI visitors puts further economic strains on sites, as their CDN costs increase to support this additional bot traffic amidst declining human referrals.

- Human visitors to sites are falling, while AI bot traffic is rising: Between Q1 and Q2 2025, there was a 9.4% decrease in human visitors across sites on TollBit. Meanwhile, AI bot/agent traffic is continuing to increase. Looking at human and AI visitors across sites, at the start of Q1, 1 out of every 200 visitors was AI. Now it is 1 in every 50, reflecting a 4× increase in the relative volume of AI visitors.

- Over the past year, there has been a 336% increase in sites blocking or redirecting AI bots/agents, as evidenced by redirections and HTTP forbidden errors served to AI bots. This is a result of bot detection solutions being deployed more extensively across publisher sites. This corresponds with hits to the TollBit Bot/Agent Paywall, which have increased around 360% from Q1 to Q2 2025 as websites have increasingly taken active measures to charge AI traffic.

- Across all AI bots, 13.26% of requests bypassed robots.txt in Q2 2025, up 4× from just 3.3% in Q4 2024.

- There is evidence that AI apps store & re-use content, but it appears it's not always efficient for them to cache data for long periods. This highlights the importance of RAG and continued content access. Here's a breakdown of what that caching pattern looks like based on our tests:

    A. Claude: cached for 16+ days*

    B. ChatGPT: cached for 30 minutes

    C. Gemini: cached for 15 minutes

*Our testing ended after this time period.

Section 1

Rise of the agentic web

Key Insights

This quarter saw the release of additional AI web browsers and new consumer agentic applications. As adoption grows, these look set to radically change the shape of web traffic with profound consequences for digital business models.

New AI applications and capabilities are increasingly being powered by ‘headless’ browsers; these are web browsers that are controlled by automated systems. Headless browsers are a more interactive form of web scraping that enables AI agents to take actions on behalf of users, including accessing dynamic websites, adding items to their cart, or booking a table. There is no human user controlling the browser, yet the traffic looks like a human visitor - despite being an agent.

TollBit data suggests that human visitors to websites are beginning to decline, with bots supplanting them. Assuming the use of headless browsers continues to grow, this trend is expected to falsely appear to reverse, with website owners seeing a growth in visitors that appear to be humans but are in fact agents serving an AI application.

This creates a new technical challenge for publishers. Since non-human traffic cannot be monetized via subscriptions or advertising, distinguishing human from non-human traffic is of strategic importance for content businesses.

It is our view at TollBit that regulatory intervention is needed to ensure the agentic ecosystem sustains the entire information value chain. Automated agents should be mandated to identify themselves. Presenting a bot as a human has no legitimate justification and should be prohibited, with disclosure embedded in user agent strings or elsewhere in the HTTP headers.

The second quarter of 2025 saw the mainstream or beta releases of AI web browsers such as OpenAI Agent Mode, Perplexity Comet, Google's Project Mariner. Fortune 1000 companies also turn to platforms like MindStudio or n8n to build agentic systems. These new technologies - which further extend the capabilities of automated systems to gather information and perform actions on our behalf - foreshadow a transformative change in how consumers engage with the online world.

While these changes are still nascent, and the signals in the data are only beginning to emerge, we can already see the contours of a future in which interactions with websites look radically different. This shift means traffic will present itself to publishers in different ways with direct implications for revenue, raising fundamental questions about the relationships between website owners, users, and the developers of AI systems that increasingly occupy an intermediary role. For these reasons, we are dedicating the first section of our Q2 State of the Bots report to the rise of agentic systems, how they will reshape online traffic, and what this means for publishers.

Headless browsers and agentic systems

Since the release of chatbots that can access the real-time internet, AI applications have needed systems that interface with websites, retrieving and synthesizing content on behalf of users when required by a prompt. In the first wave of AI this function was served by purpose-built crawlers operated by the AI developer itself. These bots declared themselves via user agent strings such as ChatGPT-User and PerplexityBot (for user agent profiles, see the appendix to this report). Much of the analysis in TollBit's State of the Bots reports has focused on these technologies.

This landscape is now changing. New AI interfaces and capabilities (chiefly agentic systems and AI browsers) and the widespread adoption of publisher IP controls (see section 5) have led developers to turn to the use of 'headless browsers', often operated by third parties. These are web browsers that are programmatically controlled via automated systems, without a human doing the browsing - hence the term, "headless." This tech has been available for many years, but until recently, was almost exclusively used for enterprise functions, such as website performance testing or QA automation.

Three types of AI headless browsers

There are three distinct categories of headless browsers:

Masked web agents

Infrastructure providers allow an AI developer's agent to access websites through headless browsers at scale. Think of this as renting browsers in the cloud, each controlled by the AI developer's agent. These visitors appear to publishers as indistinguishable from ordinary human traffic. Rather than announcing themselves via their user agent string, they typically present as a standard Chromium (Google's open-source web browser engine) visitor. Providers of this infrastructure include Browserbase and Hyperbrowser.

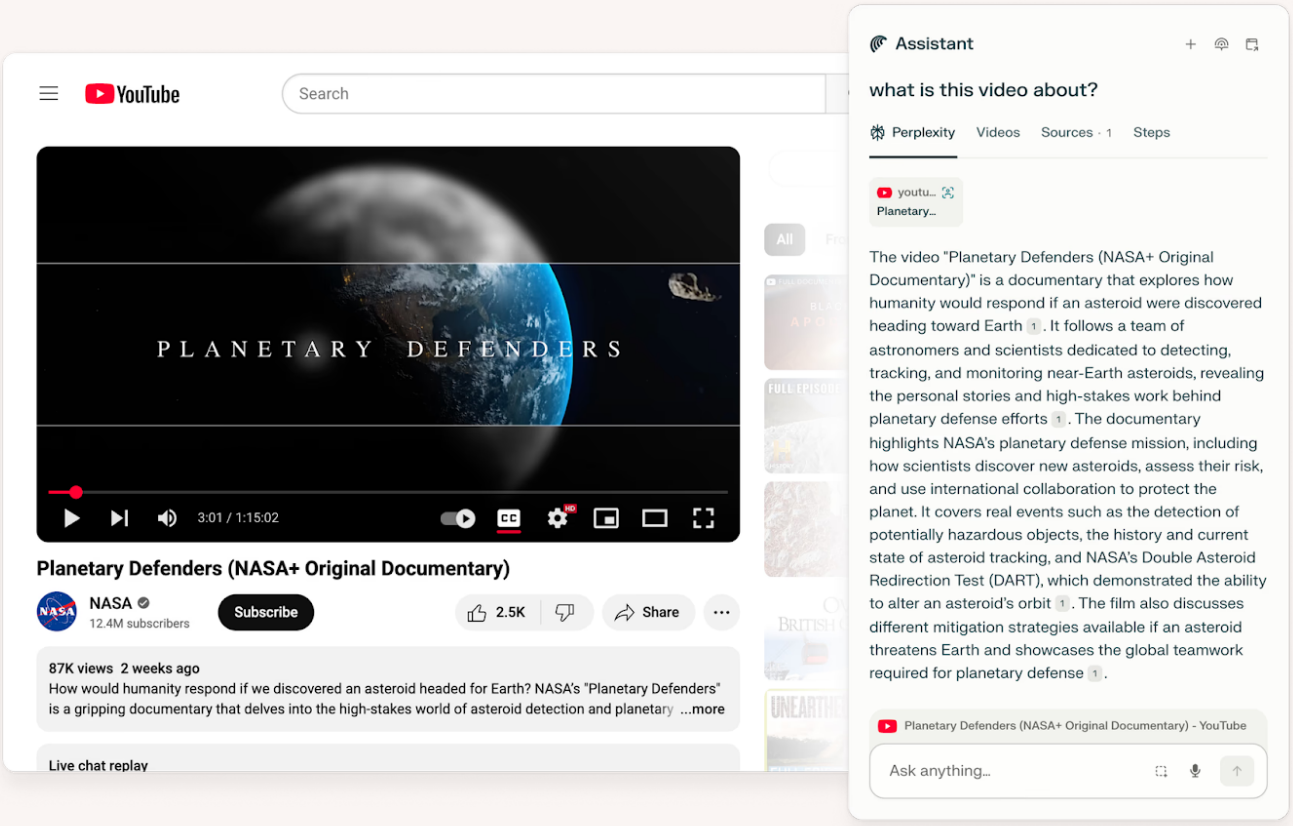
Third-party web agents

These services (e.g. Firecrawl, Apify, Zyte) act as outsourced crawlers. Instead of the AI agent itself opening a browser session (as with masked fetchers), the agent calls an API, and the third-party service performs the crawl. The AI system then consumes the processed output without ever touching the publisher's site directly. Like masked web agents, third-party web agents insert a commercial intermediary between the publisher and the AI developer; however, these go further and outsource the bot activity altogether. Part of the value prop of these services is their ability to defeat anti-bot tools to scrape high-value sites.

Browser-driven web agents

With the release of the AI-first web browsers - Comet from Perplexity and Dia from The Browser Company - come agents that operate directly inside a consumer-facing browser, using the browser itself as the interface for real-time web access (see figure 1.1). Rather than calling an API or running through a headless session, the agent automates navigation, search, and form-filling within the live browser window. From a publisher's perspective, these agents are indistinguishable from normal human browsing: the requests originate from the end-user's browser, but the actions may be initiated or guided by the AI agent, and the human behind the request may never see or interact with a visited page.

Figure 1.1. Perplexity's Comet browser



Headless browsers are being used because they provide a more reliable way to automate web access for AI developers. They execute JavaScript, maintain sessions, and mimic normal user behaviour, enabling agents to complete tasks that simple HTTP-based scrapers cannot. In practice, they can also bypass common bot-mitigation measures.

In the case of Perplexity Comet, the agent appears to use the user's own desktop to request webpages when automating workflows. This makes the requests appear to come from a user's residential IP address, which can confuse CDN companies into thinking it is legitimate. Comet is also built on the open source Chromium project, which powers Chrome, so the requests also pass along Chrome user agent strings. While this user agent string is technically correct, it does little to help distinguish automated web browsing from legitimate human traffic.

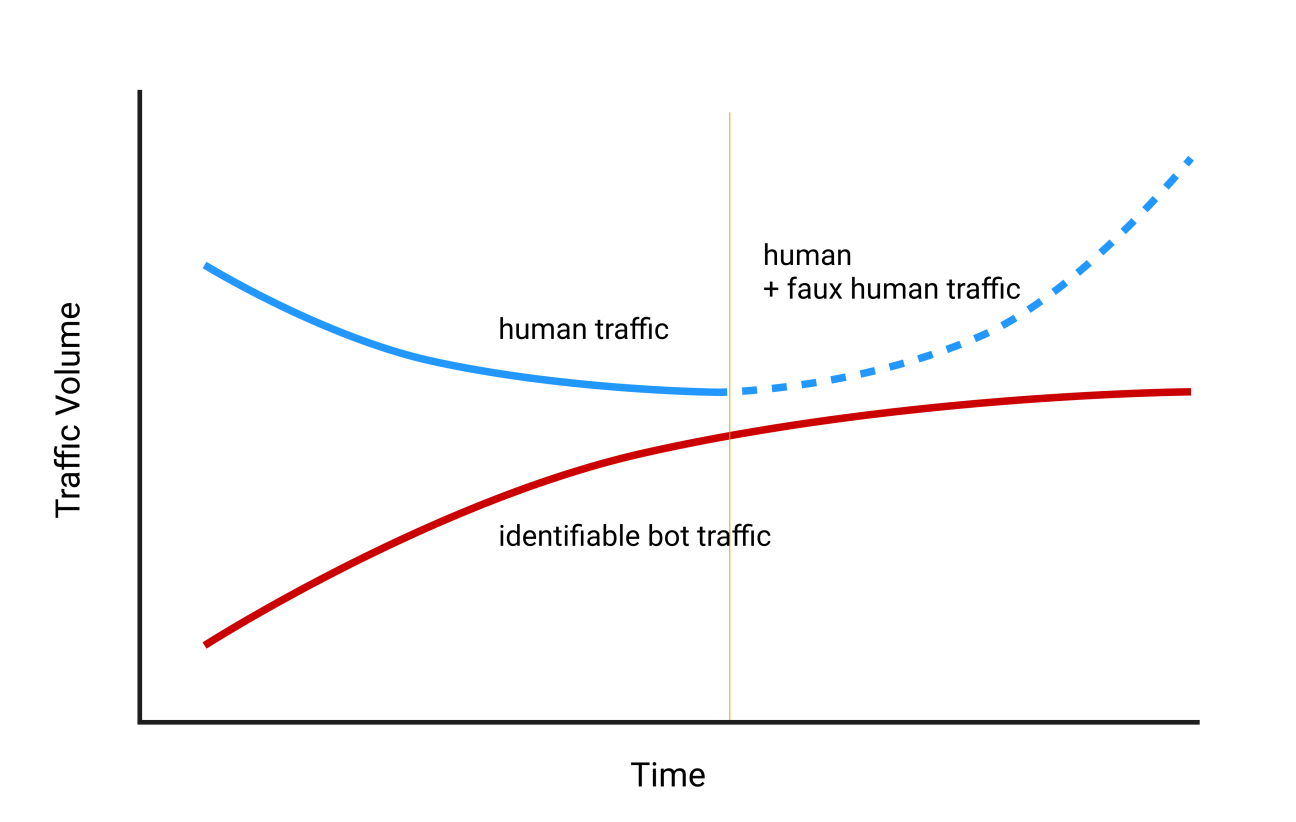
These practices create a new technical challenge for publishers. Some headless browser services can evade detection tools and even solve CAPTCHAs (tests intended to distinguish humans from bots), making automated sessions hard to separate from genuine human visitors. In some respects, this is the latest battle in the ongoing war between website owners seeking to protect their data and those attempting to scrape content for AI use. However, some applications also blur the boundary between human and automated visits, complicating the assessment of whether certain categories of traffic create value or impose cost.

Future traffic patterns

Whether a publisher, an ecommerce or an aggregator site, their advertising and subscription revenues depend upon humans engaging with content on properties. When AI systems scrape and summarize (or browse and checkout), they potentially negate the need for a user to visit the site. This corresponds to a lost monetization opportunity for the content creator. Distinguishing human from non-human traffic is therefore of profound strategic importance for digital publishing businesses.

TollBit data suggests that human traffic to publisher websites is beginning to decline, with bots supplanting it (see section 2). As the use of headless browsers (or advanced web scrapers) grows, this trend is expected to falsely appear to reverse, with website owners seeing a growth in visitors that present as humans but are in fact agents serving an AI application (see figure 1.2). This is due to headless browsers being indistinguishable from human visitors on sites. These faux human visitors will erode advertiser trust and cannot be monetized with traditional methods.

Figure 1.2. Projected future traffic composition



Looking further to the future, as agentic systems grow in complexity and sophistication (which in turn will rely on improvements in reliability and speed), this is likely to result in a dramatic expansion in this faux human traffic - far exceeding the levels seen today - with multiple agents serving a single user simultaneously; many likely to be operating autonomously and without the need for a user-trigger.

Agent-to-web protocols, distinguishing humans from bots

Ironically, to maintain the integrity of the open web, it may need to be that a parallel 'agent web' will emerge, optimized for non-human actors. This would negate the need for headless browsers to engage with human websites with all the inefficiencies that entails. This is certainly a possibility with emerging approaches such as the Model Context Protocol (MCP), structured data standards like Schema.org, and TollBit's own agent gateway, creating the means of discrete agent-to-site interfaces.

Should agent-to-site protocols become the norm, it would simplify the process of distinguishing between humans and non-humans; the onus of figuring out "bot or not" would no longer be an unfair burden placed on website owners in a web that is increasingly trafficked by autonomous visitors.

However, the novel AI applications that these systems serve are also making it harder to draw a clear distinction between these visitors. For example, if a real human is directing an AI browser that opens a tab unseen by the user, is that human traffic? Or if an agentic system books a table at a restaurant on the explicit instruction of its user, should that be considered a human visitor? (We think not, hence why we refer to this above as “faux human” traffic.)

Though there are too many unknowns now to develop a detailed plan, some strategic principles and regulatory demands are already clear.

Regulation and policy

As this section of the report has described, the rise of headless browsers and the blurring of human and non-human traffic pose clear risks for all website owners. For some publishers, value is being eroded by intellectual property leakage. For other sites, resources are diverted into costly detection, mitigation or compute costs. At the same time, these practices by AI developers are stalling the emergence of a fair and liquid content access market.

While we can already see distinctions between agentic systems, agentic automated systems, user-directed AI browsers, automated AI browsers, etc., it is clear that this taxonomy will evolve, and edge cases that blur the human-bot boundary will continue to appear. This is not a reason for inaction, though. The principle that needs establishing, underpinned by legal force, is that a website owner understands the source of a request and can serve it according to its commercial strategy. Such a rule would lay the foundations for an agentic future that sustains the entire information value chain.

It is our view at TollBit that this is where regulatory intervention is needed. As a minimum, automated agents should be mandated to identify themselves. There is no legitimate justification for any autonomous visitor on the web to present itself as human; this should be prohibited, requiring disclosure/identification embedded in user agent strings or elsewhere in the HTTP headers.

Section 2

Scale of AI scraping

Key Insights

Traffic from AI bots continue to grow strongly both in absolute terms and as a proportion of overall traffic. Out of human & AI visitor totals, we now see 1 out of every 50 visitors to a site is an AI visitor. It was 1 out of every 200 visitors to a site in Q1.

Alongside this growth in AI traffic, the number of human visitors to websites is beginning to fall, TollBit data saw a 9.4% reduction between Q1 and Q2 2025. This suggests that human visitors to websites are being replaced by AI bots, operating on their behalf.

