ENDERS | ANALYSIS

# AI, press and licensing

# Deals for the chosen few

- Big news publishers are pursuing licensing deals with AI companies, chiefly OpenAI. Not all publishers will see a substantial return; while some news may be important for training AI models, not all publisher content will be
- Litigation is a threat point when negotiations stall (see the New York Times), but the copyright status of Large Language Models (LLMs) is uncertain. In the UK, there has been no government intervention (on copyright or otherwise) that could facilitate licensing
- Publishers' bargaining position is strongest when it comes to up-to-date material that could be important in powering some AI consumer products. They should seek deals to support their journalism, while bearing in mind the risk that new products may get between them and their readers

**Niamh Burns**
niamh.burns@endersanalysis.com

**Abi Watson**
abi.watson@endersanalysis.com

**Jamie MacEwan**
jamie.macewan@endersanalysis.com

**Joseph Teasdale**
joseph.teasdale@endersanalysis.com

**Alice Enders**
alice.enders@endersanalysis.com

+44 (0)20 7851 0900

22 April 2024



**Related reports:**

Who wins from the AI boom: A dive into the value chain [2024-032]

Rebalancing terms for news on platforms: UK regime by 2025 [2023-086]

AI in the media: Rise of the machines [2023-044]

AI and search: Microsoft looks to shake up a 175 billion market [2023-029]

News publishers, Google and Facebook: An overview of interactions [2022-068]

# Executive summary

The first licensing deals are being struck between OpenAI and select press publishers, fuelling ChatGPT with up-to-date, accurate news content in a range of territories and languages. The terms of deals are confidential, and negotiations are without direct precedent. Google has significant and longstanding publisher relationships, and is also seeking content to feed its models, chatbots and enhanced search. Publishers have the greatest leverage around ongoing access to current material, augmenting LLM products so that they can respond to informational questions accurately, rather than training models on archives. Negotiations hinge on a number of unanswered legal and practical questions, as LLMs' status and their relationship to training data and outputs are all contested—now is the time to push for influence.

The marginal contribution of any one publisher to the content soup used to train LLMs will be tiny, even if publisher content *in general* is important for high-quality LLM output. When it comes to information at scale, the elephant in the room is Wikipedia, a freely available substitute for factual content. The role of news depends on the product; generative search needs news inputs more than a social media chatbot. Not all publishers will be essential even where *some* news is—there is a reason the AP was the first to land an OpenAI deal, with its timely, accurate reporting at scale. Small publishers could be left out.

The high-profile New York Times (NYT) case against OpenAI and Microsoft was triggered by the NYT when commercial licensing negotiations came to a halt. On the copyright side, the NYT's claims are *relatively* contained: a proprietary model was trained, owned and deployed by one set of defendants, and the NYT's global scale and brand strength means its place in training sets is particularly prominent. Even so, determining *where* copyright infringement might occur (in training, the model itself, or outputs, including the model's commercial deployment in search) is a minefield. In the US, the 'fair use' defence could mean courts side with the AI developers, leaving the NYT with a bill for their legal costs (and its own). High stakes on both sides raise the likelihood that the case ends with a confidential licensing deal following an out-of-court settlement. Launching legal action is not a threat point available to every publisher seeking a deal, especially when the costs are so clear and the benefits so unclear.

In the UK, there isn't the same 'fair use' barrier to publisher success in legal action, though cases around LLMs are similarly unprecedented. If legal action is necessary to assert that AI needs a licence, that is itself a significant cost barrier. Government plans for regime change on AI training and copyright have stalled, leaving publishers without the prospect of a legal backstop that would require transparency over training data, or greater publisher controls over scraping (as the EU is working towards). Even if publishers can opt out of their content being used for training new models, they cannot currently opt out of contributing to generative search responses without also opting out of being included in search results altogether.

We understand OpenAI is in talks with some UK publishers. There is a precarity to time-limited dealmaking without any legal backstop. In general, AI companies frame publisher interaction as *voluntary* sectoral support. Publishers' history with Meta offers a cautionary tale: voluntary news payments can be halted, and audiences mediated by a third party are easily lost. At the same time, there is a risk that making extensive IP licensing a prerequisite for commercially launching an LLM could end up reinforcing the existing tech oligopoly, which doesn't ultimately benefit publishers.

