

CRAWLING FOR COMPENSATION: MOVING TO A PARADIGM OF PUBLISHER REMUNERATION FOR CONTENT INGESTION

Introduction

In this document, our Data Analyst & Sustainability Lead, Dimitris Beis, shares his thoughts on the relationship between publishers and AI platforms, exploring models for remuneration and requirements for a viable solution.

Within the context of the digital advertising ecosystem, the rise of consumer-facing features and interfaces based on large language models (LLMs) has been consistently discussed as a threat to web publishers, a segment that is increasingly described as holding an unsteady position due to evolving market dynamics. The debate spans questions of intellectual property, transparency, media value, and the direction of the internet as a whole. This piece unpacks the issue and presents ideas on the requirements for a viable solution.

The problem at its core concerns the way people interact with online content. Due to the potential impact on publisher revenues, the mechanics of external referrals have long been a subject of industry debates, services, and even regulatory scrutiny. Issues have been raised over control of visibility, transparency, and dominance on the digital playing field. Concerns over publisher content appearing at the content aggregation level are nothing new - they have simply been exacerbated due to the unique capabilities that LLMs now provide: meeting users' needs at the aggregation level and making further browsing functionally inefficient or even redundant. What's changed is that the aggregation level can now offer a much more attractive value proposition to users, powered by a trove of training data, generative models, and agents that take care of browsing on their behalf.

Traffic-Level Dispute

The data shows external referrals are highly concentrated in terms of origin - most of them come from search and social media, with other aggregators trailing behind. Analysis beginning in the pre-chatbot era generally places the portion of total traffic coming from search at around 19 - 25%. Social media used to be the primary source of external referrals until it was overtaken by search around 2017.

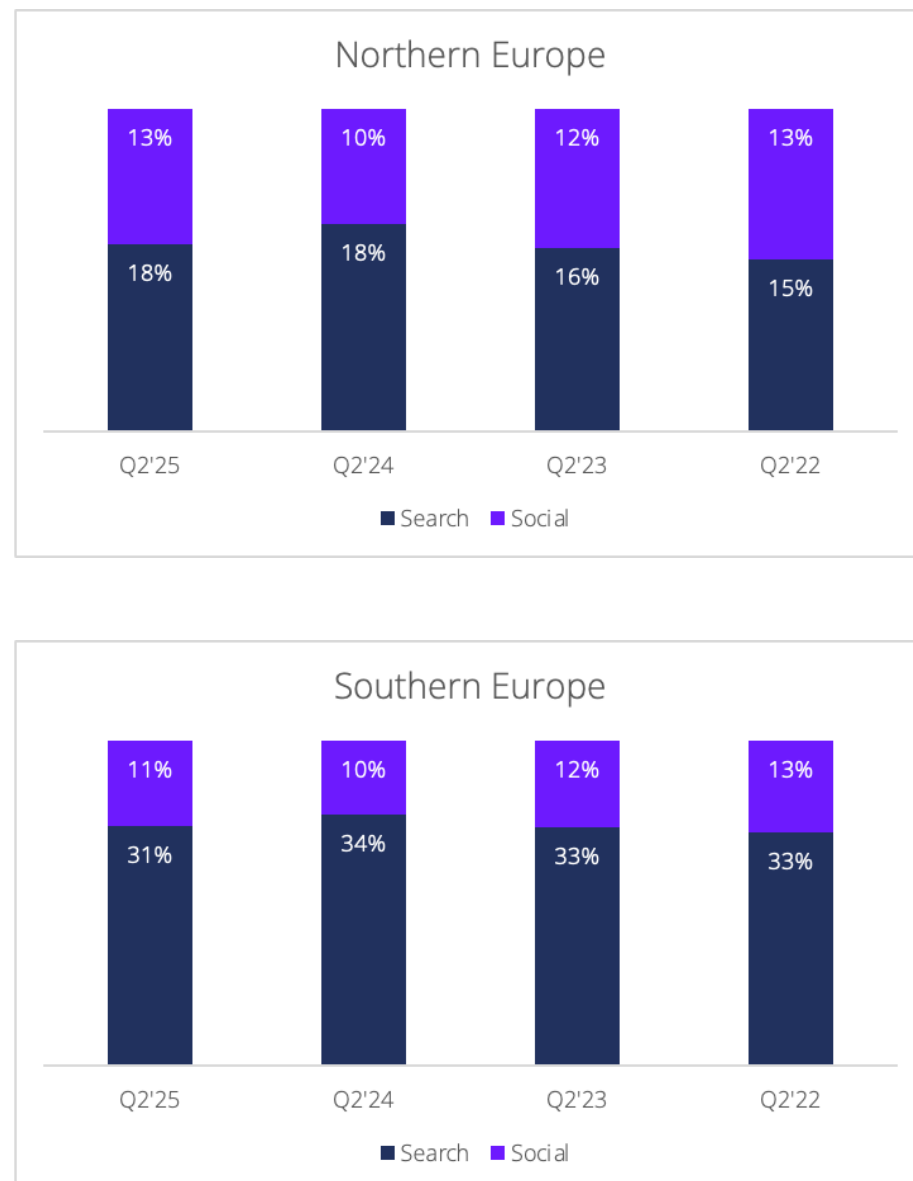
According to [Similarweb](#), referrals from AI platforms increased by 357% year-over-year in June 2025. The delta is impressive, but the overall figure of 1.13 billion visits is still far from the 191 billion visits from organic search on Google. Even though the news and media sector saw a 770% increase in traffic from AI platforms in the same period, the main concern remains whether innovations are eroding referrals from the far more consequential pool of search.

