Website Traffic in the Age of AI Crawlers: Lost Referrals, Blocking and Reliability

SIAI Research Editorial

Swiss Institute of Artificial Intelligence, Chaltenbodenstrasse 26, 8834 Schindellegi, Schwyz, Switzerland

Abstract

The study examines how responses generated by language models and AI crawlers remove traffic from websites while simultaneously collecting their content without compensation. Data from the Pew Research Center, Cloudflare, HUMAN Security and TollBit show fewer clicks, more automated traffic and crawl-to-referral ratios in the thousands. The reliability of the responses remains problematic: the European Broadcasting Union's study found significant issues in 45 percent of answers on current affairs. Major publishers respond by blocking crawlers, so the most rigorously edited sources withdraw from the material that feeds the models. The study examines open-weight models, licensing agreements and Google's pilot payment programme and concludes that without verifiable crawler identity and an auditable valuation of contributions, every payment will remain a concession rather than an obligation.

1 Introduction - Website Traffic after the Answer Engine

The public debate about reducing website traffic usually describes a change of medium. Readers move their questions from Google's search engine to apps like ChatGPT, Gemini, and Perplexity. Google is losing share, and publishers are being asked to adapt to a new distribution channel, as they had previously adapted to Facebook and mobile devices. In this reading, website traffic is not lost, it just changes route. But the accounting doesn't work. A click on a Google results page resulted, in a significant percentage, in a website that could sell advertising, a subscription, or simply win a reader who would return. A language model's response usually gets nowhere because the reading is completed within the answer itself.

Available data points in this direction, with scale differences depending on the method. The IAB Tech Lab, the body that coordinates technical standards for digital advertising, estimates that AI-generated search summaries reduce publisher traffic by 20 percent to 60 percent, with losses of up to 90 percent on niche websites, and estimates the total loss of advertising revenue at about $2 billion.[1] This estimate aggregates third-party data and should be read as a range. More testable is the measurement by the Pew Research Center, which analyzed 68,879 searches of 900 U.S. adults in March 2025. Those who viewed an AI summary clicked on a traditional results link in 8 percent of visits, compared to 15 percent when no summary was displayed, while a link within the summary itself was clicked on only 1 percent of visits.[2] The metric is for a US panel and a single product at a specific point in time, but the direction is clear and coincides with what publishers are reporting.

At the same time that people are thinning out, machines are increasing. According to Cloudflare data published by Fortune in July 2026, automated traffic surpassed human traffic in June of the same year and reached 57.5 percent of website requests. Cybersecurity firm HUMAN Security recorded a 7,851 percent year-over-year increase in traffic from AI agents performing actions such as link clicks and form filling, and a 597 percent increase in traffic from content scrapers.[3] Cloudflare also introduced an indicator that captures asymmetry more clearly than any percentage: the ratio of pages a platform crawls to back-referencing visits. For the week of June 19 to 26, 2025, the ratio for Anthropic was around 70,900 to 1, with the caveat that mobile apps do not indicate the referral source and that the ratio may therefore be overestimated.[4] Even with a brave correction, the distance from the logic of traditional search is enormous.

The present study argues that the issue is not limited to Google's loss. The removal of traffic is linked to two other developments that are usually discussed separately. The first is reliability: systems that replace click produce responses with measurable and persistent error rates, and the reader no longer has the source in front of them to judge it. The second is the reaction of publishers, who are increasingly blocking crawlers, causing the most thoughtful sources to withdraw from the very material on which the answers are based. Google hadn't done a great job, and language models don't necessarily make every query worse. The difference lies in the fact that the previous system, with all the distortions of search engine optimization, tied content quality to a fee, while the new one disconnects it.

The timing explains why the issue became urgent within three years. Google's AI Overviews summaries launched in the U.S. in May 2024[5] and expanded to dozens of markets, conversational apps gained real-time

search functions, and as of 2025, autonomous browsing agents began to replace part of human browsing. Each of these steps transferred a piece of the reading from the website to the platform. The question that follows is practical before it becomes prescriptive: what exactly mechanisms that remove the traffic, and who pays the cost.

2 How AI Crawlers and Answer Engines Remove Website Traffic

Two mechanisms work in parallel and reinforce each other. The first acts on demand: the answer replaces the click. The second acts on supply: crawlers collect the content that will feed those responses, and the collection has its own cost for the site. In traditional search, the two sides were tied together. Google crawled a page to index it, and indexing existed to show a link that someone would click. The IAB Tech Lab quantified this split in November 2025: large language models perform between 250 and several thousand crawls before referring a visit, compared to about six for Google search, and have already removed up to 60 percent of traffic from referrals for publishers in the U.S.[6] It went as far as advising publishers to block crawlers.

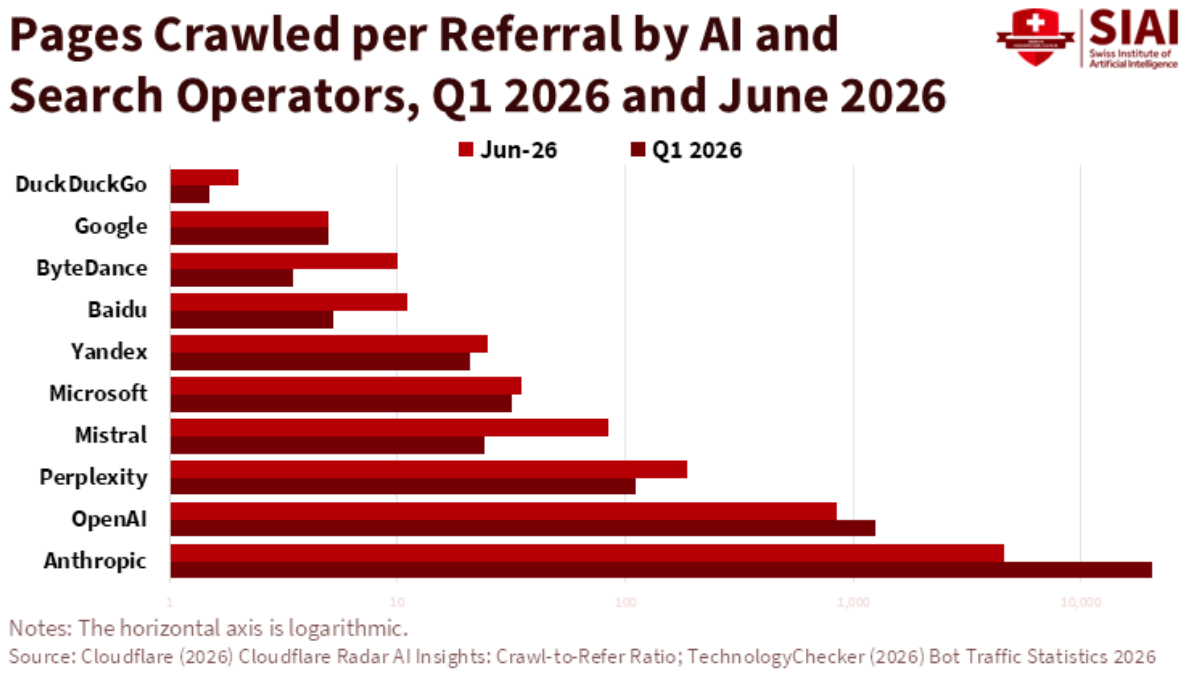

Figure 1: AI crawlers take thousands of pages for every visit returned.