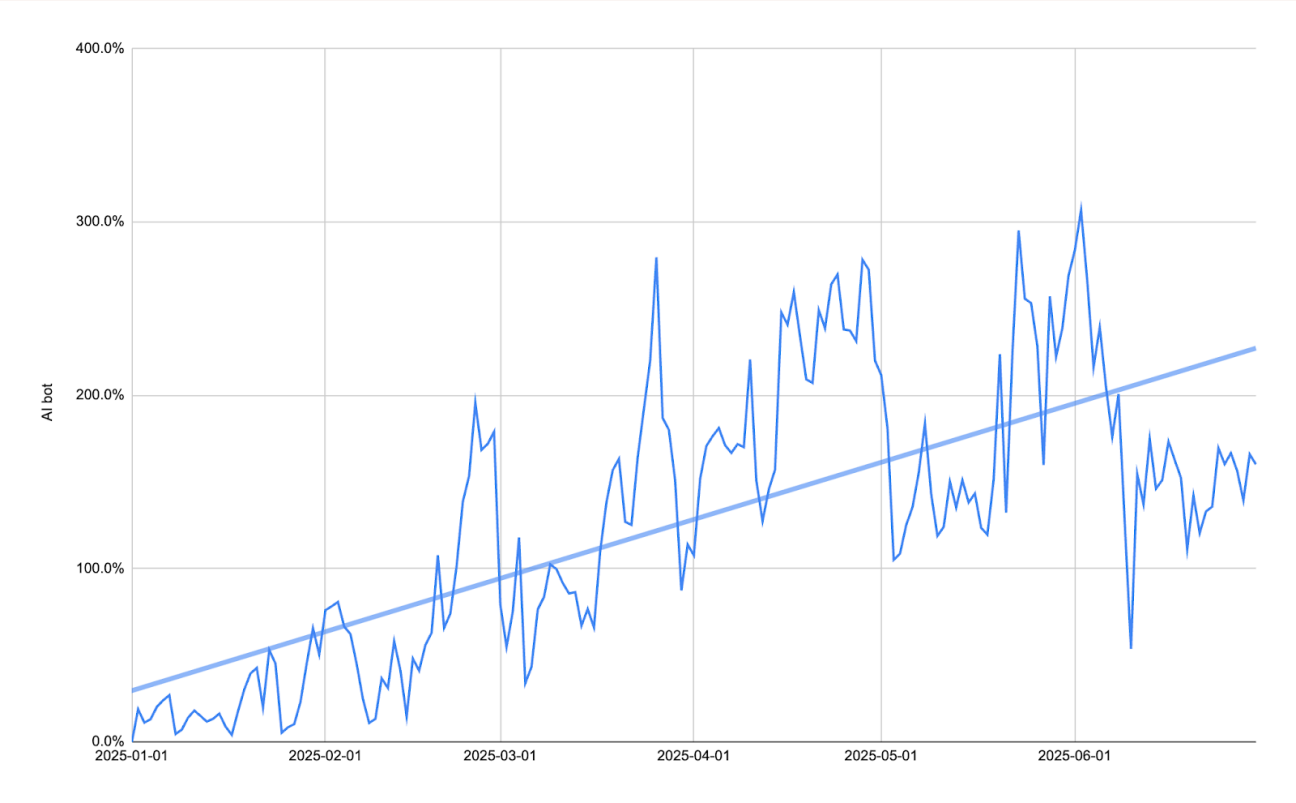
In aggregate AI bot requests have now surpassed those from Bing, the world's second-largest search engine that currently accounts for around 4% of the market.

Google's expansion of 'AI Overviews' in October 2024 resulted in 34.8% increase in Googlebot crawls. At the same time, the number of Google crawls needed to get a human visitor got worse, increasing by 24.4%. The overall increase in AI visitors puts further economic strains on sites, as their CDN costs increase to support this additional bot traffic amidst declining human referrals.

Aggregate scraping levels

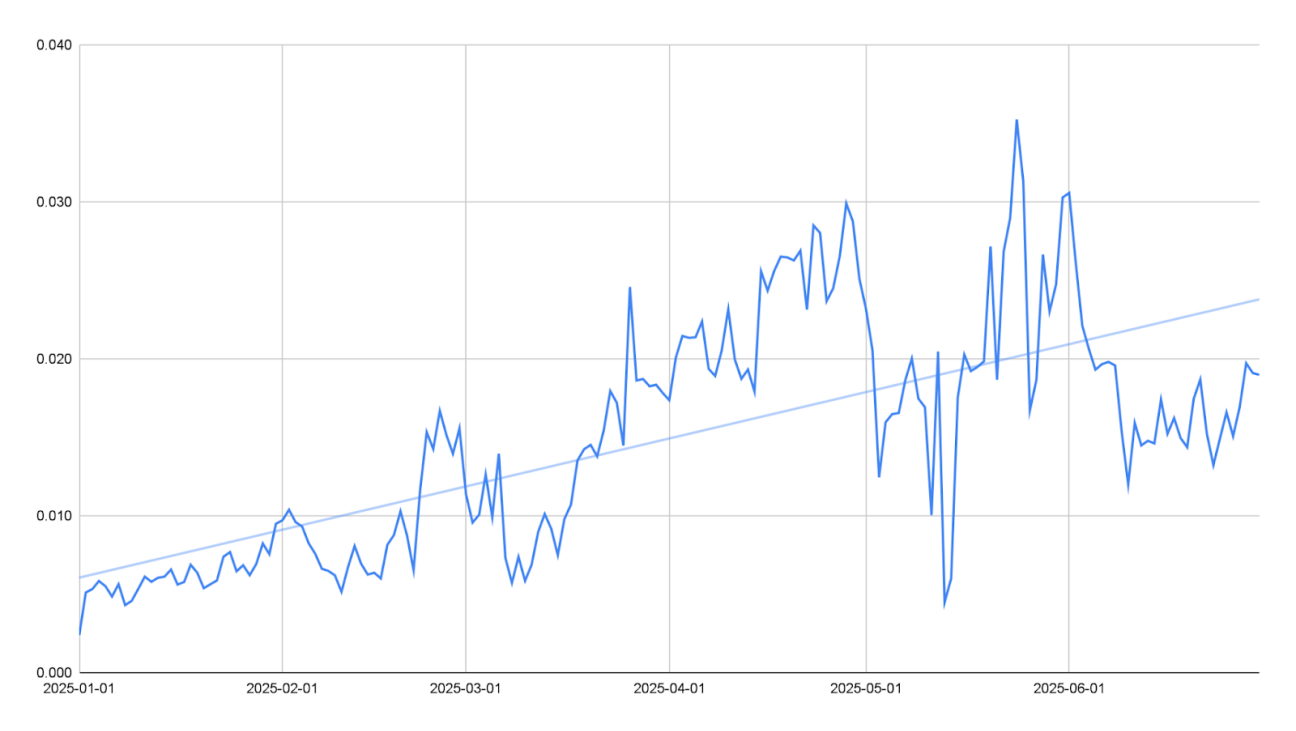
Over the second quarter of 2025, website traffic from AI bots has continued to grow strongly. When we examine the daily volume of AI bot requests per TollBit partner website since the start of the year (figure 2.1) we can see a sharp increase, with, on some days, the level exceeding 300% what it was at the beginning of 2025.

Figure 2.1. Daily per website AI bot scraping level, percent vs 1/1/2025



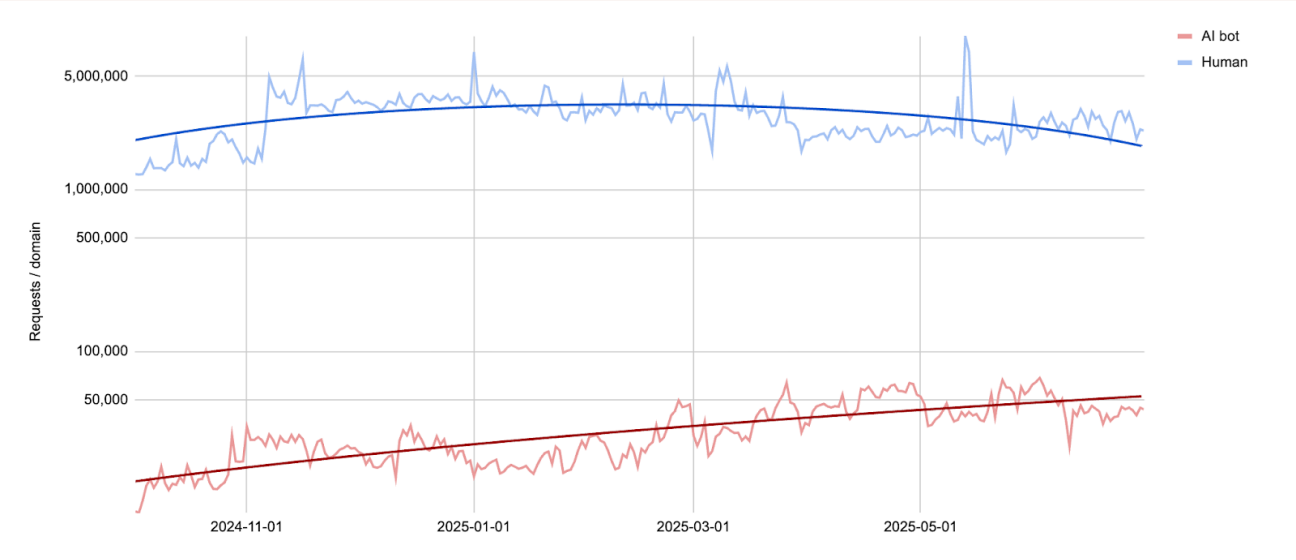
AI bot traffic also continues to grow as a proportion of all web traffic. Looking at human and AI visitors across sites, at the start of 2025, 1 out of every 200 visitors was AI. Now it is 1 in every 50, reflecting a 4× increase in the relative volume of AI visitors. (see figure 2.2).

Figure 2.2. AI to human ratio, total requests



Two factors are driving this change in publisher traffic composition. Alongside the growth in AI bot traffic we are now beginning to see human visitors decline. TollBit data saw a 9.4% drop in human requests between Q1 and Q2 2025. Put together with the AI bot traffic growth, these data suggest that substitution could be taking place - human visitors to websites being replaced by AI user agents which are operating on their behalf (see figure 2.3) whether that's through a chatbot or an AI search interface. Unless publishers can monetize this AI traffic, it might result in lost revenue.

Figure 2.3. Human vs AI bot requests, per website (log scale)

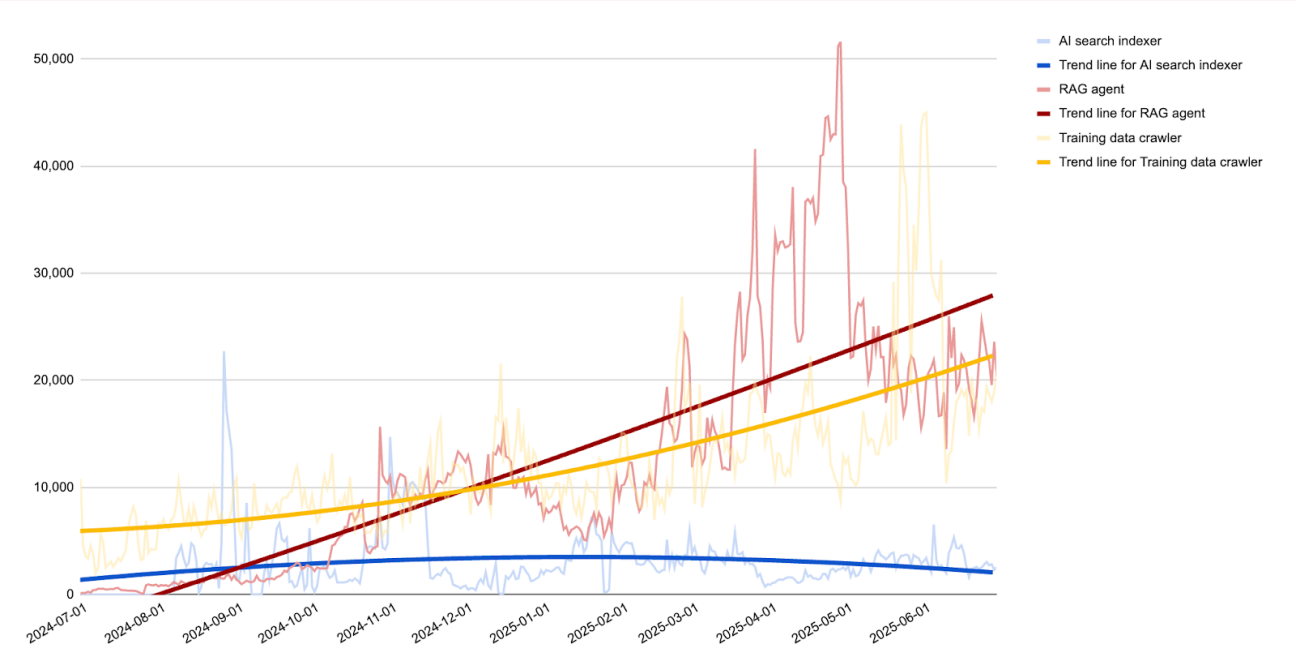


Scraping levels per AI bot category

This substitutional effect may be driven by consumers using AI applications for information retrieval, likely needs that would previously have been satisfied by the use of a search engine followed by a human visiting a publisher website. When AI applications serve these needs, they often access the web in real-time, using a retrieval-augmented generation (RAG) agent to visit and scrape websites to formulate a response. TollBit data shows that website visits from these AI RAG agents are growing the fastest (figure 2.4).

These increases in AI Bot traffic hammering sites are concerning for publishers, not only due to the potential loss in human visitors but also because this is driving up their website/CDN costs in order to support this additional traffic. AI tools often “read”, aka use & cite more content than a human would to answer a question.

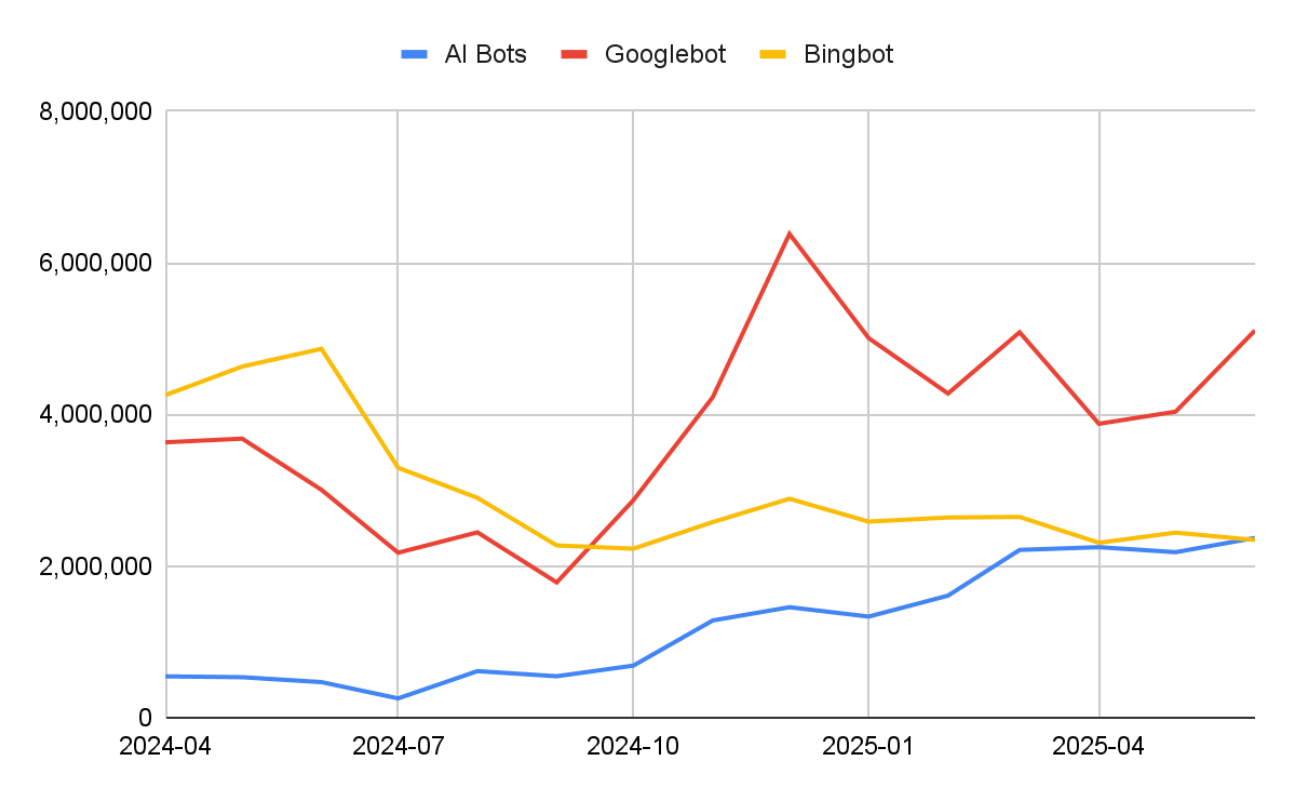
Figure 2.4. AI bot traffic by user-agent type (average, per domain)



AI and online search

When we compare the level of scraping from these new AI bots to those from the web crawlers operated by conventional search engines, we can see that, in aggregate, AI requests per website have now surpassed those from Bing, the world's second largest search engine, currently holding around 4% of the market[1] (figure 2.5).

Figure 2.5. Requests per site AI bot vs conventional search bots

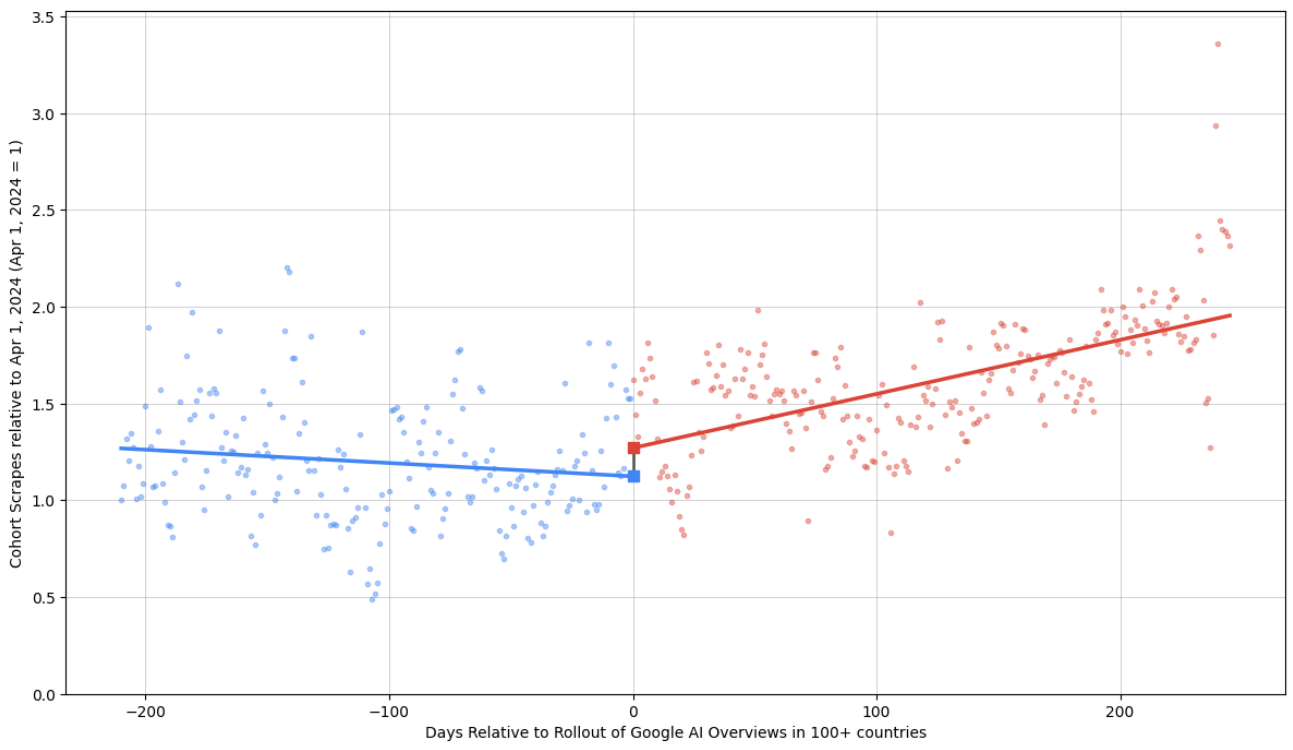


[1] Statcounter Global Stats (2025) *Search engine market share worldwide*. Available at: https://gs.statcounter.com/search-engine-market-share (Accessed: 9 August 2025).