[Speaking](#) at an event in Cannes, Matthew Prince, CEO of Cloudflare, described a fundamental shift in search using the metric of pages crawled to visitors referred. He remarked that the ratio had increased from 2:1 a decade ago to 6:1 at the beginning of the year and 18:1 in June. He also claimed that the ratio of OpenAI, the company behind the chatbot estimated to have the most users worldwide, had increased from 250:1 to 1,250:1. These figures may hint at a change in what could be described as the content trade balance between publishers and referral sources - how much content is being accessed to enable navigation across the web and how many users are being referred to the publisher's website in exchange. Furthermore, [Similarweb](#) has reported that zero-click searches, i.e., searches that do not result in any referrals, grew from 56% to 69% year-on-year in May 2025. While these metrics have caused concern, they do not paint a full picture about referrals and publisher traffic, and whether the level of overall traffic from search has decreased is in doubt.

In August, [Google](#) directly rebutted the idea, reporting that referrals from organic search have actually been stable year-over-year and that reports suggesting declines in aggregate traffic are inaccurate and often based on flawed methodologies. Research, including aggregate data from 565 US and UK news websites, conducted by [Chartbeat](#), also found that search as a source of referral has been fairly consistent in the past year. Google did disclose that for certain types of questions (the example given being "when is the next full moon"), users may not click further, which has also been the case after the rollout of other features (e.g., sports scores) that give users information at the aggregation level.

The company's perspective is therefore that while zero-click searches may rise, publishers have not been harmed in terms of referrals from search overall.

Figure 1: Referral Sources: Search vs Social as % of Traffic in Southern and Northern Europe (Chartbeat)



Another important aspect is that the value of clicks may be changing. According to [Adobe](#) research conducted between July 2024 and February 2025, visitors referred from generative AI stayed on sites 8% longer, viewed 12% more pages, and had a 23% lower bounce rate. However, they still lagged behind non-AI-referred users by 9% in terms of conversion rate. Google also claims that click quality (defined in terms of bounce rate) has slightly increased.

Overall, looking at referrals would suggest that publisher traffic is much more sensitive to changes in search than to the rise of AI platforms. Some data shows search referrals to be stable, but many

individual publishers are reporting drops, and more are expressing concern about their position. Their leverage is their content, and there cannot be any indexing without content access. Accepting the use of their content in the generation of answers comes hand-in-hand with accepting crawling for indexing purposes, and removing themselves from search results is not an option.

Value-Level Dispute

Referrals alone do not tell a complete story. The main issue is value and whether it is fairly exchanged between publishers and LLMs accessing their content. Even if referrals from AI platforms rise again by the same amount in the next year, publishers could still be getting a deal that many would describe as unfair.