The loss is unevenly distributed. Major news organizations have direct traffic from loyal readers, subscriptions, and the ability to negotiate. A niche blog lives almost exclusively from search, because no one types in the address of a travel blog by heart when looking for where to eat in Lisbon. The IAB Tech Lab cites the case of the travel blog The Planet D, which recorded a 90 percent drop in traffic in the initial phase of Gemini's integration into search, and links it to the concentration of the advertising market: according to the IAB and PwC 2024 report, 80.8 percent of online advertising revenue ends up in the ten largest companies.[7] When distribution goes through responses controlled by a few platforms, the bargaining position of the independent creator deteriorates from an already weak starting point.

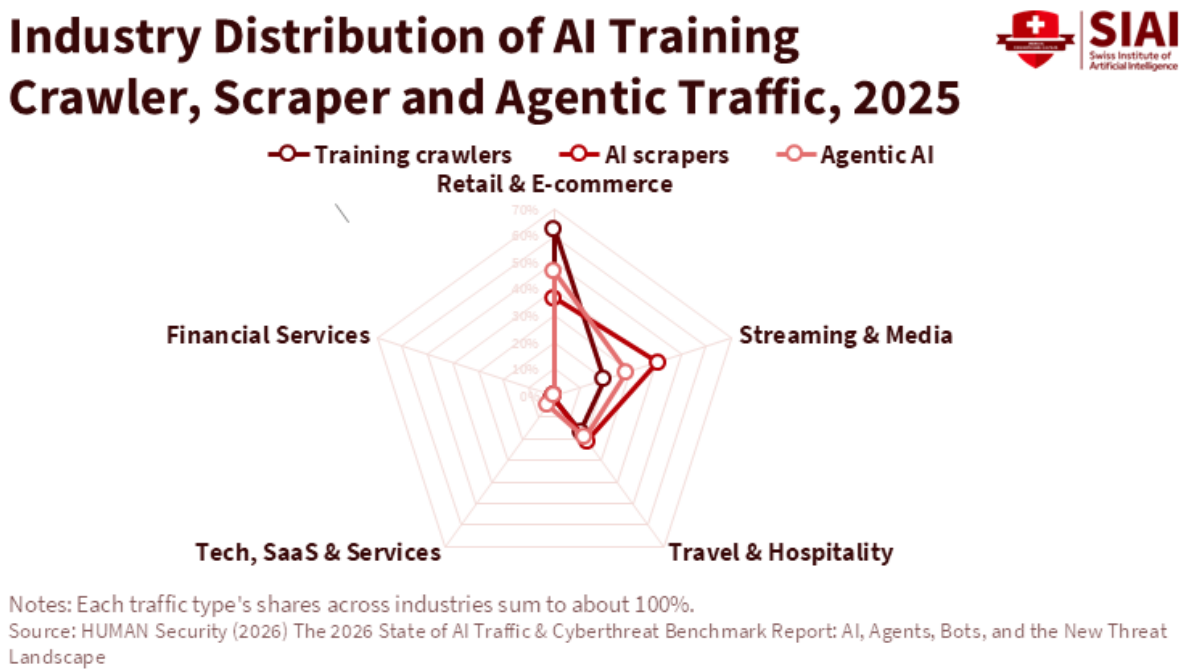

Figure 2: Media is the top target of live AI scraping.

The second mechanism became visible in an almost comical way, before becoming annoying. From the fall of 2025, websites of all sizes began to see in their statistics thousands of ”visitors” from Lanzhou, an industrial city in northwest China with no reputation as a technology center, with some of the traffic routed through Singapore. WIRED recorded in February 2026 the case of a website from Bogota about paranormal phenomena, written in a mixture of Spanish and English, in which China and Singapore had become the source of more than half of the traffic in a year, with an average visit duration of zero seconds. According to the Analytics.usa.gov platform, Landshow accounted for 14.7 percent and Singapore accounted for 6.6 percent of visits to U.S. federal government websites in the previous quarter.[8] Who is behind the traffic remained open. The geographical attribution of an IP address shows where the infrastructure is registered, not necessarily who is using it, and WIRED itself noted that experts could not explain why Landshow appears so systematically. The assumption that this is a collection of training data is plausible, but it is not proven.