In response to the growing use of AI chatbots for information retrieval, Google has been introducing AI features into its core web search product. This started with a trial of 'search generative experience' back in May 2023. This feature turned into 'AI Overviews' which were piloted, expanded to all US users then rolled-out extensively rolled-out - to over 100 countries - in October 2024.

This global expansion of AI Overviews coincided with a very substantial increase in the level of requests from Google's crawler on TollBit partner websites (figure 2.6). The daily average crawls for the cohort of websites increased by 34.8% after the rollout. This data strongly suggests that Google still needs to visit websites in real-time to power its AI search features.

Figure 2.6. Googlebot requests and the global rollout of AI Overviews (0 = 10/28/2024[2])



[2] Google (2024) *AI Overviews in Search are coming to more places around the world*. Google Blog. Available at: https://blog.google/products/search/ai-overviews-search-october-2024/ (Accessed: 9 August 2025)

A new user agent from Google?

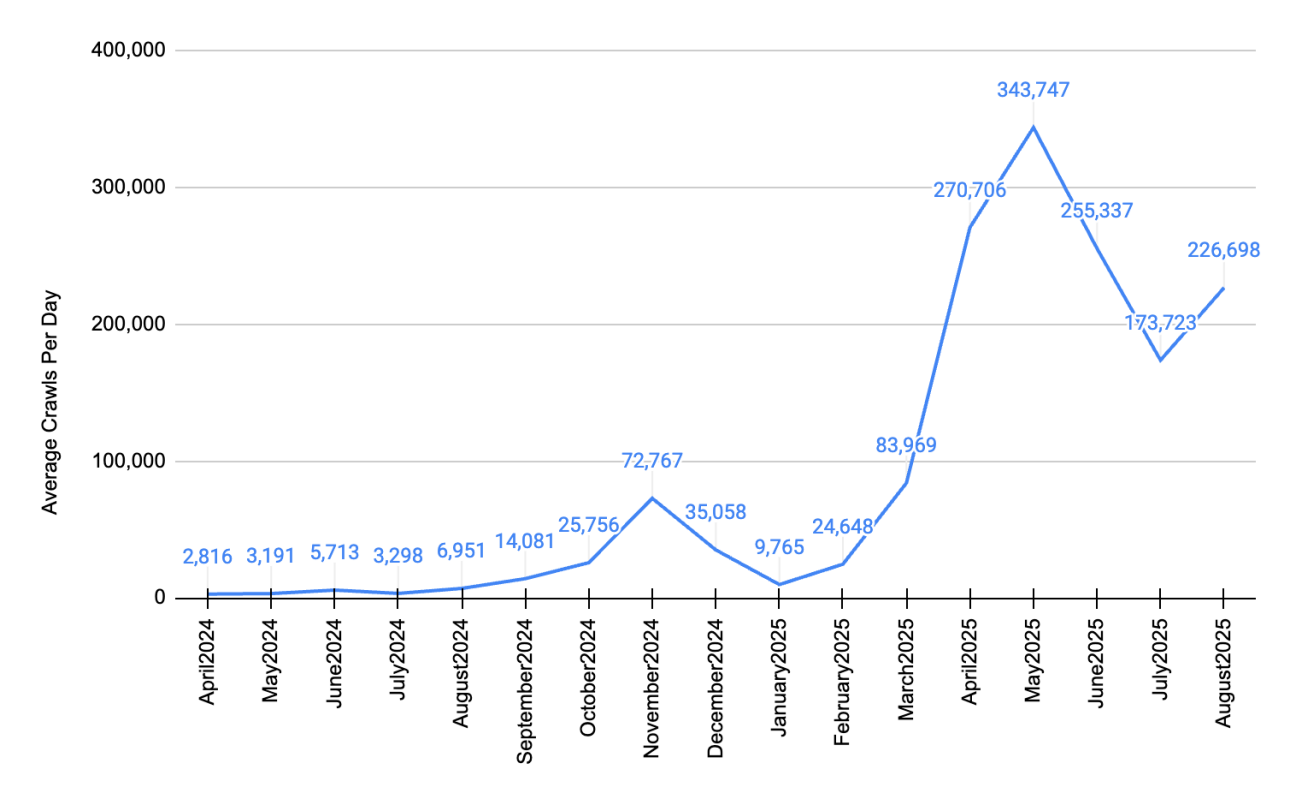
We saw ~45M requests across our publisher network from a user agent string 'Google' and some IP addresses associated with Google. We observe this user agent appearing in server logs when we use Gemini to search the website. However, Google's documentation at the time of publishing this report doesn't mention this crawler.[3]

While this IP is not listed in Google's published IP ranges, a reverse DNS lookup confirmed the IP addresses were Google, as per Google's verification guidelines.

We first observed requests from 'Google' user agent in our logs in April 2024, followed by a noticeable uptick in November 2024 and a significant rise in average monthly visits starting April 2025. These increases align with the launch of the Gemini iPhone app in November 2024[4] and the rollout of AI Mode and Deep Search features in May 2025.[5] (see figure 2.7)

[3] Google crawlers documentation. Available at: https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers
[4] Gemini App available on iPhone. Available at: https://blog.google/products/gemini/gemini-iphone-app/ (Accessed: 26 August 2025)
[5] Google Deep Search and more AI features laucnhed. Available at: https://techcrunch.com/2025/05/20/googles-ai-mode-rolls-out-to-us-will-add-support-for-deeper-research-comparison-shopping-and-more/ (Accessed: 26 August 2025)

Figure 2.7. Daily Average Crawls from 'Google' user agent

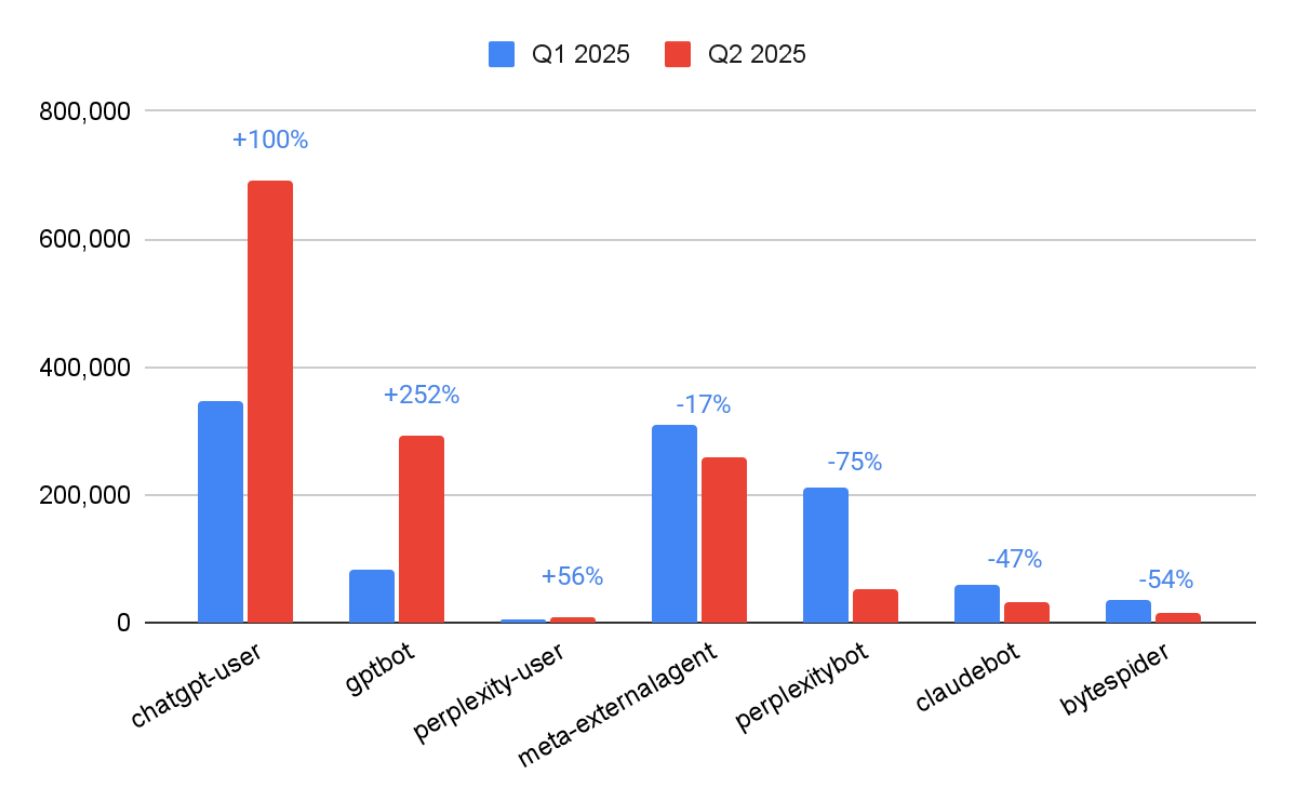


User agent scraping levels

When we examine the level of scraping by individual AI bots (figure 2.7), we can observe a notable increase in activity from OpenAI's web crawlers, both ChatGPT-User - its RAG agent - which has increased by 100% on a per-website basis, and GPTBot - its training data crawler - which has grown by 252%.

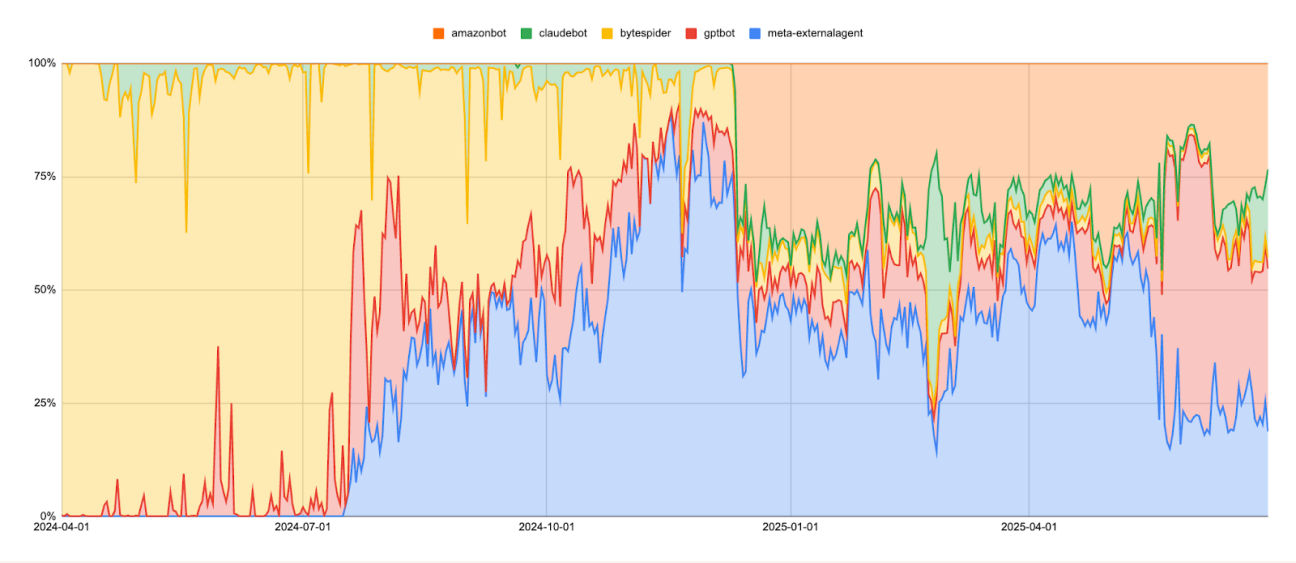
Notably, aggregate scrapes from Perplexity's user agents (PerplexityBot and Perplexity-User, introduced earlier this year) have decreased by 19% over the quarter. This could be due to alternative data acquisition methods, like third-party web scrapers or disguised user agents.

Figure 2.8. Monthly Average Scrapes per site by AI Bots – All sites from Q4 '24 vs Q1 '25



Looking specifically at patterns of activity from training data crawlers, we see periods of intense scraping activity from a particular user agent, followed by relative dormancy. This is likely to be driven by the need to collect data for a specific model training run. In the period from late May to the end of June we saw particularly intensive scraping from gptbot (figure 2.9).

Figure 2.9. Training data crawlers, user agent composition



Section 3

AI demand for content

Key Insights

Certain categories of content are subject to markedly higher levels of AI visitors:

- Relative to the level of human traffic, B2B/professional, sport, parenting and consumer technology content receive the highest level of AI scraping, suggesting these are central (publisher-disruptive) AI use cases
- AI requests to parenting content (+333%) and deals and shopping content (+111%) are growing the fastest across TollBit sites. This is likely to indicate new or growing consumer AI use.
- National news content is subject to 5× the number of RAG requests than those from training crawlers; this content is (unsurprisingly) needed in real-time to respond to prompts.

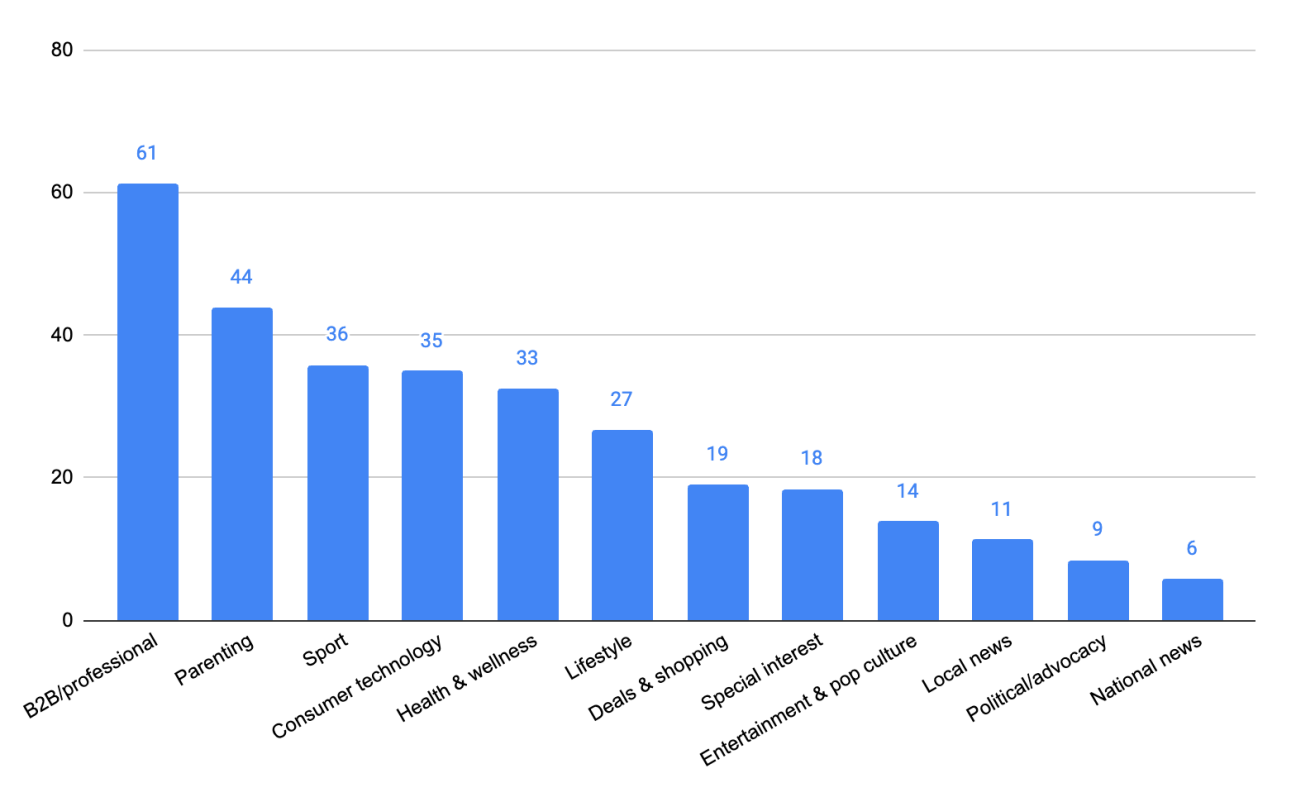
Regional analysis indicates that APAC websites are, by far, the most AI-visited domains, particularly by training data crawlers. By contrast, AI visitor levels to sites in Europe are markedly lower than the global average.

AI scraping levels by site category

AI demand for content is unevenly distributed across TollBit partner websites, with certain categories of content subject to markedly higher levels of scraping. When we examine AI bot traffic relative to the level of human traffic - a measure of the relative AI demand for any given content category - we see B2B/professional content receiving the highest level of AI scraping, followed by parenting, sport, consumer technology, and health and wellness (figure 3.1).

This AI-to-human ratio is likely to be an indication of where AI applications present a substitution risk to publisher content; a higher level (particularly if requests are from RAG agents) suggest that AI tools are being used in place of users visiting a publisher website. For example, parenting content was subject to one AI scrape per 67 human visitors in Q1 2025, which has risen to one AI scrape per 23 human visitors in Q2, potentially indicating that users are directing their parenting queries at AI applications instead of visiting publisher websites.