Publishers are able to extract a certain amount of value from a piece of original content through ad revenue or otherwise. AI platforms can also extract value from this content, and it may or may not be additional. On the one hand, an LLM response can offer users more value than the sum of the value individually extracted from the content utilised to respond, and AI platforms could generate value added. On the other hand, a shift in user behaviour towards conversational interfaces could lead to publishers being able to extract less and less value from their content, making the value taken a more accurate description. Two different problematic effects present themselves:

- A. AI platforms extract value without paying publishers, even if publishers retain their existing revenues.
- B. AI platforms reduce publishers' ability to extract value by substituting visits with conversational answers.

The solution to both is remunerating publishers for content, and different paradigms have already been put forward.

IAB Tech Lab Framework

IAB Tech Lab recently published [guidance](#) on the matter in the form of the LLMs and AI Agents Integration Framework. According to the document, the model of content access deals between publishers and LLM providers (e.g., the recent deal between the New York Times and Amazon) is not sufficient as it does not rely on market mechanisms (e.g., bidding) to discover prices and may not work for smaller publishers with less content and capacity to support such relationships. As such, the framework proposes three steps for publishers:

1. Prevent bots and scrapers from accessing content using robots.txt or Web Application

Firewall (WAF) methods. Robots.txt is a file placed at the root of a website to declare which content crawlers are permitted to access it. As not all scrapers respect robots.txt (the document mentions a 40% increase in unauthorised scraping from Q3 to Q4 last year), WAF methods are implemented as a stricter way to prevent bots from accessing content. CAPTCHA is a familiar example. Other methods rely on the detection of abnormal behavioural patterns (e.g., rapid navigation) or signals such as location, IP address, or user agent. Bot detection can resemble a game of cat-and-mouse where deployers of bots are constantly seeking new ways to spoof real users and bypass controls.

2. Develop and deploy mechanisms that enable AI agents to easily discover and understand content.

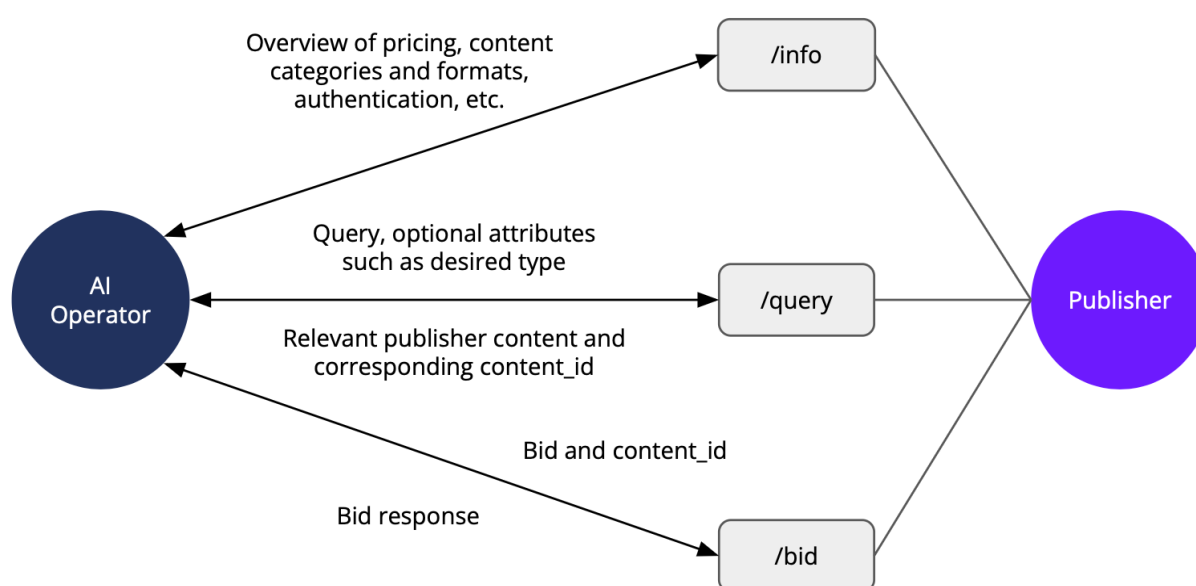
The Tech Lab framework features three components:

- A. A content access rules page containing the terms of use governing content access, instructions for scrapers, contact information, legal notices, and content metadata,
- B. a content metadata json containing useful information about the content such as a summary of the site or a mapping against the IAB content taxonomy, and/or
- C. an llms.txt markdown file that contains information, guidance, and links to other detailed markdown files and is easily digestible by LLMs. Examples can be found in [this directory](#) of sites that have adopted the standard.