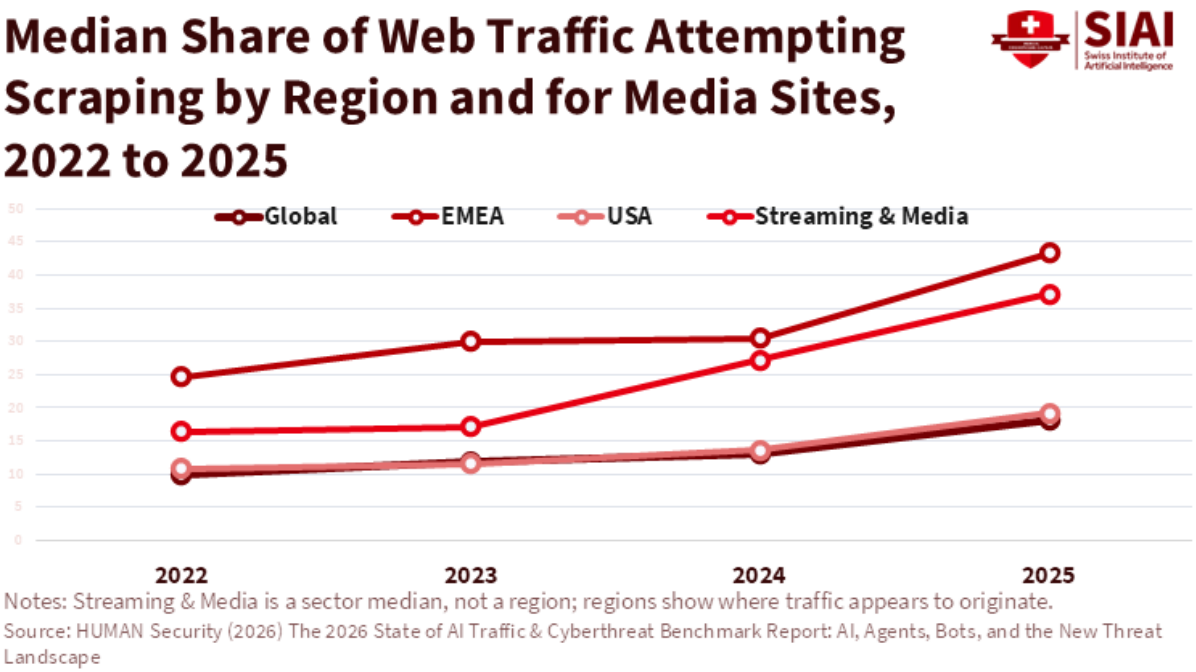

Figure 3: Scraping has more than doubled on media sites since 2022.

The scale of the phenomenon is not limited to one city. TollBit, which tracks the movement of AI robots on publisher websites, counted in the fourth quarter of 2025 one robot visit for every 31 humans, up from one for every 200 at the beginning of the same year.[9] For small websites, the consequences are tangible and have a price. The Breached.Company team described how in January 2026 one of its websites was removed from the Ezoic advertising network without explanation, and how the investigation led to the same pattern of traffic from China and Singapore. An administrator citing the same analysis recorded 127,000 daily bot visits at its peak, dropping to 2,000 after blocking entire autonomous systems belonging to Chinese cloud providers.[10] When an ad network sees hundreds of thousands of zero-length sessions, it reasonably concludes that the publisher is inflating its numbers, and the publisher is penalized for a move it didn't request.

The picture is complicated by the fact that part of this "traffic" does not even touch the website. Twenty-First Digital, which advises publishers on measurement issues, distinguishes two scenarios. In the first, automated systems actually load the page, the Google Analytics measurement code runs normally, and the session is recorded, so the traffic also appears in the server files. In the second, fake signals are sent directly to Google Analytics, without any contact with the website, and no server-level blocking can stop them.[11] The practical result is that the publisher simultaneously loses readers and the ability to count which readers are left. Engagement rates, reading time and geographical distribution, the quantities by which it decides what to write and how to sell it, are distorted.

It would be a mistake to read the phenomenon as exclusively Chinese. Many publishers are seeing similar waves from Ashburn, Virginia, the suburb of Loudoun County where the world's densest concentration of data centers is located and where Amazon Web Services, Microsoft, and Google maintain large facilities. Geolocated traffic in Ashburn is almost always cloud infrastructure traffic, and that includes the declared crawlers of American AI companies, undeclared agents, monitoring tools, and VPN networks, all mixed under the name

of a suburb with a few tens of thousands of residents. The distinction between American and Chinese origin matters to the jurisdiction, as will be seen below, but for the publisher the result is the same: the people who were reading are removed from the response, and in their place appear machines that consume bandwidth, distort the measurements and take the text.

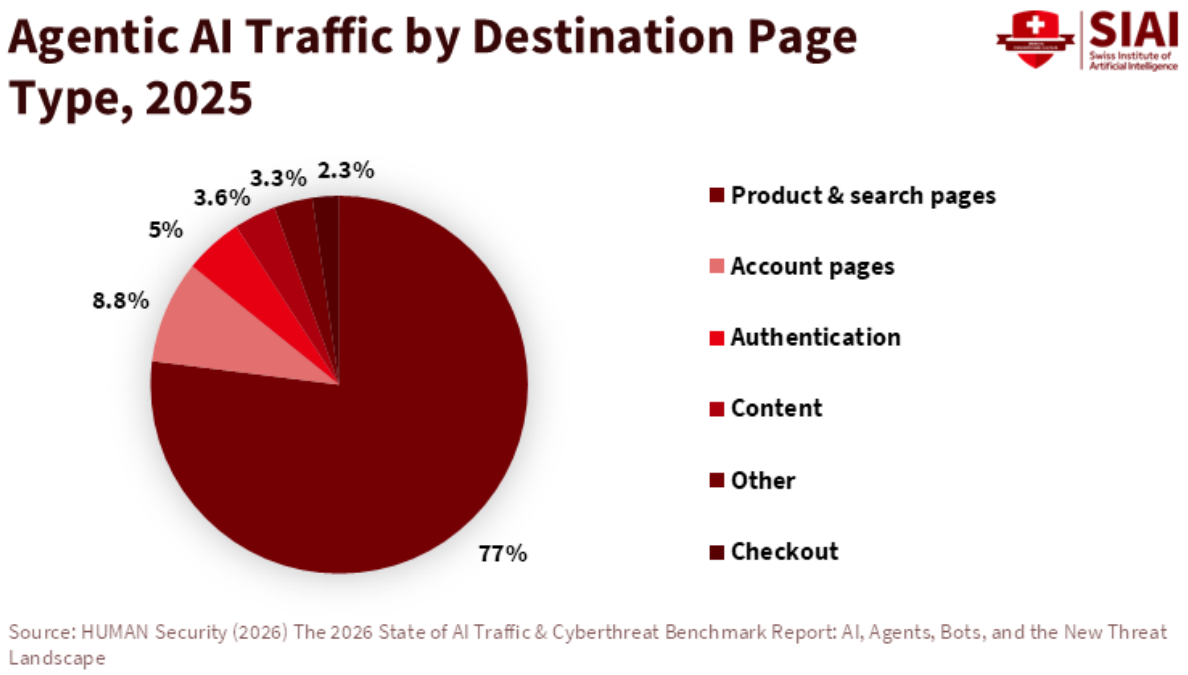

Figure 4: AI agents now reach accounts and checkouts, not just pages.

If all this were simply switching from one search medium to another, the damage would fall mainly on Google and would be a matter of competition. This assumes that the other participants, the creators, the readers and even the providers of the models themselves, are in the same or better position than before. That is what needs to be checked.