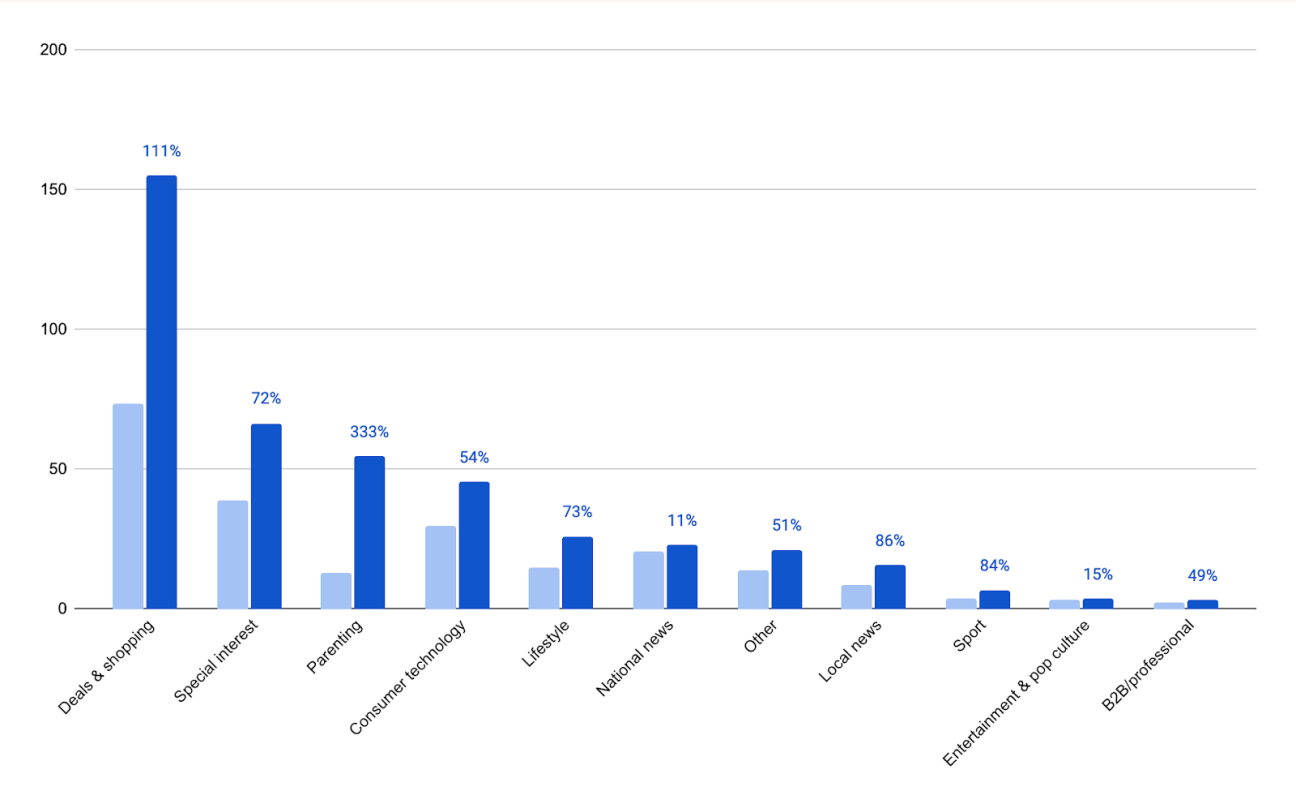
Figure 3.1. AI scrapes for every 1,000 human requests, Q2 2025



Growth in AI scraping by site category

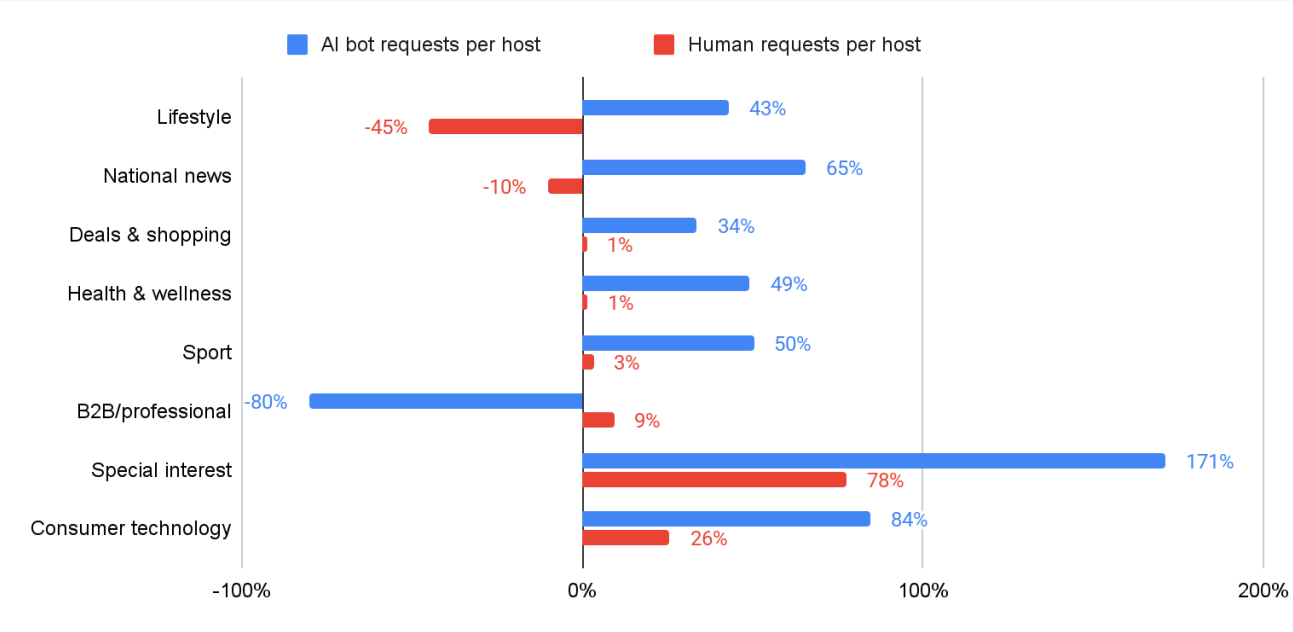
Whilst AI requests are growing across almost all content categories, this rate of growth varies substantially (figure 3.2). When analysing a cohort of websites that joined the TollBit platform in 2024 we can see that AI requests to parenting content, for example, rose by 333%. Deals and shopping is another outlier, with a +111% increase. These are likely to indicate new or growing AI use-cases with consumers turning to AI to satisfy needs that previously were fulfilled elsewhere.

Figure 3.2. Percent increase in average scrapes per page – Q1 to Q2 2025



Further evidence on these potential shifts in consumer behaviour can be found when comparing changes in human traffic and AI traffic for a given content category (figure 3.3). For lifestyle content, human requests declined by 45% whilst those from AI increased by 43%. This suggests that users are turning to AI in place of publisher websites directly. A similar pattern is observed in relation to national news content (AI traffic up 65% and human traffic down 10%). Even where human traffic is growing, this growth rate across many categories is substantially exceeded by the growth in traffic from AI bots (consumer technology, special interest, deals & shopping, sport, health & wellness).

Figure 3.3. Changes in human / AI requests, Q1 vs Q2

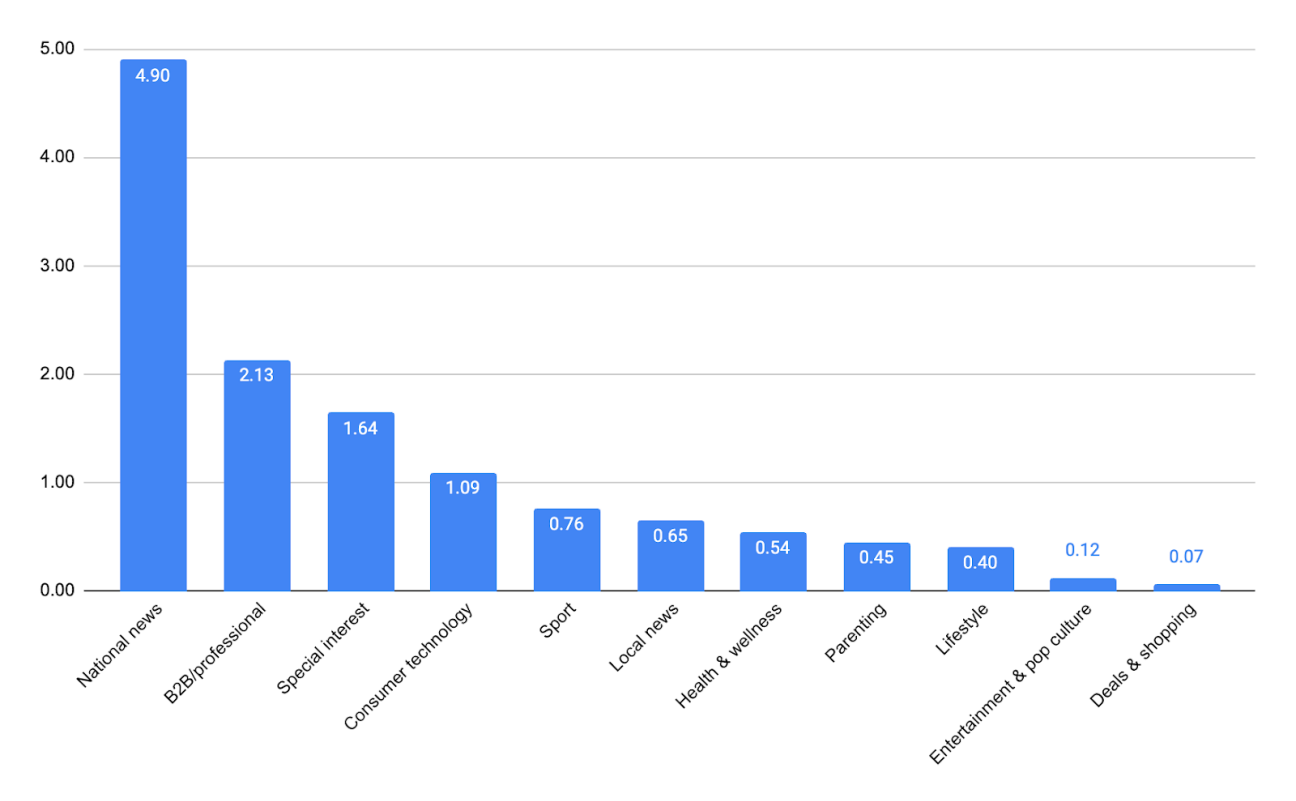


Real-time/RAG versus training

Comparing the level of demand by RAG bots with that from training bots (figure 2.5) provides an indication of the types of content that are fetched in real-time.

Unsurprisingly, given the time-sensitive nature of news content, we see almost 5 RAG scrapes per request from a training data collection bot on national news sites.

Figure 3.4. RAG vs training scrapes, site category, Q2 2025



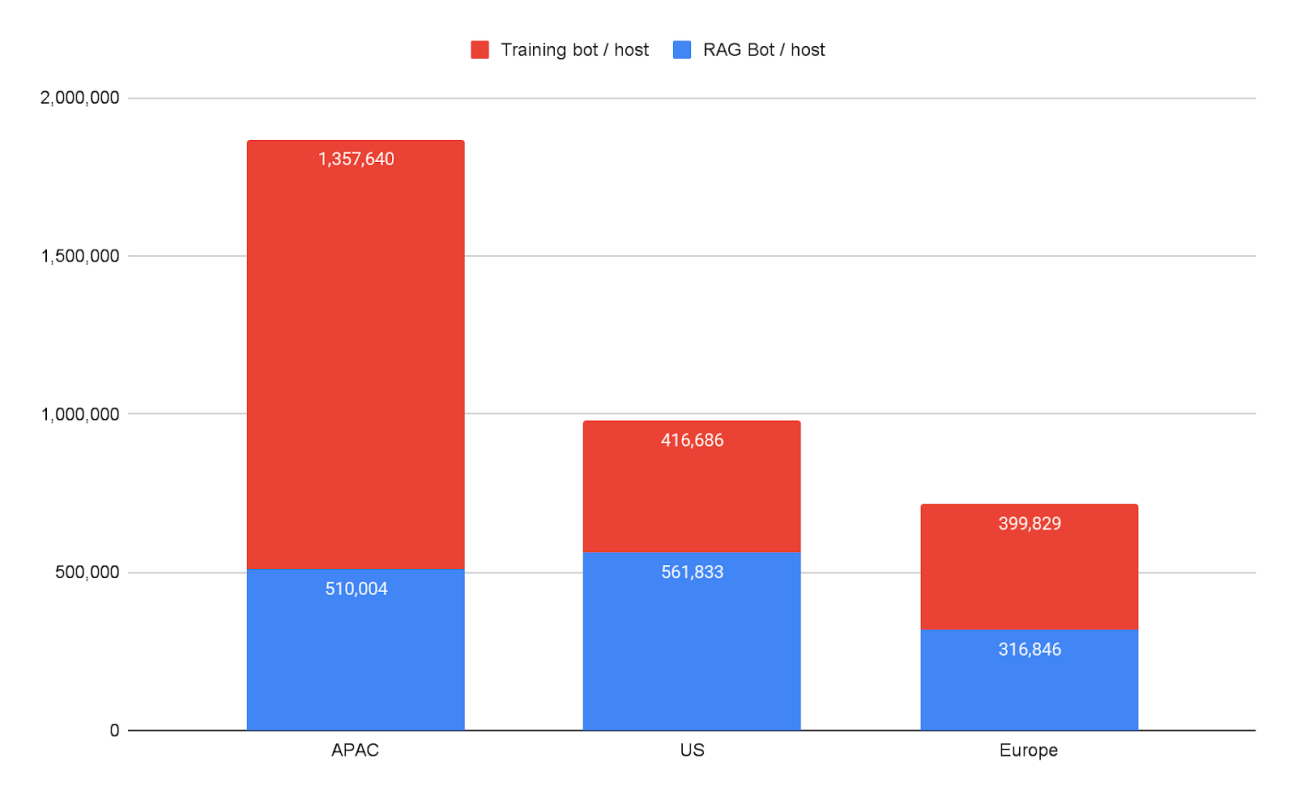
Scraping levels across jurisdictions

TollBit's footprint of media partners is growing globally. As a result, it is now possible to analyze the level of AI scraping across jurisdictions to identify any patterns or discrepancies. Doing so reveals that APAC websites are, by far, the most scraped domains, particularly by training data crawlers. On average APAC sites received over three times the number of requests in the quarter compared to US sites.

One of the drivers for this is likely to be the use of the Robots Exclusion Protocol to disallow access; we find that only 13.5% of APAC sites have disallowed GPTBot (OpenAI's training data crawler), versus 52.5% in the US and 45.8% in Europe (a category which includes the UK).

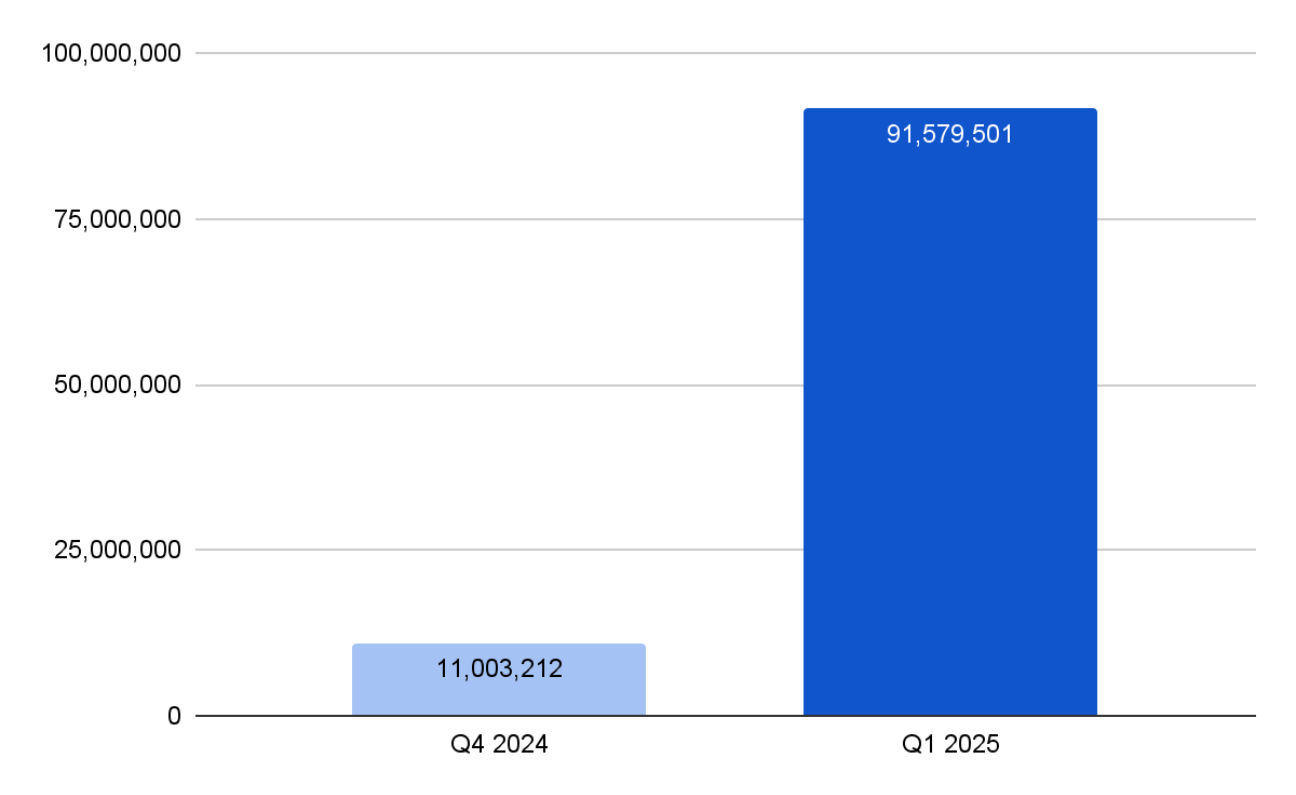
It is notable that scraping levels for European websites are lower, with these receiving 27% fewer AI requests than those for US sites owned.

Figure 3.5. Average scraping levels by website region / jurisdiction, Q2 2025



The number of bots directed to the TollBit Bot Paywall has increased by 732% in Q1 2025 versus Q4 2024. This alternative gateway of sanctioned web access helps to prevent any ads from being inadvertently served to AI visitors, ensures human visitor site metrics don't get impacted by bot traffic, and allows AI bots to be presented with an option to pay for access with added benefits.

Figure 3.6. AI bots directed to TollBit Bot Paywall



Section 4

Referral traffic

Key Insights

Google referrals are now in decline, both in absolute terms and in terms of share. They have dropped from over 90% of all external visitors in Q2 2024 to 84.1% in the same quarter of the following year.

AI applications continue to be a tiny source of publisher referrals - 0.102% of referrals in Q2 2025 came from AI apps. Google still delivers 831 human visitors for every human visitor that comes from an AI system.

The click-through rate from AI applications remains extremely small; over 91% lower than for the average of the top-10 positions on organic search.