3. Monetise content using an appropriate model. Content partnerships, covered above, usually rely on proprietary APIs. The simplest option would be introducing a static cost per crawl, which can be implemented through a third-party platform. Many platforms enable the adoption of more complicated models where the cost per crawl is a function of parameters such as how premium the content is (e.g. evergreen listicle vs. investigative piece). Tech Lab outlines a Cost-per-crawl (CPCr) API that features tiered pricing sensitive to content type, bot type (e.g. academic vs. commercial), or frequency, and supports subscriptions. Another option is the **LLM ingest content API** that supports per-query pricing based on demand through a bid / bid response exchange.

In an isolated form, retrieval-augmented generation (RAG) works by having generation include material that is drawn from a collection of embedded documents. This relies on querying the collection based on user prompts. The live RAG that is available on AI platforms is based on agents querying the web instead of a managed collection of documents. The idea behind the per-query model is that by sending the query directly to the publisher, it more closely tracks value extracted from using publisher content and facilitates a fairer deal than cost-per-crawl. The API is a blueprint for enabling per-query pricing, and the documentation covers authentication, discovery of information about the API (e.g., pricing for different content types), retrieving relevant content based on a query, bidding for exclusive or premium content, logging, and billing.

Figure 2: Some Endpoints of the Publisher LLM ingest API



Apart from being a welcome and well-received proposal in a space with clear gaps in terms of technical specifications, the Framework is also incredible food for thought - Tech Lab is clear on the fact that aspects like pricing models will evolve in the market, but has put forward a system to unpack and iterate upon.

Observations

1. Controlling content access effectively is a clear prerequisite for the ingest API to work.

Otherwise, AI operators may be incentivised to scrape content and bypass the API. Most large publishers already implement robots.txt and some WAF methods to limit bot traffic, as it is costly regardless of whether the content is scraped and how it is used. As aforementioned, these methods are imperfect and would likely need to be complemented by a commitment from AI operators to go through the API. Many AI companies state that they respect robots.txt, but training data is often not made public, and multiple investigations have suggested otherwise. This problem is at the root of publishers' concerns and has to be solved independently (and likely on a regulatory level) to any conversation around what fills the gap.

2. Auction dynamics will differ from ad space.

In advertising, auctions involve many buyers competing for exclusive, time-sensitive slots. Content access differs: auctions usually involve only a single AI operator, access is not inherently exclusive, and the publisher does not control when requests occur. In this setting, an AI operator's rational strategy could be to find the publisher's bid floor by offering incrementally higher bids. To prevent this dynamic, throttling or sealed-bid mechanisms would need to be introduced so that multiple bids cannot be trialled per query.

Licensing also raises verification challenges. The value of content depends heavily on how it is used: incorporation into a long-term training set, for instance, is worth more than a single retrieval for a live query. Once access is granted, however, publishers have limited visibility into actual use. This makes trust-based solutions ineffective. A zero-trust approach would be needed to ensure that content licensed for one purpose is not repurposed for another.

3. Valuation will be complex.

What is the marginal benefit of an additional piece of content being used to generate a response from an LLM? Furthermore, how is the marginal benefit determined without actually accessing the content?

On the publisher side, setting price floors might appear to be simple. However, bid floor algorithms on the publisher side would have to internalise the potential trade-off between revenue from content ingestion and revenue from website traffic. Publishers would need more information to estimate how much revenue from other sources they are parting with by enabling content access.

Pricing will also be complicated on the AI operator side, even if there is infrastructure that enables differentiating price depending on how the content will be used, as the level of traceability differs between the two aforementioned use cases (training data and live RAG). It will be harder for AI companies to determine the value of content additions to the training corpus at a granular level, and the feedback loop will be longer.