3 Who Gains from Models Trained on Website Data

A simple profit and loss account shows that the losers are more than one. Search engines are losing clicks, and Google itself is replacing results pages that brought ads with summaries that bring them harder. Content creators are losing traffic with the mechanisms described. Readers are gaining speed, but they cannot know if the answer they are reading is correct. The last leg is the least visible and perhaps the most serious, because it concerns people who have no financial connection to the dispute between publishers and platforms.

The largest assessment of AI assistants in news to date was coordinated by the European Broadcasting Union and led by the BBC and published in October 2025. Twenty-two public broadcasters from 18 countries, in 14 languages, evaluated more than 3,000 ChatGPT, Copilot, Gemini, and Perplexity responses to current affairs questions. 45 percent of responses had at least one major issue, 31 percent had serious source rendering issues, meaning sources that were missing, misleading, or incorrect, and 20 percent had serious accuracy issues, with made-up details and outdated information. Gemini had significant issues in 76 percent of responses, mostly

due to source attribution.[12] The significance of the outcome depends on how many already rely on these tools. The Reuters Institute's Digital News Report 2025 recorded that 7 percent of those who update online use AI assistants weekly for news, a figure that rises to 15 percent for those under 25.[13] The rates are small in absolute terms, rising precisely at the ages that will shape the habits of the next decade.

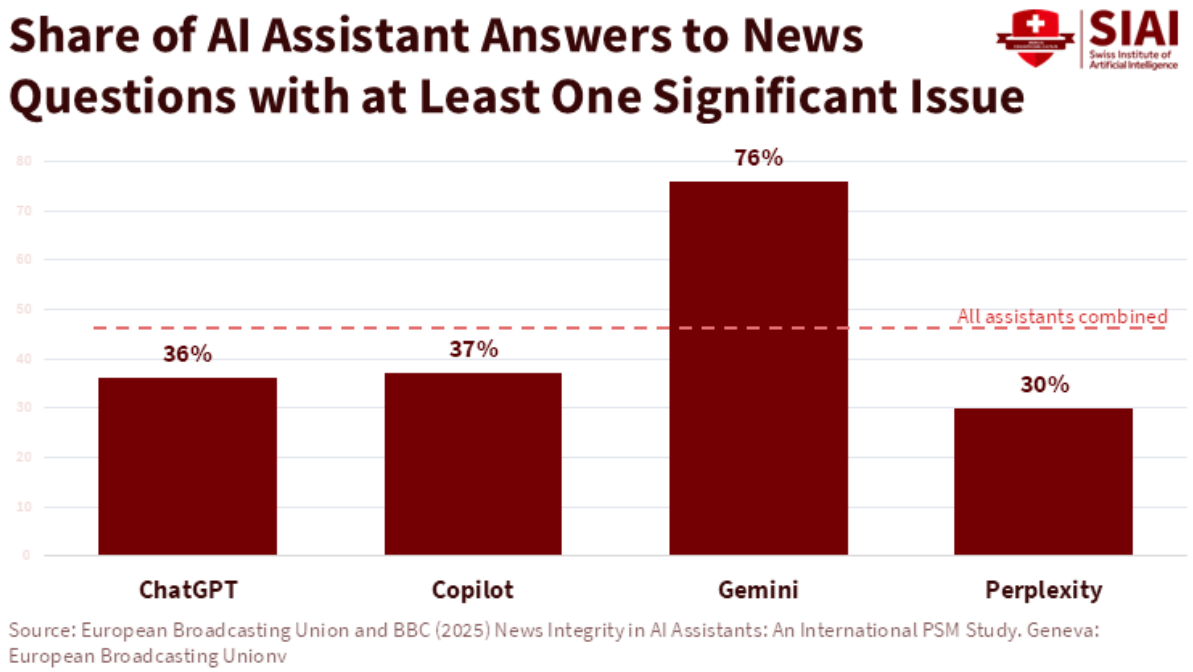

Figure 5: Gemini's answers carried significant issues more than twice as often as ChatGPT's, Copilot's or Perplexity's.

Columbia University's Tow Center for Digital Journalism had come to a similar conclusion with a different method. It tested eight search tools based on language models asking them to identify the title, publisher, date, and address of articles from snippets, and found that overall they answered more than 60 percent of queries incorrectly. Paid versions of some tools gave incorrect answers with greater certainty than free ones, and several tools built links that led nowhere.[14] The finding for certainty is critical. A tool that says "I'm not sure" transfers to the reader the responsibility of checking. A tool that answers incorrectly in the same tone that answers correctly removes it.

It could be assumed that the problem is solved when the model works on specific documents rather than from its "memory." A study from Northwestern University and the University of Minnesota tested exactly this scenario: it gave ChatGPT, Gemini, and NotebookLM a corpus of 300 documents about the legal and regulatory dispute surrounding TikTok in the U.S. and asked for journalistic research-type tasks. 30 percent of the responses contained at least one hallucination, with about 40 percent for ChatGPT and Gemini and 13 percent for NotebookLM. Most of the errors were not invented entities or numbers. It was what the authors called interpretive overconfidence: the models added characterizations of sources that the sources did not support and converted opinions attributed to specific actors into general findings.[15] For a publisher this carries particular weight. An opinion article arguing that a policy will fail may appear in response as a given that the policy failed, without a name, without a link, and without the reservation that the author had written.

If readers come out uncertain and creators lose, the obvious next step is to look at model providers as beneficiaries. The available financials don't easily support this picture. According to shareholder documents cited by The Information, OpenAI consumed $3.7 billion in cash flow in the first quarter of 2026, with revenue of $5.7 billion over the same period.[16] Each answer has a computation cost, and in much of free use that cost is not covered by revenue. The bleeding is, quite literally, in tokens. Margin accumulates upstream, in hardware and infrastructure. Nvidia reported revenue of $130.5 billion for the fiscal year ended January 2025 of which $115.2 billion came from the data center division.[17] In this chain, those who sell chips, data center space, and paid development work are in a much better position. The other links, from the editor of a blog to the provider of the model that subsidizes the answers, are either losing or betting on future profitability.