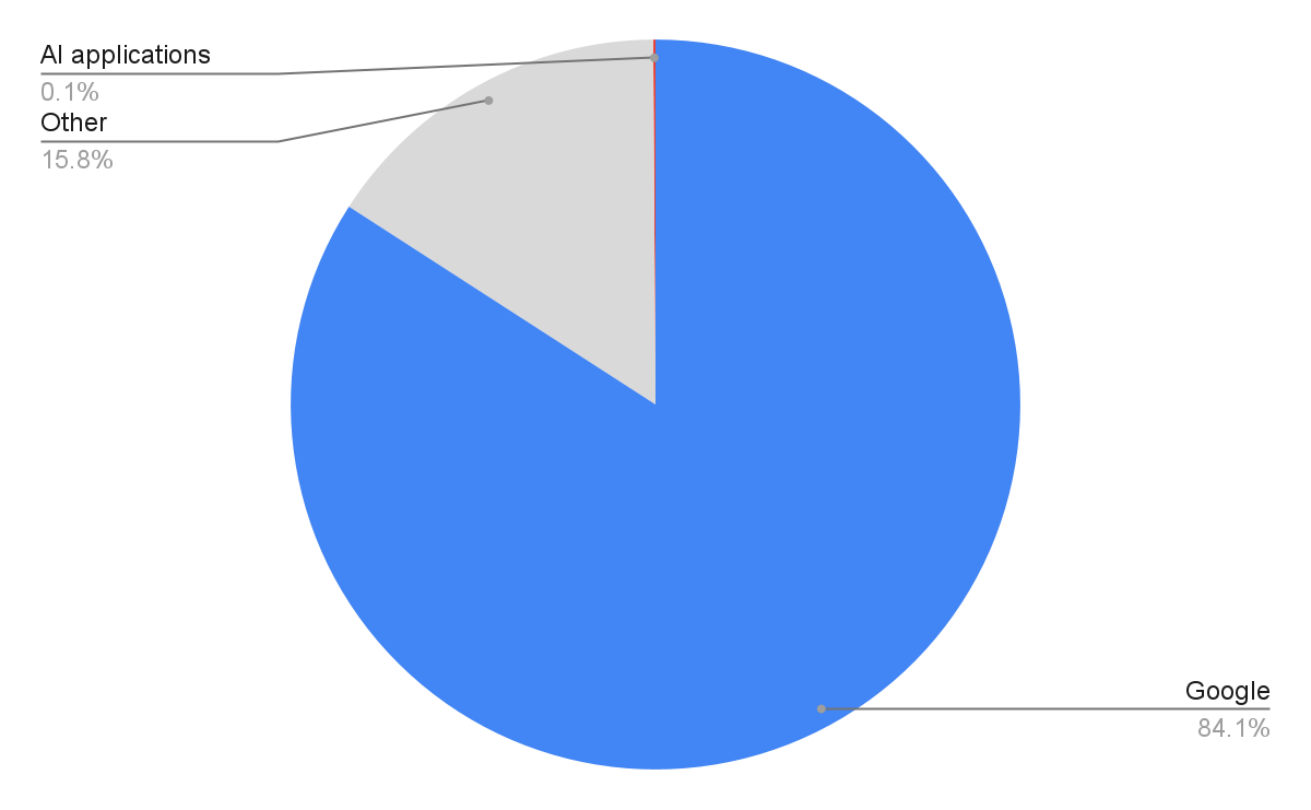
Within this low click-through rate, AI sends proportionately more human visitors to national news and B2B/professional content, and proportionately less to entertainment & pop culture, parenting, sport and lifestyle.

Referral rates from ChatGPT are markedly stronger for publishers with OpenAI licensing agreements in place. Although the difference is closing as more deals are signed.

Aggregate referrals from AI applications

Whilst referrals from AI increased from 0.042% of all referrers in Q1 to 0.102% in Q2, this still represents only a tiny fraction of overall traffic to publisher websites with Google delivering 831 visitors for every single visitor from an AI application (figure 4.1).

Figure 3.6. AI bots directed to TollBit Bot Paywall

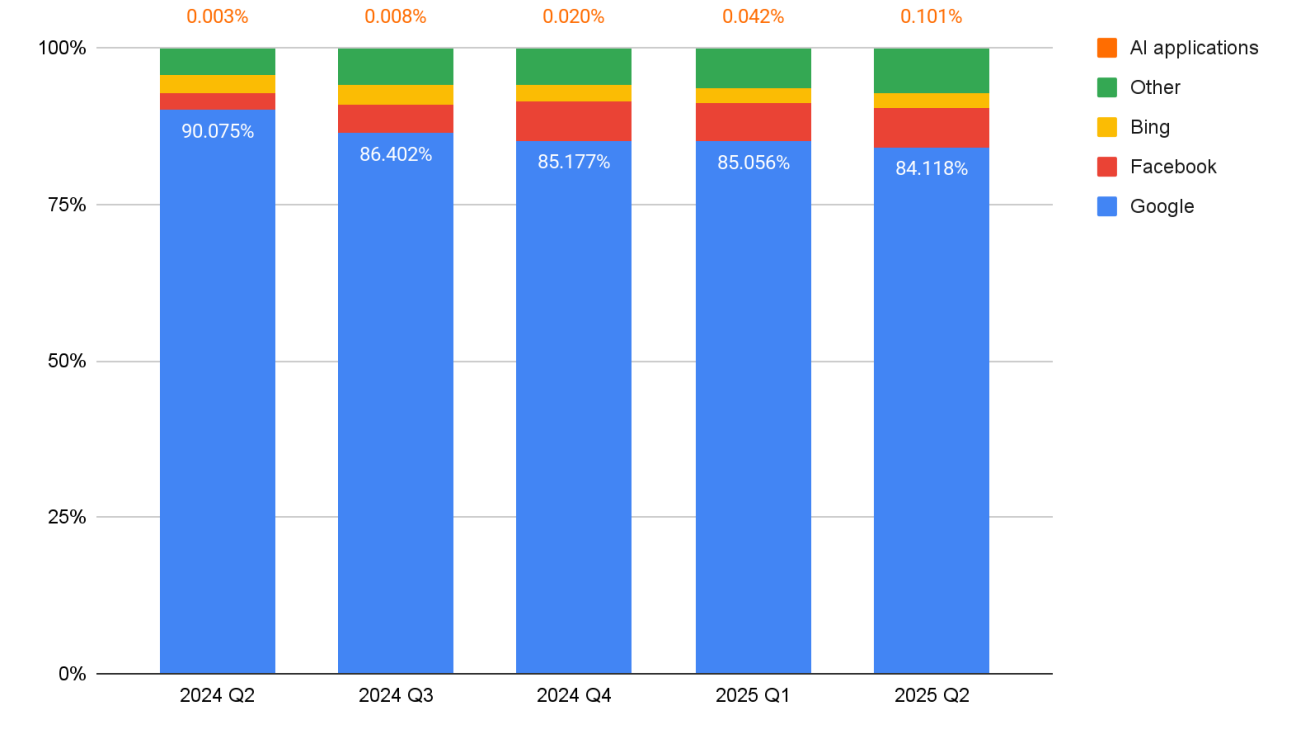


Google remains the dominant source of external traffic; however, TollBit data indicates this is now in decline on both a relative and absolute basis. When we examine referrals from a cohort of sites that joined TollBit prior to July 2024, we find that visits from Google are down 9.17% between July 2024 and June 2025

100 crawls from Googlebot resulted in 454 referrals in Q2 2024, yet one year later, in Q2 2025, 100 crawls only resulted in 312 referrals in Q2 2025. The same number of crawls now yields 142 fewer referrals.

When we analyze Google's share of all external referrals to TollBit partner sites, these have also dropped from over 90% in Q2 2024 to 84.1% in the same quarter of the following year (figure 4.2). This decline in traffic from Google is taking place when crawls from googlebot (its combined user agent) is increasing, and markedly so since the introduction of AI Overviews (see figure 2.6).

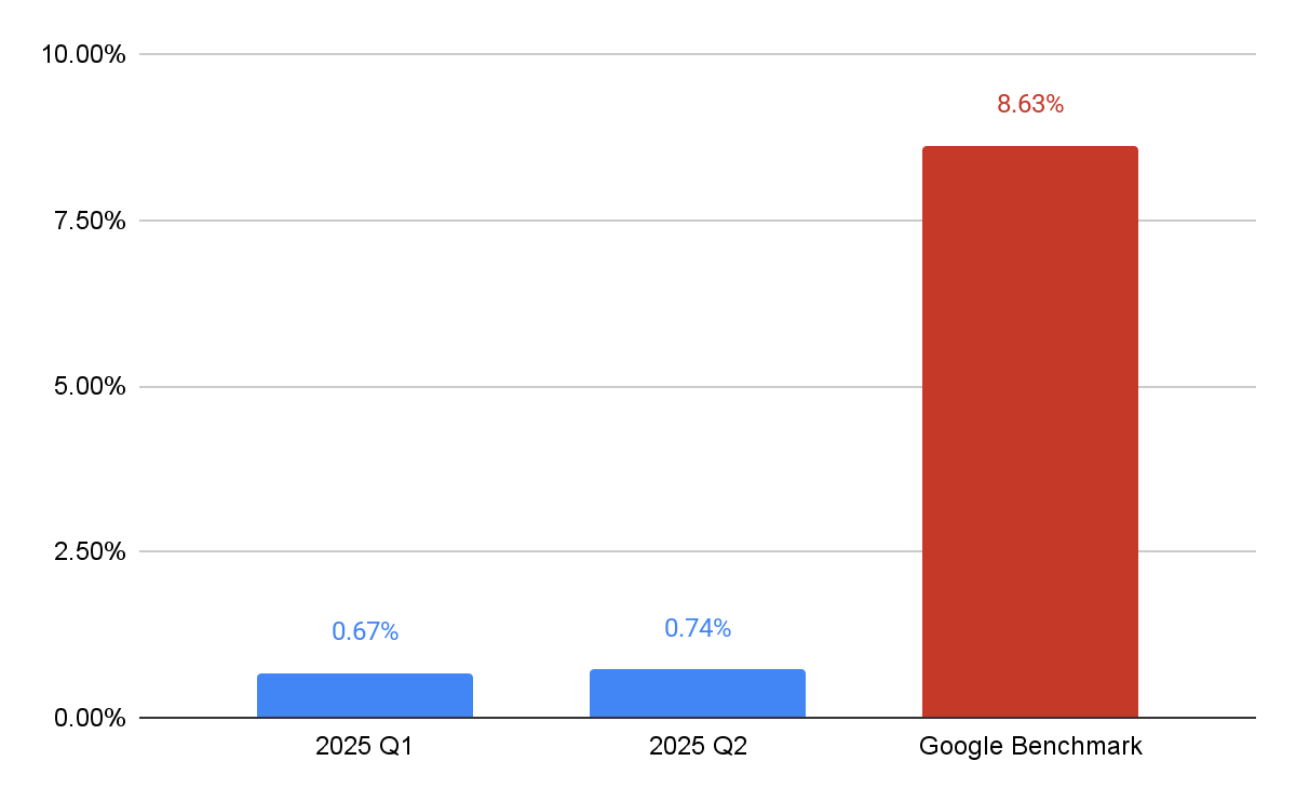
Figure 4.2. Composition of referrals over time



Comparing referral rates

Whilst in aggregate publishers are seeing growth in the number of visitors from AI applications, the click-through rate from these interfaces have not meaningfully improved. Whereas Google's organic (non-AI) search delivers, on average across the top-10 results positions, a click 8.6% of the time, for AI applications, this figure remains below 1%. On average it requires 135 real-time scrapes from an AI application for a single visitor to arrive.

Figure 4.3. AI application click-through rates



Although the total number of human visitors coming from AI sources remains negligible, it's interesting to note that the distribution of these referrals differs from Google (figure 4.4).

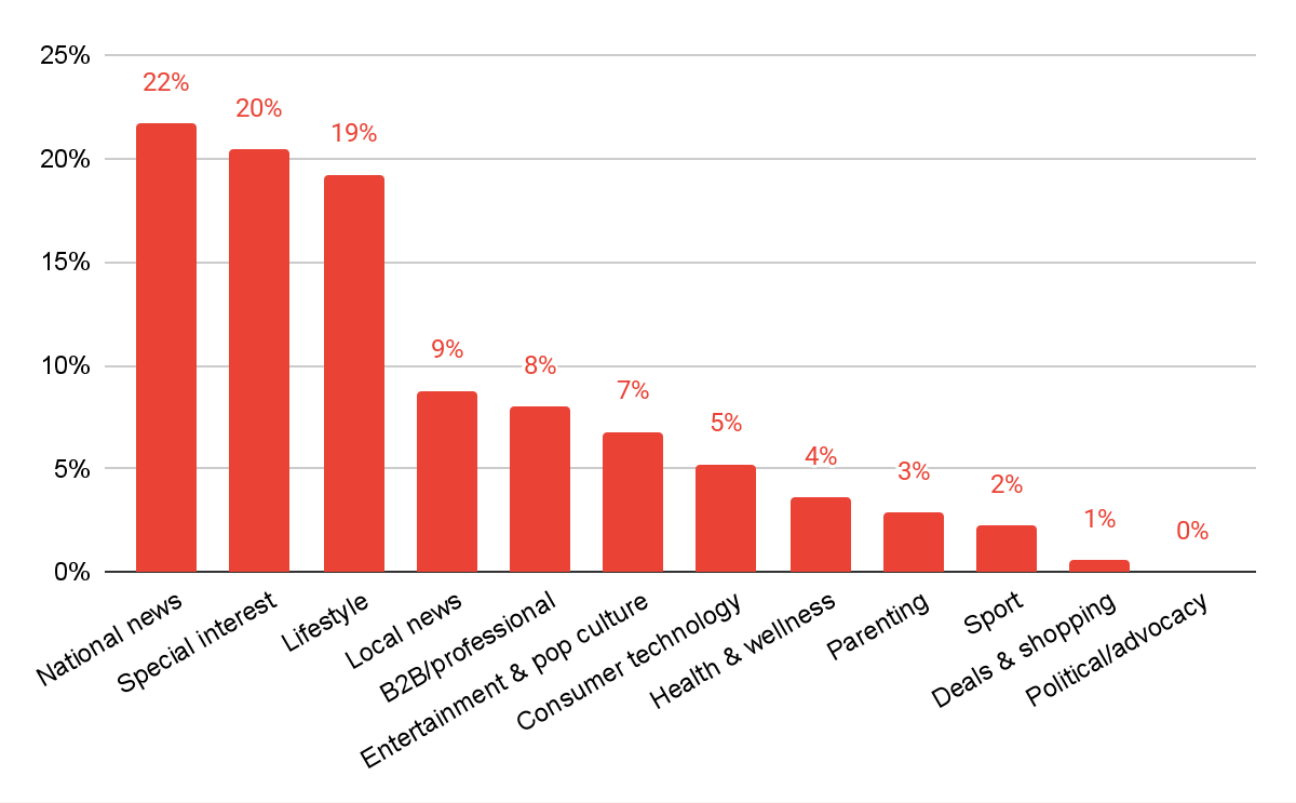
Of the few referrals that AI platforms send, the breakdown of referrals by category is quite different from Google search: For example, 45%of AI apps' total referrals are being sent to National News. While across the same set of sites, only 22%of Google's referrals are being sent to National News.

Two factors may explain this difference:

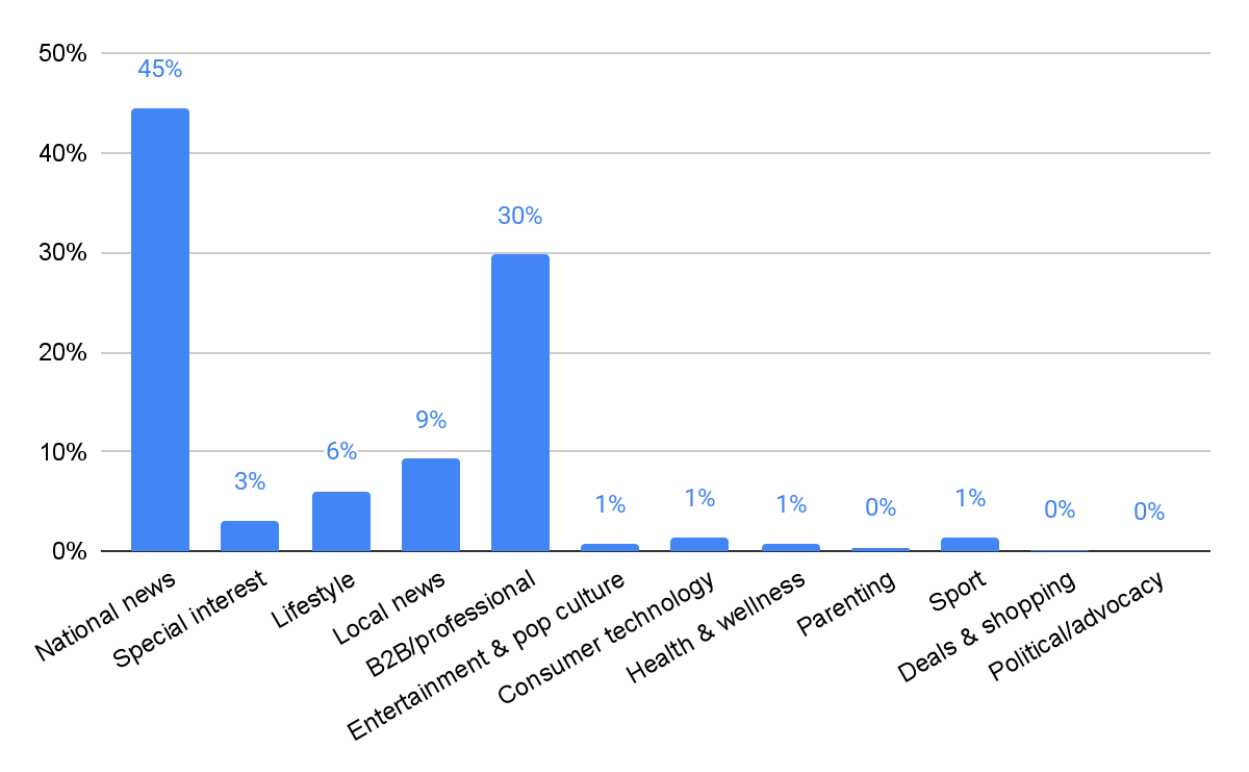
- Different click-through patterns - AI systems may produce a different rate of click-throughs for similar queries versus Google.
- Different starting points for information needs - users may be disproportionately turning to AI tools, rather than Google, for certain categories of queries.