Finally, the process is going to be probabilistic if based solely on metadata. An article (in its plain or tokenised form) may contain all the information required to answer a user's question, or it may be a dud - pricing decisions will have to take that into account. An alternative would be to allow models to verify whether content is useful before purchasing the right to use it, effectively splitting access and licensing into two steps. However, that model again introduces a requirement for a zero-trust solution to verify that access used for generating bids is not abused.

Other Models

Alongside the Tech Lab framework, several other models have been put forward to channel remuneration back to publishers:

Revenue-sharing subscriptions.

AI company Perplexity offers a subscription service where 80% of user fees are distributed to participating publishers. Allocation is based on engagement signals such as clicks, citations, and usage of content in responses. The model attempts to align publisher compensation with actual audience demand on the AI platform.

Bilateral licensing agreements.

High-profile deals between publishers and AI platforms represent a direct negotiation model. These agreements typically involve large upfront payments or structured revenue shares, but because they are negotiated one-to-one, they concentrate benefits among large publishers with the leverage to secure favourable terms.

Collective or statutory licensing.

Borrowing from the precedent of music rights societies and broadcast royalties, some policymakers and industry voices have proposed collective schemes. AI platforms would pay into a central pool under a compulsory licence, and funds would be distributed to publishers according to measured or estimated usage. This model promises universality and scale, but requires regulatory action and consensus on allocation.

Each of these paradigms approaches the same problem from a different angle, but none is yet dominant. Together, they may signal that remuneration is unlikely to converge on a single mechanism, and publishers should anticipate a patchwork of models depending on market position and jurisdiction.

Requirements

To close the gap between the value extracted from publisher content and the current lack of remuneration, any viable usage-based model would need to meet three conditions:

Effective and verifiable content access control.

Publishers must be able to reliably block unauthorised scraping and enforce terms of access.

Assurance of purpose-limited use.

Content licensed for a single query should not be stored or repurposed for training. Without this safeguard, per-query APIs could be exploited as a de facto cost-per-crawl system.

Transparency in pricing and trade-offs.

Publishers require visibility into how their content is being used and valued in order to weigh short-term ingestion revenue against potential long-term losses in traffic. If AI platforms control this information unilaterally, market dynamics will skew in their favour.

These requirements are not currently satisfied. Unauthorised scraping continues to rise and remains the root cause of publisher concern. Most publishers lack visibility into how their content is used once accessed, and as Tech Lab has noted, only the very largest can secure protections through bespoke agreements with AI operators.

[Cloudflare](#) has recently introduced the ability to block AI crawlers and is piloting systems in which AI platforms declare the purpose of content access while publishers control permissions. The company is also working on technologies such as signed requests and mTLS to strengthen crawler identification. Such measures move the industry in the right direction, but ultimately the requirements outlined above may need to be mandated. Tony Katsur, CEO of IAB Tech Lab, has [argued](#) for regulatory intervention and urged publishers to advocate for their interests.

Summary

The challenge of publisher content in the age of generative AI reflects two intertwined issues: long-standing disputes over how content is surfaced at the aggregation layer, and the absence of a clear framework governing AI access and use. The debate centres both on whether search traffic is in decline, a point contested by recent aggregate data, and on the lack of remuneration when AI platforms ingest and use publisher content in generation. Multiple paradigms are being tested to address the revenue gap, but structural solutions that enforce access control, transparency, and verifiable usage are likely to be prerequisites before any remuneration model can function at scale.


IAB Europe's Artificial Intelligence Working Group is looking for European publishers to collaborate with. To learn more about the Working Group and how you can get involved, please reach out to Dimitris Beis at [beis \[at\] iabeurope \[dot\] eu](mailto:beis@iab europe.eu).


Dimitris Beis

Data Analyst & Sustainability Lead

beis@iabeurope.eu

iab europe
Rond-Point Robert
Schumanplein 11
1040 Brussels
Belgium

 [@iabeurope](https://twitter.com/iabeurope)

 [iab-europe](https://www.linkedin.com/company/iab-europe/)

iabeurope.eu