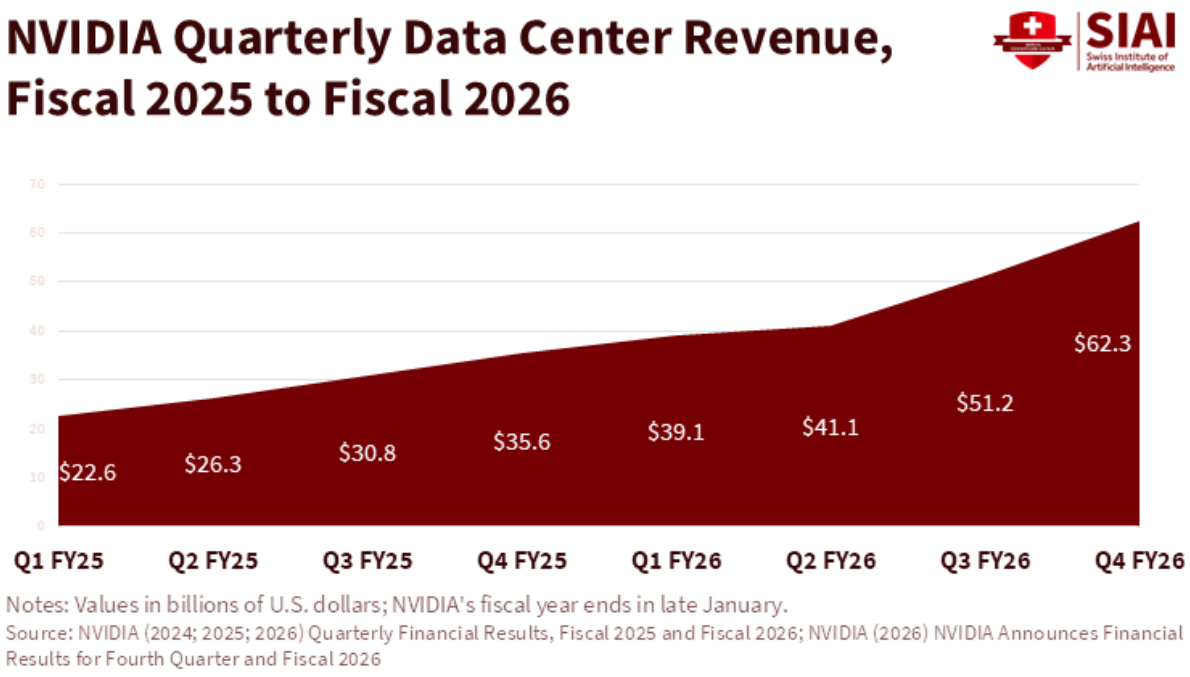

Figure 6: Hardware revenue kept climbing while model makers burned cash.

The strongest counterargument deserves serious consideration. Summaries and conversational responses save time, give unqualified people access to explanations that would otherwise require hours of reading, and serve queries for which no one really needed to visit a page. Click-free search, after all, didn't start with AI. Google's answer snippets and knowledge boxes kept users on the results page for years, and Pew itself recorded that about two-thirds of searches ended up without clicking on a result, regardless of the summary display. The argument correctly describes a portion of the value to the user. It doesn't answer two things: that the summary almost doubles the loss for queries where the user would step somewhere, and that the value for the user is produced with material that someone else paid to exist.

This leads to the question of whether the models can be made reliable enough for the replacement of the source to be acceptable. The explanation given by researchers at OpenAI and Georgia Tech in September 2025 is enlightening and less reassuring than they might have liked. Hallucinations, they argue, arise as binary classification errors under the natural statistical pressures of pre-training: where incorrect propositions cannot be distinguished from true ones, the model will produce plausible falsehoods. They persist because most

evaluations score in a way that rewards guesswork over assuming uncertainty. For events without an internal pattern, such as a date of birth, the hallucination rate of a baseline model is expected to be at least as much as the percentage of events that appear just once in the education.[18] Duke University Libraries, in a January 2026 analysis, came to a compatible conclusion: hallucinations thicken where data is sparse, contradictory, or of low quality, and the model does not inherently know which sources are reliable.[19]

The scarcity of data on a topic therefore increases errors. But it is not the root. A linguistic model produces the most likely continuation of a text based on statistical regularities, not a claim that has been tested against some representation of the world, and this difference is not eliminated by adding more texts. It is improved by retrieving sources, by training that rewards abstention, and by better calibration, but it remains a property of the method. Herein lies the essential difference from Google's results page, which was full of inadequacies. In front of a list of ten links, the reader decided which one to open, saw the name of the publisher, held a rudimentary judgment about whether a medical question was better answered by a hospital or by a forum. In front of an answer, the choice has already been made by the system, and the reader simply reads. The rate of 1 percent of click-through visits within the summary shows how rarely the reader returns to the source to check. The responsibility for verification shifts from the reader to a system that does not bear it, neither legally nor commercially.

4 Why Publishers Lack Control over AI Use of Their Content

The word "correction" assumes that there is someone who can intervene and a mechanism through which the intervention reaches the result. For hallucination, the providers themselves talk about reduction, not elimination. OpenAI, presenting the analysis discussed above, stated that GPT-5 has significantly fewer hallucinations, especially when it uses reasoning, that they still occur, and that they remain a fundamental challenge for all major language models.[20] In a system that answers hundreds of millions of queries a day, a small error rate is a large absolute number of incorrect answers, and there is no way for the reader to know which category the one they are reading belongs to.

For the remuneration of creators, the answer is even less encouraging, and a comparison with the previous regime helps to show why. Google didn't pay publishers for indexing. But it sent them traffic, and the traffic was converted into revenue through advertising or subscriptions. The race for the first page of results had its pathologies, from content farms to texts written for algorithms instead of humans, but it maintained a basic connection: better content was more likely to be read, and reading had value for whoever wrote it. The answer engine maintains the first half relationship, since it selects and leverages useful content, and cuts the second.

An inability to check is what makes the situation different from a simple decrease in revenue. A publisher cannot decide whether their text will be used in a response that falsifies it. They cannot choose in which responses and next to which other sources to appear. If a response attributes something they did not write, there is no correction mechanism equivalent to a correction request to a newspaper. If they update an article because the information has changed, they cannot ask the provider to replace the old version that was already absorbed in the training. Often they do not even know if and when its content was detected. Since 2 August

2025, the EU AI Act has required providers of general-purpose models to publish a summary of their training content, using a template the European Commission issued on 24 July 2025. The template asks for a narrative summary of sources and a reference to the most important domain names from which data was collected, not a per-site archive of what was collected and when.[21] It's a market-level transparency step. For the individual website, almost nothing changes.