Figure 4.4. Google search vs AI app referral distributions

4.4.1. Google's search referral distributions

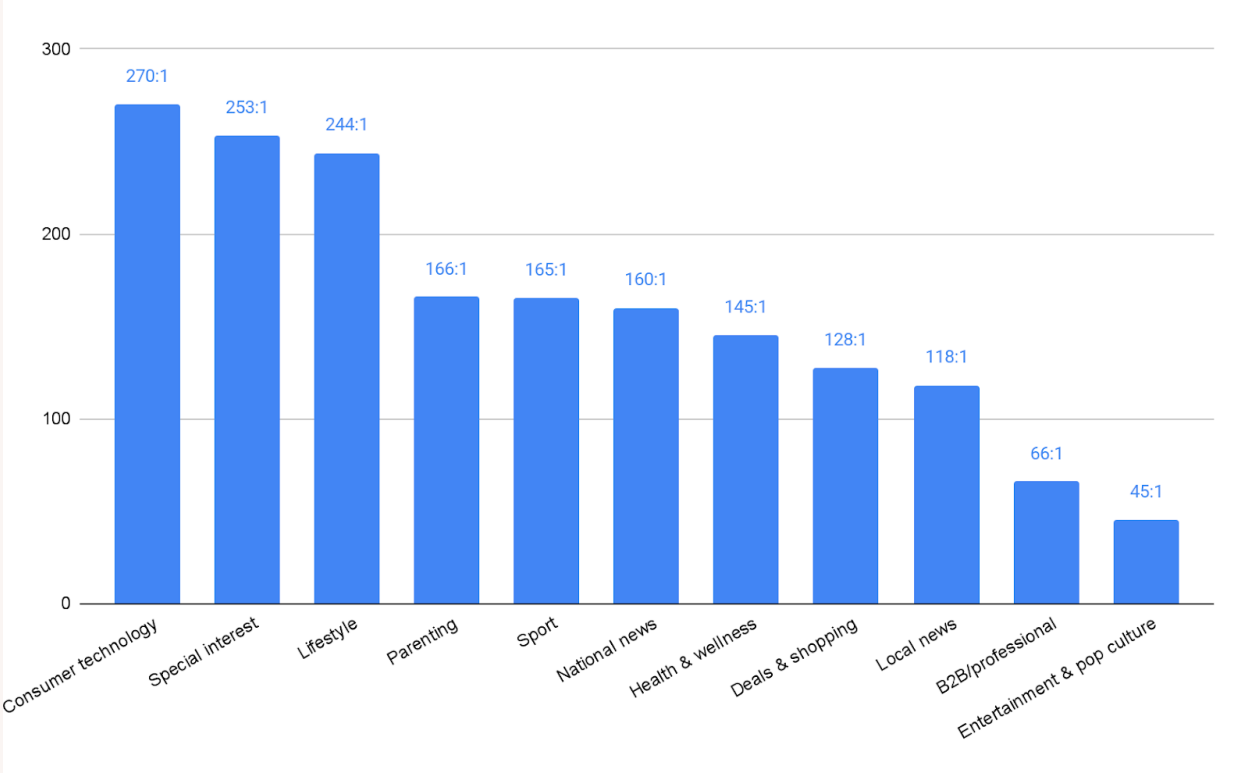


4.4.2. AI applications' referral distribution



By examining the RAG scrape to referral ratio (figure 4.5), we can test whether it is the first of these two factors at play (i.e. whether AI applications are more likely to satisfy specific categories of user need and thus depress the click-through rates for that content). This analysis shows that consumer technology, special interest and lifestyle receive far fewer referrals per scrape (likely indicating that AI answers are likely to be sufficient and further information that would lead to a human click-through is not required). Conversely, B2B / professional and entertainment and pop culture content delivers, in proportion, substantially more clicks than scrapes, indicating the opposite.

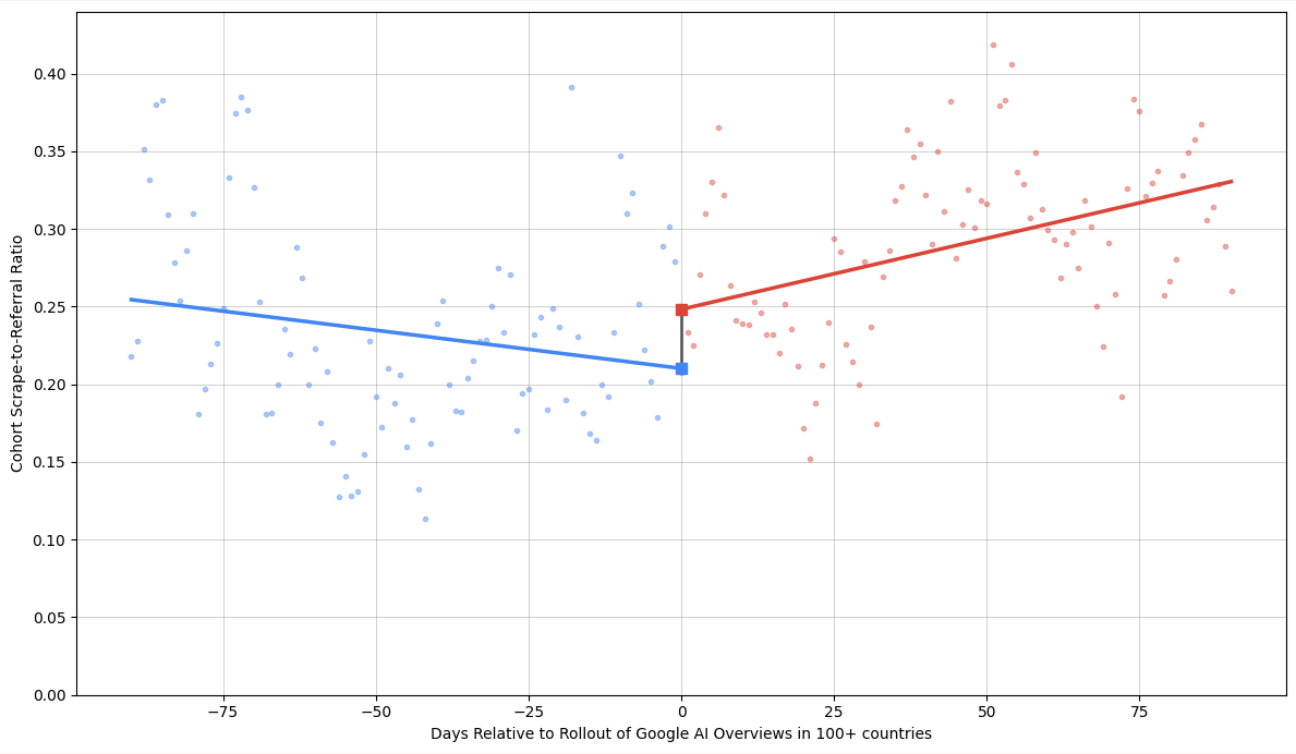
Figure 4.5. Scrape-to-Referral Ratio, Q2 2025



Shifting Google value exchange

As well as corresponding to an increase in the number of requests to publisher websites (see section 2.3) from its crawler, Google's expansion of AI Overviews in October 2024 also resulted in an increase in the crawl-to-referral ratio by 24.45% (figure 4.6). Googlebot is used for a multitude of purposes and Google makes extensive use of offline caches of content so this shift is directional, rather than truly indicative of the number of uses of content that result in traffic delivered. However, it does strongly suggest there is a change to the value exchange, with fewer referrals delivered for each time Google accesses website content.

Figure 4.6. Scrape-to-referral ratio (Googlebot) after global rollout of AI Overviews (0 = 10/28/2025)

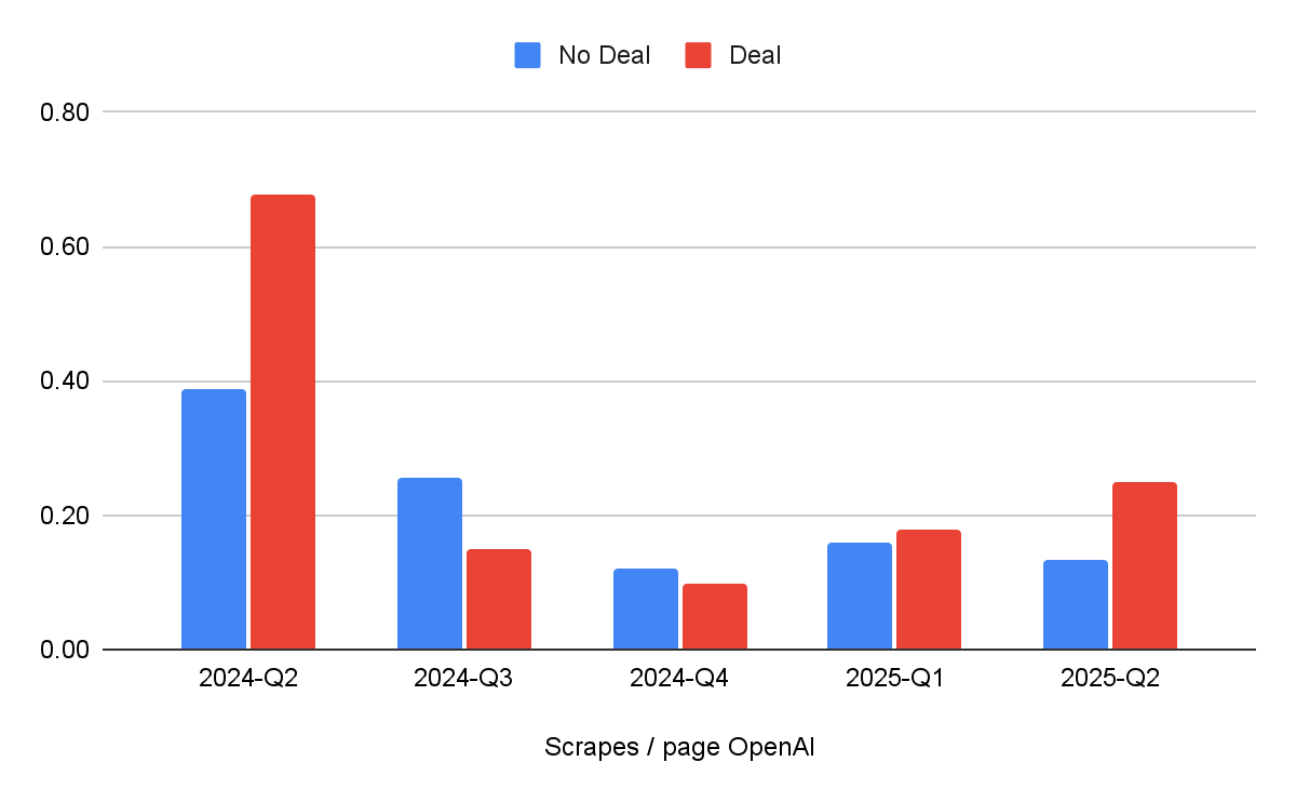


Effect of licensing deals on AI usage and referrals

OpenAI has been establishing new partnerships with publishers since it made its first licensing deal with Associated Press in July 2023. At the time of writing, it has almost 40 similar agreements in place with publishers. While limited public information on these deals is available, we have analysed TollBit data to understand whether having a deal corresponds to more scraping (figure 4.7) or more referrals (figure 4.8) from ChatGPT.

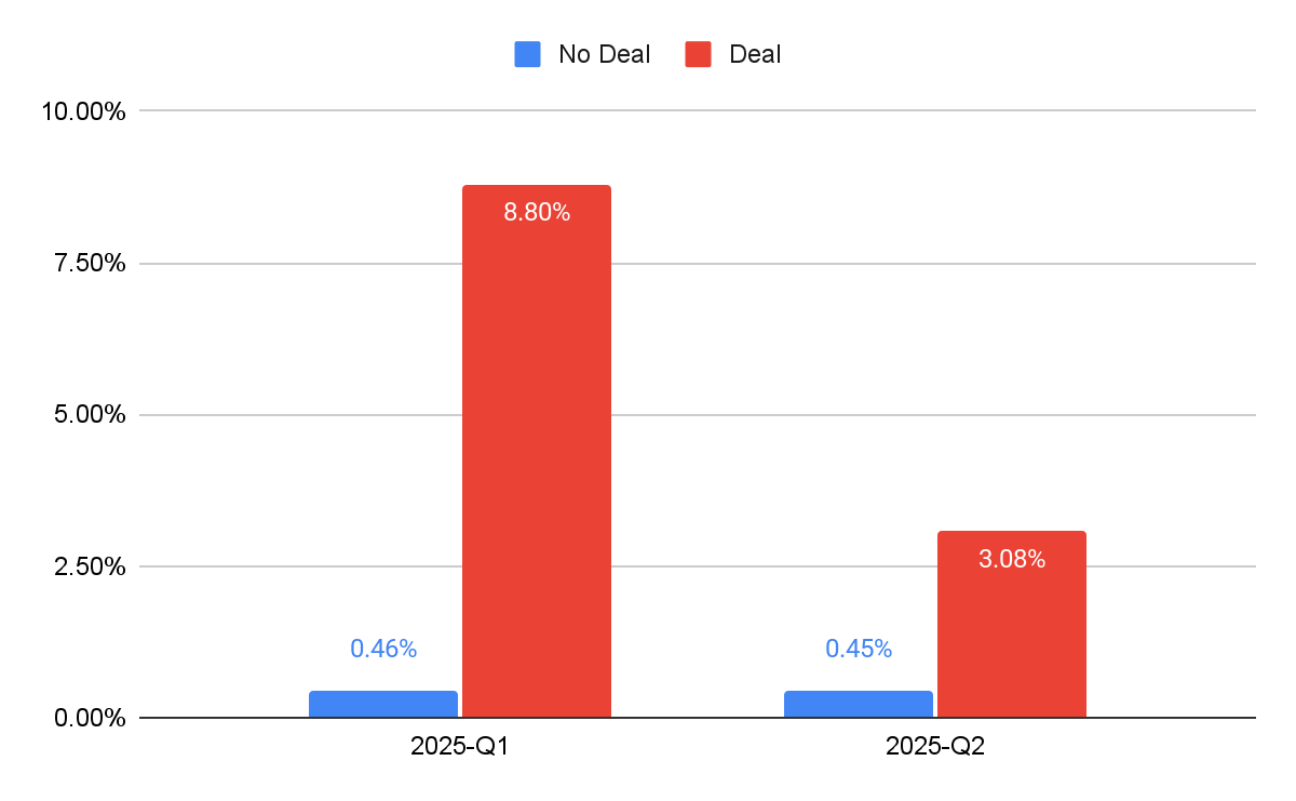
In doing so, we find that in the last quarter, publishers with deals saw 88% more scraping on a per-page basis than those without agreements. This rose from a difference of just 10% in Q1 2025.

Figure 4.7. ChatGPT-User, scrapes per page



Referrals, on the other hand, are markedly stronger for publishers with licensing agreements in place (figure 4.8). In Q1 2025 the click-through rate was comparable to the average of the top-10 organic search positions. AI companies may be placing logos and names of publishers they have done deals with front and center in AI app citation lists. However, the referrals have dropped in Q2 likely as the number of deals expands and the positions with greater prominence are shared amongst a greater number of publishers.

Figure 4.8. OpenAI click-through rates



Section 5

IP controls and errors

Key Insights

We can observe a growing number of forbidden server errors and redirections being served to AI visitors - likely as a result of bot detection solutions being deployed more extensively across publisher sites.

Publishers continue to increase their use of robots.txt to signal that AI bots are not permitted to scrape their content, although the changes this quarter are relatively modest, with most publishers having already implemented directives to the most widely known bots prior to 2025.

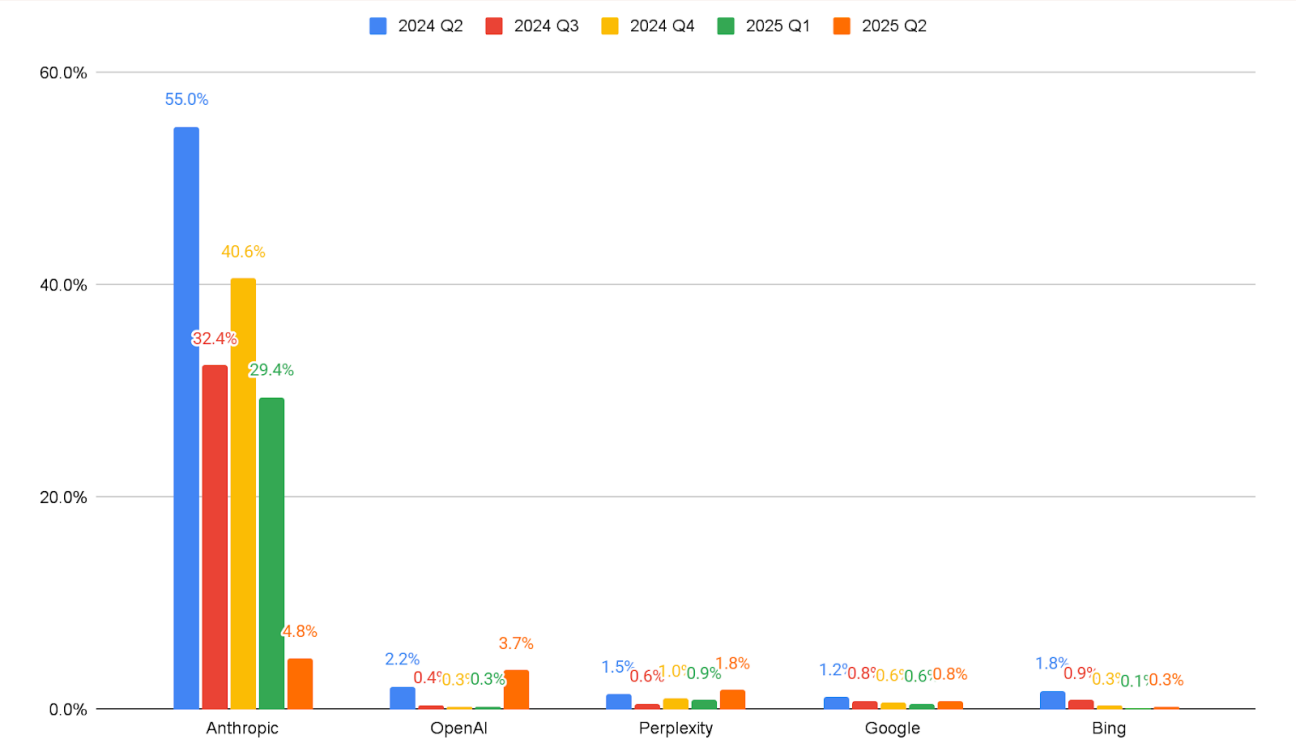
AI crawlers still routinely ignore website robots.txt signals. Crawlers from ByteDance, OpenAI, Meta, Perplexity, and Amazon all bypass publisher signals at times. Across all AI bots, 13.26% of requests bypassed robots.txt in Q2 2025, this has risen from just 3.3% in Q4 2024.

Server responses – HTTP 404 error rates

Each time a request to a website takes place, the site's server responds with a code which indicates the status of the response, for example, whether the request was successful, blocked or was towards a page that doesn't exist. Monitoring the distribution of these codes - and how they change over time - can provide some insight on the nature of the requests coming from AI bots and how website owners are handling them.

404 errors indicate that the page requested could not be found. In the case of search engines, this is often caused by a broken link or the deletion of a page. AI applications additionally may generate 404 errors by hallucinating URLs, directing users to pages that never existed. Two notable changes in 404 error rates from AI applications have occurred over the last quarter (figure 4.2). Firstly, the data indicates a marked increase in OpenAI's figures, rising from 0.3% of responses in Q1 to 3.7% in Q2. Meanwhile, Anthropic's 404 error rate has dropped dramatically from 55% in Q2 2024 (meaning most of the pages requested by users were not found) to 4.8% in Q2 this year. The timing of this improvement corresponds with Claude - Anthropic's chatbot - being given access to the real-time web, allowing it to retrieve and cite live URLs rather than relying on extrapolations from training data.

Figure 5.1. HTTP 404 error rates, 2024 Q2 to 2025 Q2

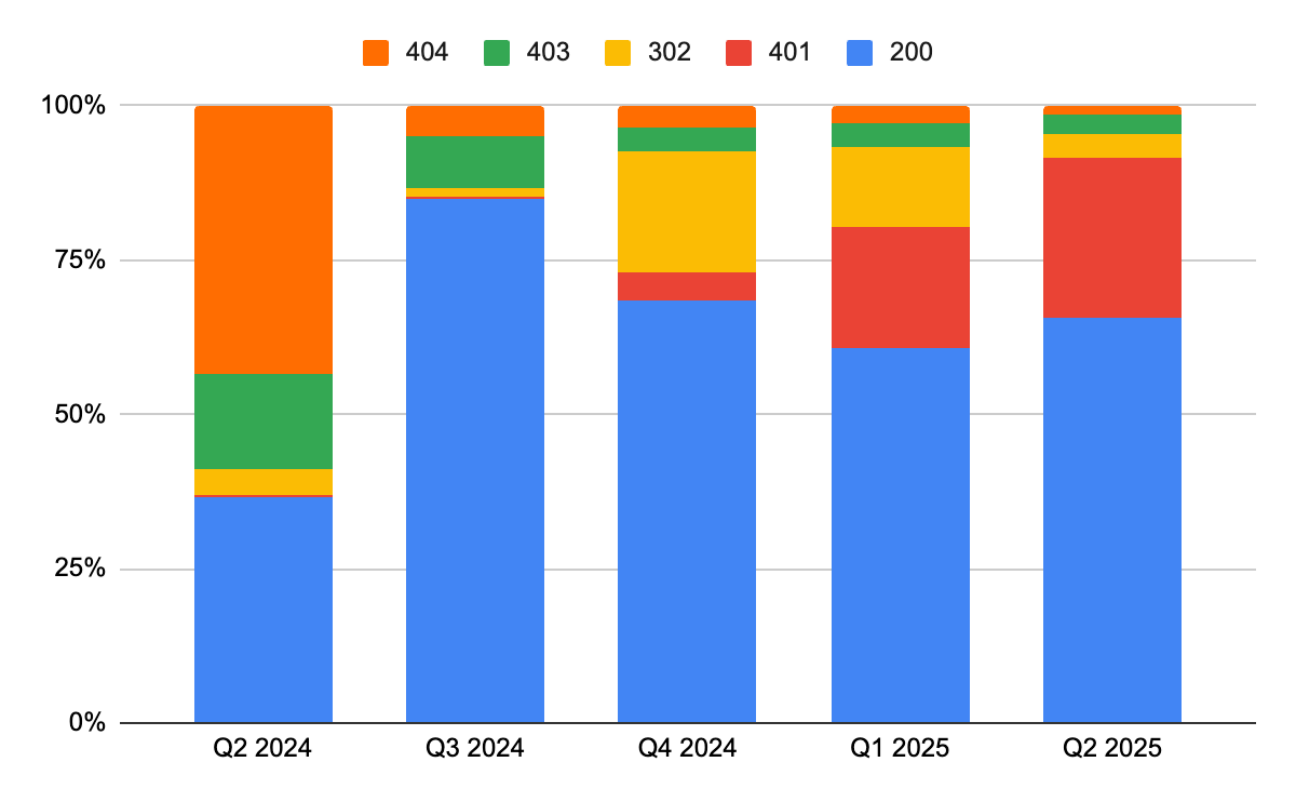


Server responses – Blocking or redirecting error rates

Over the last year, 302, 402, 403, and 401 response statuses[1] have increased by almost 4× on TollBit partner websites (see figure 5.2), likely due to publishers deploying security measures (bot detection/WAF) that prevent non-human visitors such as AI bots from accessing content. TollBit provides basic bot enforcement and partners with various cybersecurity companies, including Fastly, DataDome, and HUMAN Security to provide advanced bot blocking options for sites.

[1] 302 = temporary redirect; 401 = not authenticated; 402 = payment required; 403 = access forbidden/blocked

Figure 5.2. All domain quarterly changes in 403 errors



Publisher use of disallow signals

Publishers continue to extend their use of the Robots Exclusion Protocol (AKA 'robots.txt') to signal that AI bots are not permitted to scrape their content. While most publishers implemented changes on the most widely known bots prior to 2025, we can observe a growth in disallow requests as new AI scrapers are released and become known (figure 5.3 and 5.4). For example: As of June 2025, 81% of the cohort of sites are blocking CCbot and 73% block GPTBot.

Figure 5.3. Total AI bot disallow requests

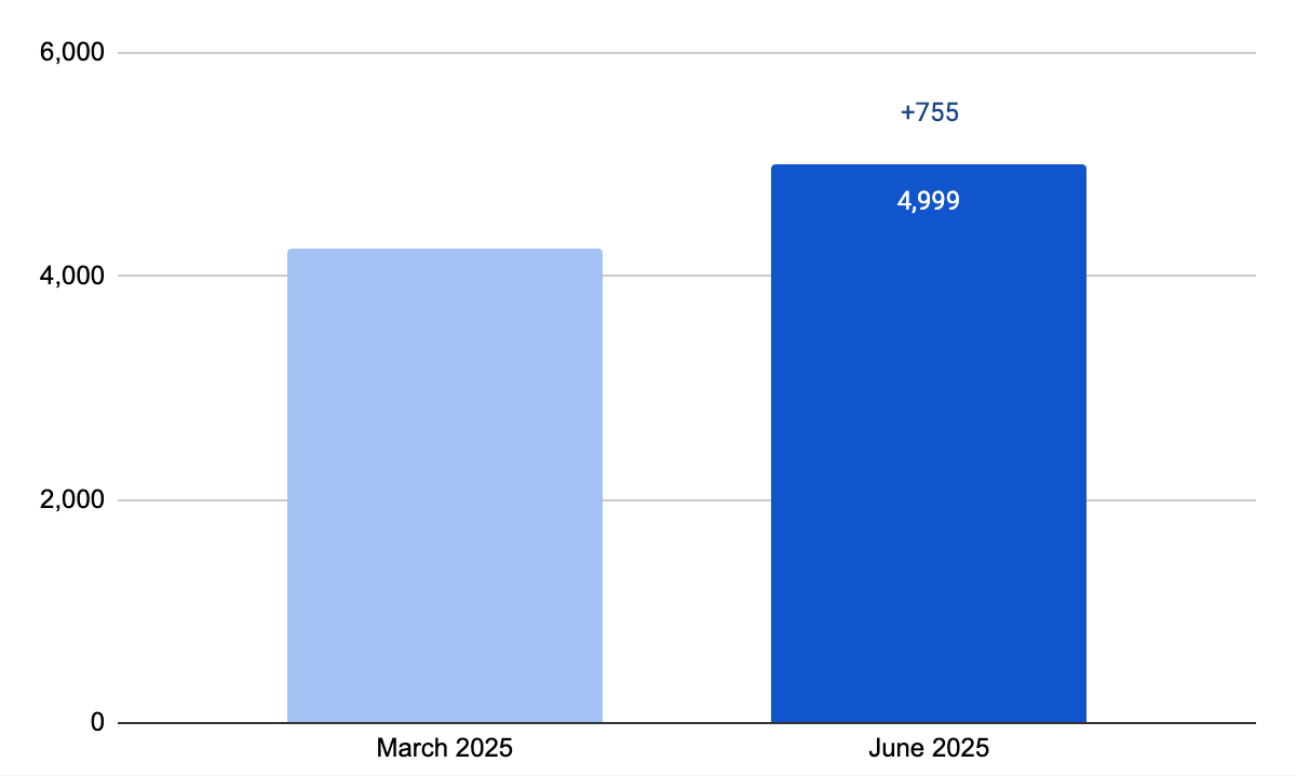
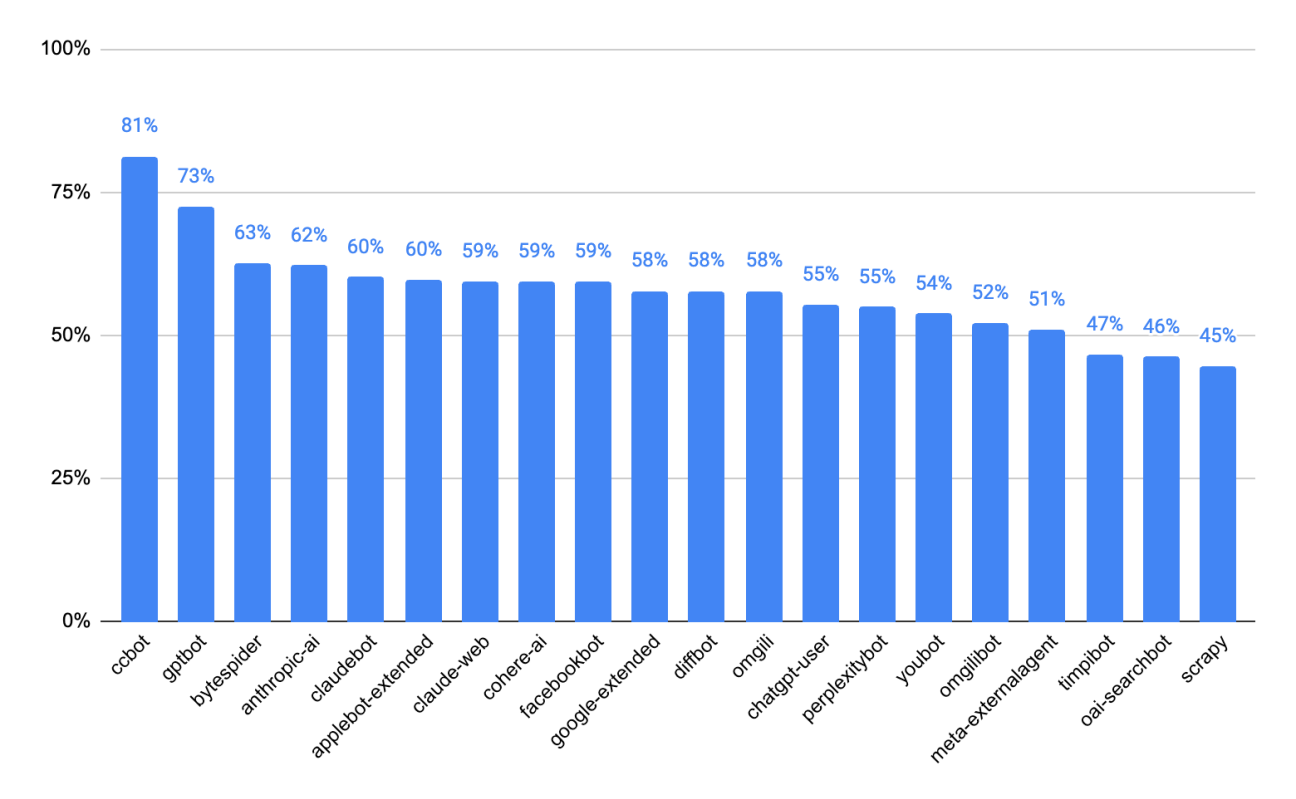


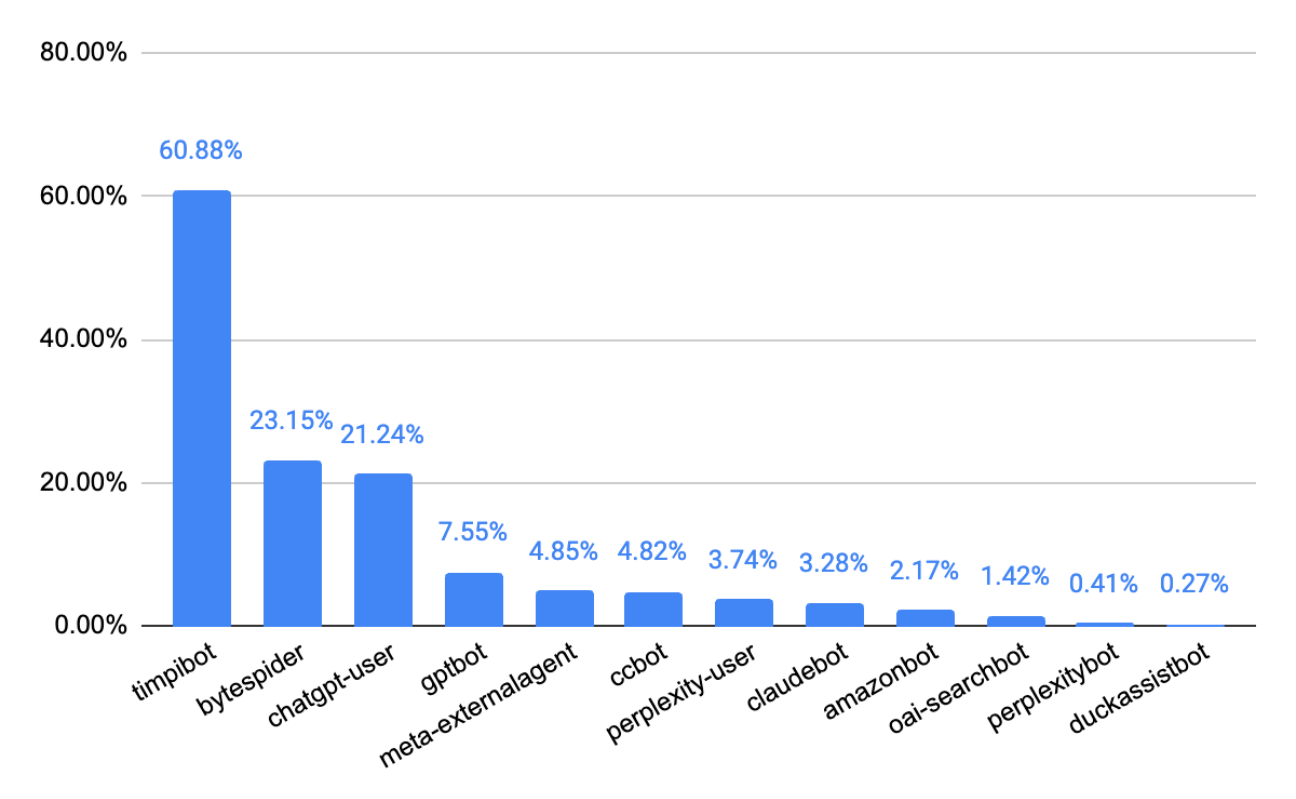
Figure 5.4. Disallowed Bots via Robots.txt, cohort of TollBit partner websites



Robots.txt compliance

TollBit data shows that many AI crawlers still appear to ignore these publisher robots.txt signals, scraping content even when explicitly requested not to. This behavior varies significantly by bot, with crawlers from ByteDance, OpenAI, Meta, Perplexity and Amazon all bypassing publisher signals (figure 5.5) at times. Across all AI bots, 13.26% of requests bypassed robots.txt in Q2 2025, this has risen from just 3.3% in Q4 2024.

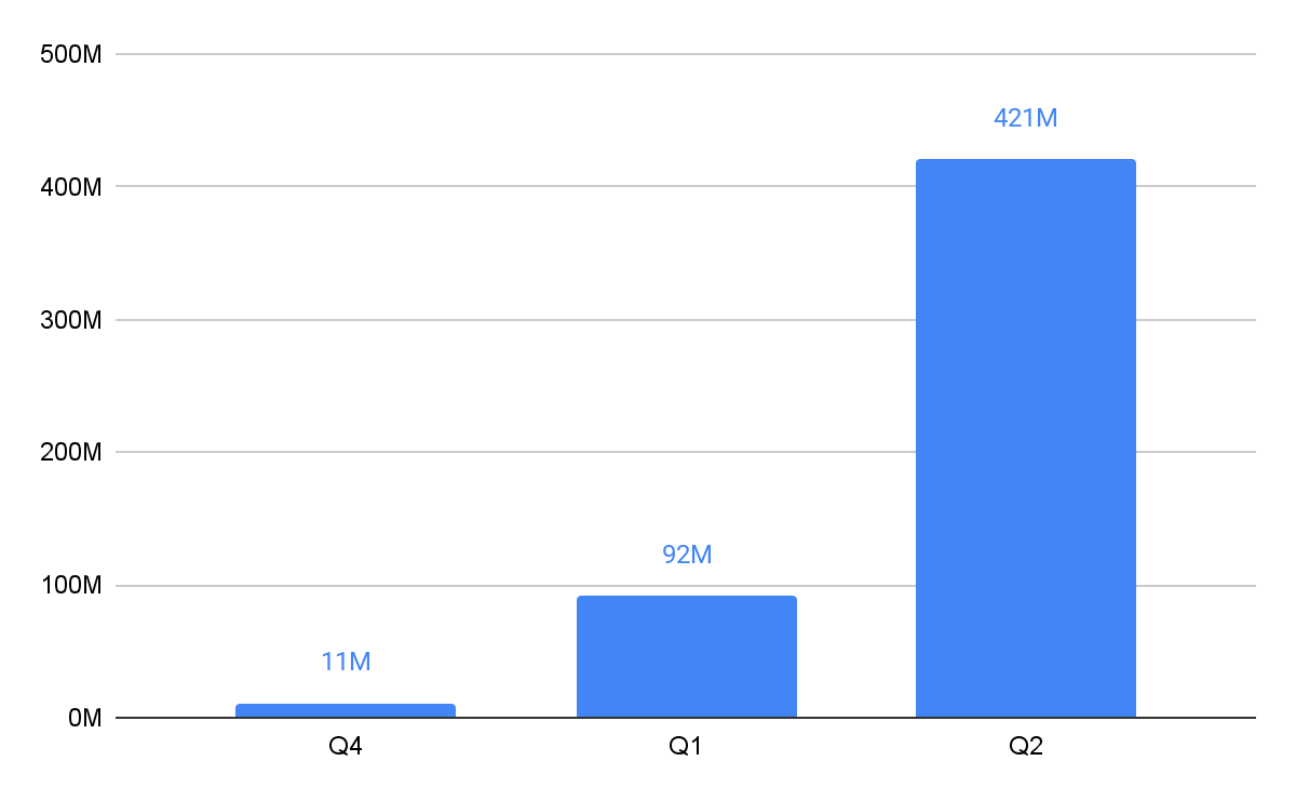
Figure 5.5. Requests that bypass robots.txt signals, percent Q2 2025



Redirects to Bot Paywall

The TollBit platform works with websites' existing tech stacks to redirect bots which are not permitted to access a given site or page to a Bot Paywall with the price for access set by the publisher. Hits to these paywalls have been growing steeply (360% from Q1 to Q2 2025) as publisher partners deploy this functionality (figure 4.6). This delivers an HTTP 402 - Payment Required to the AI bot or agent.

Figure 5.6. Hits to TollBit Bot Paywall



Section 6

Future trajectory of AI bot traffic

Key Insights

There is evidence that AI apps store & re-use content, but it appears it's not always optimal for them to cache data for long periods. This highlights the importance of RAG and continued content access. Here's a breakdown of what that caching pattern looks like based on our tests:

- ChatGPT: cached for 30 minutes across user accounts
- Gemini: cached for 15 minutes at a user account level
- Claude: cached for 16+ days* across user account

*Our testing ended after this time period

AI demand for professional content is only set to grow as this technology is deployed in new applications and consumer adoption continues to accelerate. However, identifying this demand by observing the activity on the publisher servers will become harder as AI developers increasingly store and reuse content by making use of offline caches of the content. The TollBit team has been running experiments to understand - and explain - each of the top AI bots usage of caching for our State of the Bots report readers.