The tool that is supposed to give control, the robots.txt file, is a polite request that nothing enforces. Cloudflare published an analysis in August 2025 that Perplexity used undeclared crawlers, changing IP addresses and browser identities, to access websites that had explicitly requested not to be crawled.[22] A January 2026 Press Gazette analysis noted that robots.txt guidelines can be ignored or bypassed through third-party companies that collect content on demand, and that the IAB Tech Lab, reviewing much of the publishers' guidelines, found a host of typos and incorrect settings that render them ineffective. The same analysis also identifies a structural problem with Google: the Google-Extended option allows the publisher to opt out of Gemini training, but not from using the content in AI Overviews summaries. To avoid summaries, a publisher would have to block Googlebot, i.e. disappear from search.[23] Crawl for search and crawl for responses are tied, and the publisher can only deny both together.

Under these circumstances, blocking became the dominant strategy of major publishers. In February 2024, 61 of the 106 largest news sites in the UK and the US were blocking at least one AI crawler.[24] In January 2026, Buzzstream platform's analysis of nearly a hundred top websites in the two countries found that 79 percent were blocking at least one training tracker and 71 percent were blocking crawlers that retrieve real-time content for responses. Among the 50 largest, 34 percent blocked all crawlers examined, with the BBC, the New York Times, the Daily Mail, the Telegraph, the Associated Press and the Wall Street Journal among them, while Anthropic's training crawler was the least allowed, with access to just eight of the fifty. The Telegraph explained its position in terms of exchange: almost no value returns, language models are not designed to send traffic, and companies are not willing to pay for the content they were trained on. Belgian public broadcaster VRT, which participated in the European Broadcasting Union study, used its findings to justify restricting AI assistants' access to its news content. At the infrastructure level, Cloudflare started on July 1, 2025, by default blocking AI crawlers for the new domains it serves, while offering a pilot bill-per-crawl system.[25] Millions of websites that had never opened a robots.txt file found themselves blocked without deciding for themselves.

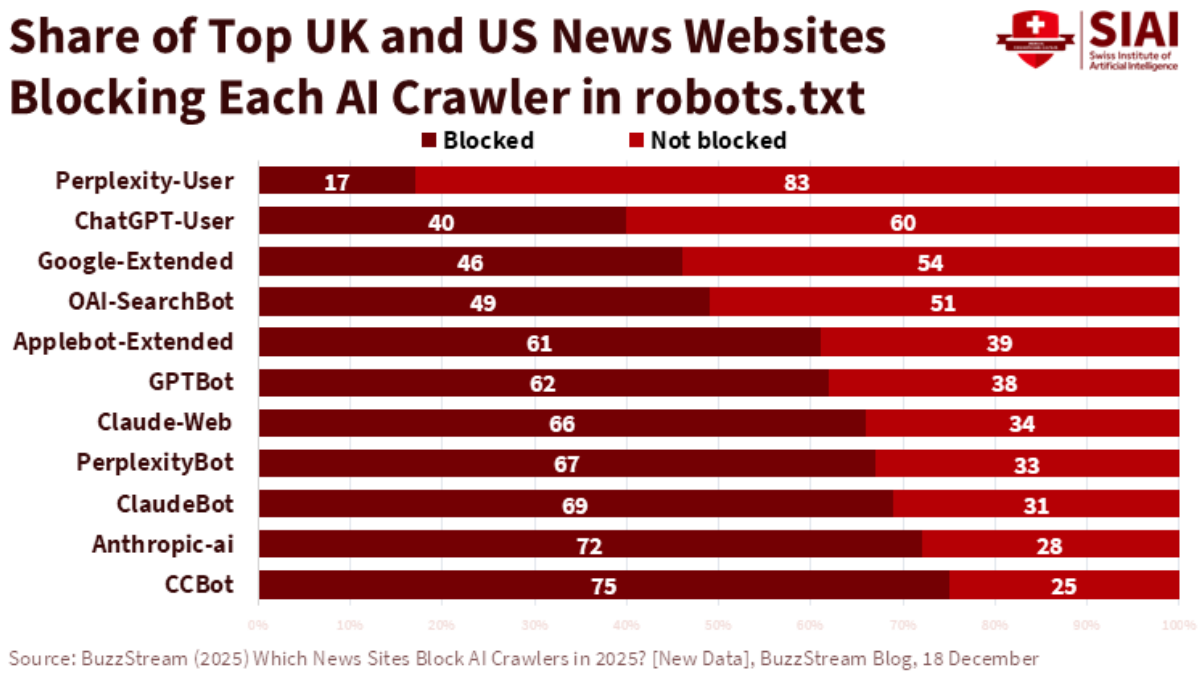

Figure 7: Publishers block training bots far more than retrieval bots.

The academic literature on content ”control” in relation to artificial intelligence gives little help to the editor who sees their text absorbed. A typical proposal, presented at a workshop at the NeurIPS 2025 conference, describes a multi-agent system, with design, production, revision, integration, and protection agents, that incorporates subtle watermarks during the production process itself.[26] The approach is useful for the traceability of material produced with AI tools. But it protects the output of a creator who already uses such tools, and offers nothing to the author of a text written by a human, openly published, and collected for training. The problem of input remains untouched.

Here the two threads of the study, traffic and credibility, are tied together in a loop that has no obvious output. The sources that invest the most in checking their data, the big news organizations, the public broadcasters, the specialized publications with paid editors, are the ones that exclude first, because they have the resources and the bargaining power to do so. The more they withdraw, the greater the part of the answers is based on what remains open: on websites that live off search engine optimization, on republications, on copies. The Tow Center found that some tools displayed content from publishers that had blocked their crawlers, drawing it from republished versions on other sites. Blocking, in other words, does not fully protect either the publisher, who sees their material circulating without attribution, or the reader, who gets a second-hand response. And for the traffic that is geolocated in Lanzhou or Singapore, where it is not even clear who is crawling, there is no counterparty to negotiate, no competent authority to complain or a standard that is committed to respecting, so the publisher resorts to the only mechanism available, the blocking of entire networks, at the cost that along with the bots, it also blocks any real reader from the region who happens to pass through the same providers.

5 What Needs to Change: Open Weights and Payment for Creators

Two directions emerge from the analysis, and none is sufficient on its own. The first concerns the nature of the product. A general-purpose language model, with the exception perhaps of its capabilities in code, where companies have invested extensively in their own data and in human labeling, is largely a distillation of texts written by millions of people without being asked. From this perspective, the model is a common product and not an exclusive property, and the minimum obligation of anyone who trains it is to make it available with open weights, if not as fully open software with published training data. The argument has gained practical weight. In January 2025, DeepSeek launched the R1 reasoning model with open weights under the MIT license, proving that a model competitive with closed weights can be freely available.[27]

Open weights, however, solve a different problem than that of the creator. They return value to the public in the form of access: researchers can control the behavior of the model, public organizations can host it without depending on a provider, small businesses can adapt it. They don't pay any editor, they don't give them control over the use of their text, and they make further use even more undetectable, since a model with open weights can be embedded anywhere. Nor do they reduce hallucination, which, as it turned out, are a property of the method. There is a more uncomfortable side: the very idea that the written production of humanity is a common good can be used as a justification for the very uncontrolled collection described in the first chapter. Whether openness compensates for expropriation or simply makes it more acceptable depends on what else accompanies it.

The second direction is payment, and its problem is who gets paid. The licensing agreements that have become known concern organizations with pooled records and bargaining power. Reddit agreed with Google in February 2024 to make its content available for training for about $60 million per year, according to Reuters, followed by a deal with OpenAI in May of the same year.[28,29] But the texts were written by Reddit users, who received nothing. In the book space, the settlement in the Bartz v. Anthropic case amounted to $1.5 billion. That's about $3,000 for each of about 500,000 works.[30] The case showed that using a work for training can gain a price when asked for by a court, but it involved books that had been acquired by pirate libraries and protected authors with publishers behind them. A single blog editor has neither the resources for a corresponding trial nor a catalog of works large enough to interest a buyer. If the licensing market remains a large archive market, independent online writing will continue to feed models for free until it ceases to be produced.

The most advanced step towards payment at the individual page level has been taken, perhaps unexpectedly, by Google. According to Digiday in September 2026, the company is scaling up a pilot program within Search Console that pays publishers when their content contributes "significantly" to Gemini, AI Overviews, and AI Mode responses. Participants see a monthly amount of earnings with no explanation of how it's calculated, the program has attracted more interest from small and medium-sized publishers than large groups, and is expanding beyond news. Publisher executives described the bids as low, and one of them warned that participation could weaken publishers' bargaining position, since Google will be able to claim that it is already compensating.[31] The logic of the program, payment per contribution rather than per file, is exactly what a small creator needs. Its implementation, with unilateral valuation, non-transparent form and participation only by invitation, leaves

the power to set the price in the same place it was before.

Other schemes have emerged around this initiative, with a different philosophy. Cloudflare introduced a per-crawl fee in 2025, which turns blocking into a price. The IAB Tech Lab formed a working group on AI-powered content utilization protocols, which included Google, Microsoft, and Meta, while OpenAI, Anthropic, and Perplexity, according to the agency, did not respond. The Really Simple Licensing standard, introduced in September 2025 with the support of Reddit, Yahoo, and Medium among others, it allows a website to declare in a machine-readable format the licensing terms of its content.[32] In the Internet Engineering Task Force, the Working Group on AI Usage Preferences is working on a common vocabulary by which a website can distinguish between education, search, and other uses.[33] None of this is mandatory for anyone crawling, and this is their common limit.

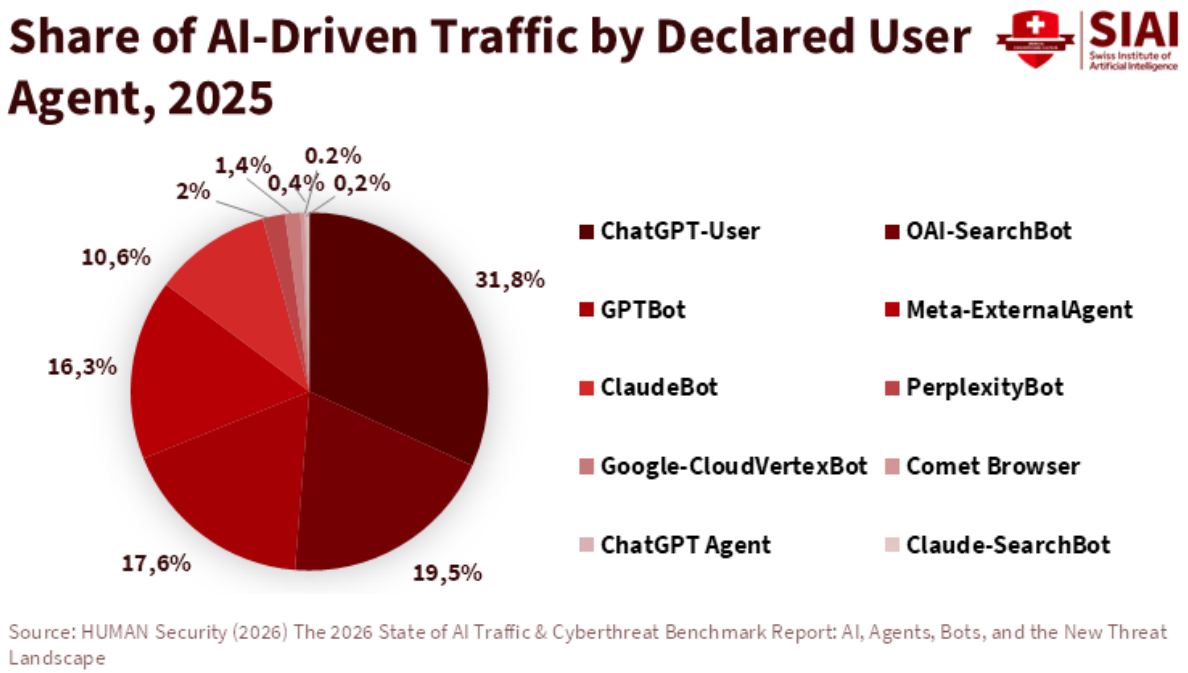

Figure 8: Firms absent from licensing talks generate most AI bot traffic.

The necessary changes are distributed among different actors, and the confusion of roles is part of the problem. Companies that develop models can, without any legislation, publicly separate their crawlers by purpose, cryptographically sign their requests so that their identity is verified, give each website access to the file of crawls concerning them, and open a correction channel for claims attributed to a specific source. They can also apply what OpenAI's own analysis suggests, i.e. scoring that rewards abstention when sources are insufficient and make visible attribution with a link by default to any response based on retrieved content. Google and other answering engines experimenting with payments should publish how the contribution is measured, show data by page address and open the programmes on any website that meets the minimum criteria, without an invitation. Publishers and independent creators have an interest in organising themselves into collective licensing bodies, along the lines of music collecting societies, because only pooled producers acquire a catalogue of commercial interest. The IAB Tech Lab has argued that without scarcity there is no market, and concerted exclusion is the only way to create scarcity, but at the risk of being seen by competition authorities as a cartel.