How We Set Up the Experiments*

To understand how AI applications retrieve web content —and whether they rely on live scraping or cached (aka stored) data—we ran a series of structured experiments across four AI applications: ChatGPT, Gemini, Claude, and Perplexity.

The Test Website

We created one-page websites that showed a jumbled sequence of letters (example site here: https://getcode-gem.onrender.com/). This "code" changes every second and on every load or refresh. Since the site shows a unique code that changes every second, the code returned by the AI application tells us when it accessed the page. If an AI model responded with a code that matched the exact time a prompt was sent, it indicated real-time scraping. If it returned a code from an earlier time period, it suggested the application had cached/stored the content from a previous request.

- The test website did not block any bots in the robots.txt file.
- The test website wasn't submitted to Google Search Console, so it couldn't be pre-cached.

Repeating this process across hundreds of queries and cross-referencing with server logs let us map cache refresh intervals, identify cross-user caching, and observe whether models behaved consistently across different prompt styles.

We used the following setup for each AI application/ chatbot:

- Ran between 200 to 500 prompts
- Queried each application using two separate user accounts to test for cross-user caching.
- Prompt style was designed to trigger RAG behavior. Example: "Find the latest code displayed on [URL]."
- Prompts were spaced at varying time intervals—from a few seconds to a few hours—and made from different user accounts to test both time-based and user-based cache behavior.
- We cross-referenced each prompt with server logs to verify whether and when the page was accessed by the chatbot's respective AI bots.

OpenAI ChatGPT

What we tested

- We prompted ChatGPT (GPT-5 and o4-mini-high) over 500 times using multiple OpenAI accounts.

What we tested

- Cache duration: The cache duration was approximately 30 minutes. During 30-minute intervals, ChatGPT returned the same cached code regardless of prompt time or user account.
- Cross-user caching: ChatGPT consistently cached content across users, returning the same code even when queried minutes apart from different accounts.
- Scraping time: Real-time scraping occurred only after cache expiry. Once the cache expired, the respective AI bot(s) fetched a fresh version of the code at the time of the prompt.
- Citation: The webpage was cited as a source in the response. Even when the content was cached and not freshly scraped, ChatGPT displayed the website in the citations provided to the user.

What showed up in our logs

- User agent: Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko); compatible; ChatGPT-User/1.0; +https://openai.com/bot
- IP Address: IP addresses used were part of OpenAI's public list and registered to Microsoft Limited

- Additional scraping activity: We observed OpenAI's SearchBot (OAI-searchbot) accessing /robots.txt multiple times, outside active test windows. This suggests background scanning unrelated to prompt timing.

Google Gemini

What we tested

- We prompted Gemini 2.5 Pro over 200 times using three different Google accounts.

What we observed

- Cache duration: Cache duration was approximately 15 minutes. During 15–minute intervals, repeat prompts returned the same code — the page was not scraped again.
- User-level caching: Caching occurred at the individual user level. Each Google account saw its own cached result. The cache was not shared across users.
- Scraping time: Real-time scraping occurred only after cache expiry. Once the cache expired, the respective AI bot(s) fetched a fresh version of the code at the time of the prompt.
- Citation: Gemini included the test page as a citation, even when the response was served from cache.
- Other model behavior: Gemini 2.5 Flash cache duration did not differ from Gemini 2.5 Pro

What showed up in our logs

- User agent: Google (Note: This user agent string is not listed in Google's official documentation but was consistently used during Gemini's scraping activity.)
- IP Address: Although this IP address is not listed in Google's published IP ranges, a reverse DNS lookup confirmed that it was associated with Google's IP addresses, as per Google's verification guidelines.

Anthropic Claude

What we tested

- We prompted Claude (Sonnet 4 and Opus 4.1) over 200 times using multiple Anthropic accounts.

What we observed

- Cache duration: Even after 16 days of testing and hundreds of queries, Claude continued to return the same stale cached content.
- Cross-user caching: Claude cached responses across users and across time.
- Scraping time: The model never scraped the test page again after the initial prompt, regardless of changes in time gaps between prompts.
- Citation: Claude included the test page as a citation, even when serving cached responses.

- Other model behavior: Both Claude Sonnet 4 and Claude Opus 4.1 showed identical behavior throughout testing.

What showed up in our logs

- User agent: Mozilla/5.0 AppleWebKit/537.36 (KHTML, like Gecko; compatible; Claude-User/1.0; Claude-User@anthropic.com)
- IP Address: Anthropic doesn't publish IP addresses it uses. The IP addresses were registered to Google Cloud.
- Additional scraping activity: Claude-User checked the/robots.txt page before scraping the website.

*Note: These AI tools are evolving rapidly. The caching behaviors and bot activity we observed may change over time, and results may differ based on prompt wording, account setup, or system updates.

APPENDIX

AI USER AGENT PROFILES

1. Categories of AI user agent

As our State of the Bots reports chart, the internet is crawled by an ever-growing number of AI user agents. These have different functions for the applications they serve and consequently, behave differently online. Here is a quick overview of the principal forms of AI user agent, what they do, how they work and why they matter to publishers.

1.1 Retrieval augmented generation (RAG) agent

These bots retrieve information in real-time to respond to user prompts. They use an index of the web, gathered by an indexing crawler (either proprietary or third-party such as Bing or Google) to locate the relevant content which is then retrieved and synthesized into a response.

Publisher effects

When a RAG agent accesses a site, it's using that content to provide a response to a user's prompt, typically in an AI chatbot (e.g. Chat-GPT) or AI search (e.g. Perplexity) application. This might be substitutional to a human visitor.

1.2 Training data crawling agent

Large language models - such a Llama from Meta or GPT-5 from OpenAI, both of which power a multitude of consumer applications - are trained on vast quantities of data. Training data crawlers move around the web - following links from websites to websites or working through sitemaps - downloading content which is then processed and stored for offline use.

Publisher effects

The inclusion of content in the training data means that a model has this 'knowledge' to answer prompts. Training data typically cuts off several months prior to model release so the substitutional risk is limited for content with a longer shelf-life.

1.3 AI search indexing agent

AI systems with access to the real-time web need an index of the internet. This is used to direct RAG agents to the right sources when collecting data needed for responses to prompts. AI search indexing crawlers build these indexes by systematically navigating the web, collecting and organizing content and metadata.

Publisher effects

These crawlers only ensure that a website's content can be navigated to by a RAG agent, should a relevant prompt require that.

1.4 Hybrid agent

Some AI developers use a single bot (or at least software which identifies itself with the same user agent metadata) for more than one of these purposes. For example, PerplexityBot appears to act both as an indexing crawler and a RAG agent.

Publisher effects

This makes it hard for publishers to apply granular controls to the access and use of their IP by AI applications.

2. AI bot profiles by operating organisation

2.1 OpenAI

Developer of ChatGPT, OpenAI's chatbot was first-to-market and still holds a ~60% share[1] with 700M weekly active users[2] as of August 2025. Its foundation models - and therefore user agents - power both ChatGPT and an array of third-party applications, including Microsoft's Co-Pilot and Bing search engine.

User agents

ChatGPT-User

ChatGPT-User accesses websites in real-time and on-demand so that ChatGPT can formulate responses to

user prompts based on the live web. It visits websites to gather information then processes, summarises and synthesizes this to provide the output. It is not used to gather data for model training.

OAI-SearchBot

OAI-SearchBot is used to power ChatGPT's search capabilities. Similar to ChatGPT-User, it accesses the internet in real-time but is optimized for search scenarios, delivering the raw links and search results alongside summaries.

GPTBot

GPTBot gathers data from the web for the training of OpenAI's large language models. It operates continuously in the background and the data it gathers is collected and used offline for model development, rather than real-time responses to user prompts.

Robots.txt policy at time of publication

OpenAI respects the signals provided by content owners via robots.txt (see boxed text), allowing them to disallow any or all of its crawlers.

Publisher partnerships

OpenAI has an extensive publisher partnerships program, having signed bilateral deals with ~40 publishers at the time of writing[3]. It is understood that these deals include both a real-time access component and data for model training.

Robots Exclusion Protocol or ‘robots.txt’

This index also refers to the Robots Exclusion Protocol or ‘robots.txt’. This mechanism allows website owners to give instructions to bots about accessing a website. It uses a machine-readable file (named robots.txt) which specifies - for individual bots or all bots collectively - which pages or sections they can or cannot crawl. Robots.txt operates simply as a signal though and does not actively block access. Not all developers program their bots to comply with these instructions.

2.2 Perplexity

Perplexity has developed an AI answer engine, effectively AI-powered search that provides users with a natural language response to a prompt alongside a list of links and sources. It has 22 million active users[4]. The standard free product primarily uses a proprietary model, whereas the premium, paid-for service includes access to a range of models including OpenAI’s GPT-4 Omni, Claude 3.5 from Anthropic, Llama 3 from Meta and Grok-2 from xAI. In all instances Perplexity retrieves information in real time from the web via its own user agent. It has also been reported that Perplexity makes use of unofficial user agents[5]. These are discussed at the end of this section.

User agents

PerplexityBot

This bot gathers data from across the web to index it for

Perplexity's search function. It has also appeared to act in real-time, gathering data to respond to specific user queries as they are placed, however this function is now - at least in some circumstances - performed by Perplexity-User.

Perplexity-User

This user agent accesses web pages in real-time to respond to user prompts.

Robots.txt policy at time of publication

Perplexity claims that its PerplexityBot user agent respects robots.txt. However, there have been widely-reported complaints from publishers that it has ignored their signals, leading to an investigation from Amazon, its cloud provider. In its FAQs, Perplexity explains that 'if a page is blocked, we may still index the domain, headline, and a brief factual summary[6]. It is not clear that it is possible to exclude a page or domain from this activity. Its policy makes no mention of the behaviour of its second user agent Perplexity-User.

Publisher partnerships

In 2024 Perplexity launched a publisher partner program under which it provides a share of advertising revenues generated by responses that are based upon the content of media partners. Around 20 deals have been signed so far, mostly (although not exclusively) with smaller or niche publishers. The revenue shared with publishers is reported to be capped at 25%[7].

2.3 Anthropic

Founded by seven former OpenAI employees, Anthropic is an AI developer with a focus on privacy, safety and alignment (with human values). Its foundation models power its Claude chatbot - which has a free and premium version - and a multitude of third-party applications.

User agents

ClaudeBot
Anthropic uses ClaudeBot to gather data from the internet for AI training.

Claude-SearchBot
This user agent indexes online content to power Claude's web search. This allows Claude-User (see below) retrieve up-to-date information when responding to user queries.

Claude-User
This bot fetches a web page on demand when a user asks a question that requires a live lookup.

Robots.txt policy at time of publication
Anthropic's bots respect publisher signals in robots.txt files. Notably they also respond to any disallows for Common Crawl's CCBot, which gathers web data in an open repository that is widely used by AI developers.

2.4 Google

Google is one of the foremost AI developers with its own proprietary models that have been integrated extensively into its consumer and enterprise applications, including search. It also operates Gemini as a standalone AI chatbot with real-time web access. This has around 14% of the chatbot market[8] with an estimated 42 million active users[9].

User agents

GoogleBot

This is Google's all-purpose search bot. It creates the index of the web that the search engine relies and evidence suggests it is now also operating in real-time to gather information from websites for both AI Overviews and AI Mode.

Google-Extended

This bot is used to gather data to train and improve Google's AI models. It operates independently of the crawlers used to power Google's search product. It should be noted that both Gemini, when it requires data from the live web, and AI Overviews (the natural language AI response to search queries) do not rely on Google-Extended for real-time data retrieval and therefore disallowing this bot does not control whether a publisher's content is used to inform the outputs of these products.

Robots.txt policy at time of publication

Google's bots respect robots.txt signals. However,

publishers do not have granular controls over the use of their content in real-time for AI search or Gemini's as these products use the data collected for Google's general search product. In order to prevent content being used for these applications publishers need to use the no-snippet directive or signal Google to stop indexing a page for search entirely. Both of these would have negative effects on prominence and referral traffic.

Publisher partnerships

Whilst Google has signed a $60M AI content licensing deal with Reddit, the large number of partnerships it has with publishing businesses are built around its News Showcase product, rather than content for AI. As competition authorities examine its conduct around Gemini and search, it may soon have to start securing explicit authorization for access to content to fuel these products. This may in turn lead to a programme of deal-making for access to content for AI. Reporting suggests this may have already started[10].

2.5 Meta

Meta has developed the Llama family of open-source AI models. These are used in Meta's own products and extensively across third-party applications. Whilst historically it has relied on external datasets for model training, it has recently launched a new web crawler to collect data for its LLMs.

Bots

Meta-ExternalAgent

Meta describes this user agent as crawling the web for 'use cases such as training AI models or improving products by indexing content directly'.

Meta-ExternalFetcher

This bot accesses websites in real-time in response to actions by users. The precise products or functions which it serves are unknown. Meta is transparent that this crawler may bypass robots.txt on account of being user-'initiated'.

FacebookBot

Meta described this bot as crawling public web pages to improve language models for its speech recognition technology. It has since removed this description from its developer site and it is not known whether the bot is now being used for different purposes.

Robots.txt policy at time of publication

Meta's crawlers respect robots.txt signals although the Meta-ExternalFetcher bot may bypass the protocol because it performs crawls that were user-initiated.

Publisher partnerships

Meta has signed an AI licensing deal with just one publisher - Reuters. This agreement is focused on real-time access to Reuters content, allowing Meta's chatbots to answer user questions about current events.

2.6 Apple

Apple has been investing in AI - including the development of its own foundation models - to enhance its products, particularly in privacy-centric applications and on-device AI capabilities.

Bots

Applebot
This user agent acts as both a search indexing crawler and a RAG agent feeding information to power features across Apple's ecosystem, including Siri, Spotlight and Safari.

Applebot-Extended
Apple's primary user agent is AppleBot. This is used to collect data to feed into a variety of user products in the Apple ecosystem, including Spotlight, Siri and Safari. Applebot-Extended is a secondary user agent that allows publishers to opt-out of their content being used to train Apple's foundation models. Applebot-Extended does not crawl webpages; it is only used to determine how Apple can use the data crawled by the primary Applebot user agent.

Robots.txt policy at time of publication
Applebot-Extended respects robots.txt directives, allowing website owners to control the use of their content for AI training.

Publisher partnerships

Apple has not publicly announced any AI content licensing deals at time of writing but it is reportedly in negotiations with a number of publishers including Condé Nast and NBC News[11].

2.7 Amazon

As well as a strategic partnership with Anthropic, Amazon has developed its own 'Nova' family of AI models which emphasize speed and value.

Bots

AmazonBot

Amazon describes AmazonBot as being used to improve services, 'such as enabling Alexa to answer even more questions for customers'. There are no published details of how it operates, or what the data it captures is used for.

Robots.txt policy at time of publication

AmazonBot respects standard robots.txt rules.

Publisher partnerships

Amazon is reportedly in licensing negotiations with a number of news outlets for access to content that will give a revamped Alexa the ability to answer questions about current events[14].

2.8 ByteDance

ByteDance - owner of TikTok - has developed both text and video generation models. Most of these are for research purposes although its Doubao text models power its chatbot, only available in China.

User agents

Bytespider

This bot has been scraping the web at a high rate since it first appeared in early 2024. ByteDance has published no information on the function it serves or what the data collected is being used for.

Stated robots.txt policy at time of publication

ByteDance has no published robots.txt policy. There are widespread reports of Bytespider ignoring robots.txt signals.

2.9 Other notable AI User Agents

DuckAssistBot

DuckAssistBot is DuckDuckGo's web crawler. This bot crawls pages in real-time to source information for answers by DuckAssist - the search engine's AI answer feature. According to the information published by DuckDuckGo, data collected is not used to train AI models and it respects robots.txt signals.

Timpibot

Timpibot is the web crawler for Timpi, a decentralized search index, accessible for a cost to businesses. The data collected by Timpibot is also available to AI developers for model training[12]. At the time of writing there is no published information on Timpi's robots.txt adherence.

YouBot

YouBot is the crawler for You.com, an AI-powered search engine that integrates AI query responses alongside conventional search links. It indexes websites to provide AI-driven search capabilities. At the time of writing there is no published policy for YouBot's compliance with robots.txt.

Diffbot

Diffbot is a web crawler focused on extracting data from web pages which it then converts into structured datasets for businesses and developers. The stated policy at the time of writing is that Diffbot fully respects robots.txt directives, allowing granular control over its crawling behavior.

[1] First Page Sage (n.d.) *Top generative AI chatbots*. Available at: https://firstpagesage.com/reports/top-generative-ai-chatbots/ (Accessed: August 2025). (Accessed: August 2025).

[2] OpenAI (2025) *OpenAI: ChatGPT on track to hit 700 million weekly active users*, CNBC, 4 August. Available at: https://www.cnbc.com/2025/08/04/openai-chatgpt-700-million-users.html

[3] Brown, P. (n.d.) AI partnership tracker. Available at: https://petebrown.quarto.pub/pnp-ai-partnerships/ (Accessed: August 2025).

[4] Business of Apps (n.d.) Perplexity AI statistics. Available at: https://www.businessofapps.com/data/perplexity-ai-statistics/ (Accessed: August 2025).

[5] Mehrotra, D. and Marchman, T. (2024) 'Perplexity is a bullshit machine', Wired, 19 June. Available at: https://www.wired.com/story/perplexity-is-a-bullshit-machine/ (Accessed: August 2025).

[6] First Page Sage (n.d.) *Top generative AI chatbots*. Available at: https://firstpagesage.com/reports/top-generative-ai-chatbots/ (Accessed: August 2025).

[7] Guaglione, S. (2024) 'How Perplexity calculates publishers' share of ad revenue', Digiday, 17 December. Available at: https://digiday.com/media/how-perplexity-calculates-publishers-share-of-ad-revenue/ (Accessed: August 2025).

[8] First Page Sage (n.d.) *Top generative AI chatbots*. Available at: https://firstpagesage.com/reports/top-generative-ai-chatbots/ (Accessed: August 2025).

[9] Business of Apps (n.d.) *Google Gemini statistics*. Available at: https://www.businessofapps.com/data/google-gemini-statistics/ (Accessed: August 2025).

[10] Love, J. and Miller, H. (2025) *Google in licensing talks with news groups, following AI rivals*, Bloomberg News, 22 July. Available at: https://www.bloomberg.com/news/articles/2025-07-22/google-seeks-licensing-talks-with-news-groups-following-ai-rivals (Accessed: August 2025).

[11] Mullin, B., Mickle, T. (2023) 'Apple discusses AI with news publishers', *The New York Times*, 22 December. Available at: https://www.nytimes.com/2023/12/22/technology/apple-ai-news-publishers.html (Accessed: August 2025).

[12] Timpi (2025) *Data API,* Timpi [online]. Available at: https://timpi.io/data-api/ (Accessed: August 2025).