Public intervention is justified where the market cannot coordinate on its own, not in price setting. A register of crawlers with a mandatory verifiable identity, the legally binding nature of machine-readable opt-out statements that will result from the Internet Engineering Task Force standards, and the right of every website to ask a provider if and when it has collected its content, as an extension of the European summary model, are measures that no single company has any incentive to adopt first. Competition authorities have reason to consider linking crawling-to-search and crawling-to-responses in the case of Google, because that's where the publisher has no real choice. For traffic coming from jurisdictions where none of this is applicable, transnational cooperation is the only avenue, and progress in this area has so far been minimal.

6 Conclusion - Content and Computation Both Carry a Cost

Artificial intelligence is not free. It is calculated on chips, energy, data centers and salaries, and the companies that offer it today pay more than they receive. Nor is content free. It was paid for by someone, with advertising that relied on readers, with subscriptions, with the unpaid time of a person writing about a topic they knew well. The transition from the results page to the answer page maintained the cost of the computational side and implicitly removed the mechanism that covered the cost of the content.

The analysis showed that the damage is not limited to Google. Creators lose readers and the ability to count those who remain. Readers get answers with errors that are measured in double-digit percentages, formulated with certainty that does not correspond to their accuracy, and without the source in front of them. Sources with the strictest control are withdrawn, and answers are increasingly based on what is left open. Deterioration of credibility and loss of traffic are aspects of the same process.

Open weights return public access, not remuneration. Licensing agreements pay those with large files. Google's pilot program is the first to pay per contribution, with terms it unilaterally sets. Whether these will evolve into an working system or remain symbolic gestures cannot yet be judged.

The minimum necessary point is clear. Without verifiable identities of crawlers, binding exclusion statements, and a third-party verifiable contribution assessment, each payment scheme will function as a concession rather than an obligation, and independent writing will continue to feed systems that answer with certainty, often incorrectly.

References

[1] IAB Tech Lab (2026) *LLM Content Ingest API Initiative.* New York: IAB Technology Laboratory.

[2] Chapekis, A. and Lieb, A. (2025) *Google users are less likely to click on links when an AI summary appears in the results.* Washington, DC: Pew Research Center, 22 July.

[3] Fortune (2026) 'Turns out Dead Internet Theory was right: AI agents are eating the Web, growing by nearly 8,000% and rewiring the Internet's business model', *Fortune*, 23 July.

[4] Belson, D. and Rhea, S. (2025) 'The crawl before the fall of referrals: understanding AI's impact on content

providers', *The Cloudflare Blog*, 1 July.

[5] Google (2024) 'Generative AI in Search: Let Google do the searching for you', *The Keyword*, 14 May.

[6] Mumbrella (2025) "'Why buy the cow?' IAB Tech Lab warns publishers over AI traffic losses', *Mumbrella*, 19 November.

[7] IAB Tech Lab (2025) *Dude, AI ate my traffic.* New York: IAB Technology Laboratory.

[8] Yang, Z. (2026) 'A wave of unexplained bot traffic is sweeping the web', *WIRED*, 12 February.

[9] Levinson, A. (2026) 'A surge of strange bot traffic from China has website owners alarmed. Here's what it means for your data', *Inc.*, 13 February.

[10] Breached.Company (2026) *We got hit by the mysterious Lanzhou bots: here's everything you need to fight back*, 17 February.

[11] Twenty-First Digital (2026) *Seeing a spike in direct traffic from Gansu, China? Here's what to do next.*

[12] European Broadcasting Union and BBC (2025) *News Integrity in AI Assistants: An International PSM Study.* Geneva: European Broadcasting Union.

[13] Newman, N., Ross Arguedas, A., Robertson, C.T., Nielsen, R.K. and Fletcher, R. (2025) *Reuters Institute Digital News Report 2025.* Oxford: Reuters Institute for the Study of Journalism.

[14] Jaźwińska, K. and Chandrasekar, A. (2025) 'AI search has a citation problem', *Columbia Journalism Review*, 6 March.

[15] Hagar, N., Agustianto, W. and Diakopoulos, N. (2025) *Not wrong, but untrue: LLM overconfidence in document-based queries.* arXiv:2509.25498.

[16] Woo, E. (2026) 'OpenAI burned $3.7 billion in first three months of 2026', *The Information.*

[17] NVIDIA (2025) *NVIDIA announces financial results for fourth quarter and fiscal 2025.* Santa Clara, CA: NVIDIA Corporation, 26 February.

[18] Kalai, A.T., Nachum, O., Vempala, S.S. and Zhang, E. (2025) *Why language models hallucinate.* arXiv:2509.04664.

[19] Duke University Libraries (2026) 'It's 2026. Why are LLMs still hallucinating?', *Duke University Libraries Blogs*, 5 January.

[20] OpenAI (2025) *Why language models hallucinate.* San Francisco, CA: OpenAI, 5 September.

[21] European Commission (2025) *Explanatory Notice and Template for the Public Summary of Training Content for General-Purpose AI Models.* Brussels: European Commission, 24 July.

[22] Cloudflare (2025a) 'Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives', *The Cloudflare Blog*, 4 August.

[23] Tobitt, C. (2026) 'Eight in ten of world's biggest news websites now block AI training bots', *Press Gazette*,

22 January.

[24] Maher, B. (2024) 'Revealed: Which of the top 100 UK and US news websites are blocking AI crawlers', *Press Gazette*, 27 February.

[25] Cloudflare (2025b) *Cloudflare just changed how AI crawlers scrape the internet-at-large; permission-based approach makes way for a new business model.* San Francisco, CA: Cloudflare, 1 July.

[26] Khan, H. and Asif, S. (2026) *Generative AI agents for controllable and protected content creation.* arXiv:2601.12348.

[27] DeepSeek-AI (2025) *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning.* arXiv:2501.12948.

[28] Reuters (2024) 'Exclusive: Reddit in AI content licensing deal with Google', *Reuters*, 22 February.

[29] OpenAI (2024) *OpenAI and Reddit partnership.* San Francisco, CA: OpenAI, 16 May.

[30] *Bartz v. Anthropic PBC* (2025) No. 3:24-cv-05417, United States District Court for the Northern District of California, class settlement.

[31] Digiday (2026) 'Google rolls out pay-per-value AI licensing program to publishers', *Digiday*, September.

[32] RSL Collective (2025) *Really Simple Licensing (RSL) standard.* RSL Collective, September.

[33] Internet Engineering Task Force (2025) *AI Preferences (aipref) Working Group Charter.* Fremont, CA: IETF.