Licensing revenue might be material for publishers, but it is unlikely to be commercially transformative. Long-term demand effects could be more significant, and must be borne in mind when publishers are designing their approach to deals. Many news publishers in the UK have a substantial reliance on Google Search for discovery by users, while others operate subscription-based direct reader relationships. Google has just started generative search trials in the UK; could a more advanced version adversely affect publisher traffic? It is impossible to know right now, but if internet users become accustomed to using AI products to access journalistic content (or content like reviews and recommendations), it could severely weaken publisher-reader relationships. They will have to adapt their commercial model to a changing online economy; AI is sharpening existing challenges rather than creating new categories of problems.

# Despite legal and practical uncertainty, publishers are already negotiating commercial deals

Select news businesses have already made licensing deals—selling authorised access to their content archives or ongoing access to their output. News businesses' revenue from content licensing could increase (and already offers a flashy immediate-term bump for other online media businesses: Reddit pointed to a deal with Google worth $60 million per year in advance of its IPO). Many UK publishers are in talks. The amounts involved are unlikely to be game-changing for most publishers, and the benefits won't be spread equally across the industry. There are also some longer-term risks to publishers if AI *products* can substitute for those of news providers.

Publishers are considering the following issues:

- How should publishers go about valuing their material? What value lies in 1) archival material and 2) ongoing access to new material? This will vary by publisher, and some could be left behind
    - For model pre-training, individual publisher content might not be worth much to LLM providers. However, there is less downside risk to publishers licensing archival content that receives little traffic. Once a publisher has allowed its content to be part of model pre-training, this cannot be straightforwardly withdrawn after the model is trained. For current models, LLMs have in practice already been trained on publisher activities by the time the deal is made. For training new models, publishers could have more control
    - Access to content for the ongoing 'grounding' of models with up-to-date material[1] is where far more value lies for LLM providers, and concerns content that is actively monetised by publishers on their own sites
    - Quality, original journalism matters here—the kind that requires dedicated, well-funded human newsrooms. There is a reason the AP was the first major publisher to make a deal: it produces timely, accurate reporting at scale. Reuters is another good candidate. Publishers cannot be complacent about the quality of their product: lower-quality or unoriginal output will be worth much less to AI providers (just as it will be much less distinguished to consumers as AI-generated content proliferates online)
    - Opinion content, even where it is high-quality, may be less valuable for licensing than factual hard news—the value of newspaper opinion having already been diluted by the rise of digital and social media, which makes this kind of content very easy to source
- How are UK publishers positioned for dealmaking?
    - A number of UK national publishers are in licensing talks with AI companies—the UK has a range of high-quality English-language titles with international reach, brand recognition and coverage, though those titles will have to compete with international ones for licences with global AI companies
    - Local content will not have the same value to LLM providers as quality national content, but depending on the product, some local informational coverage could be of use (e.g. in search integration). Consolidation in local press has resulted in Newsquest, National World and Reach plc having an approximately 70% market share by number of titles, which puts them in a slightly better negotiating position in the absence of formal collective licensing mechanisms

[1] Once deployed, a model might also be 'grounded' with material that wasn't part of its original training set, like a search LLM having ongoing access to websites at the point of user query, so that its output is relevant and up-to-date.

    - Paywalled content might be at an advantage: there is a risk that AI developers view much of the open internet as fair game, though many publishers have required 'rights and permissions' to use their content for years, even if it is not paywalled

- Should these deals be exclusive?
    - The terms of existing deals are unknown, but the more modest size of reported payout offers beyond the highest-profile deals suggest many are *not*
    - Exclusive deals would be a way for highly competitive AI companies to differentiate their offerings, if a given data source gives them extra capabilities or provides a level of brand trust
- How much of an early mover advantage is there among publishers?
    - AI developers might not need *all* publishers' content to ensure quality outputs, and a select group of global news brands making early deals could mean that latecomers do not have the same leverage
    - The AP reportedly built in a most-favoured nation clause into its contract with OpenAI, meaning the terms can be reassessed if another news brand got a better deal[2]
- What controls, attribution or limitations should publishers push for in licensing contracts?
    - It's hard to say what AI products will even look like a year from now, so some caution in the licensing terms is appropriate. Publishers are looking to share in the value they create, but what that value will be is still totally uncertain. Publishers must minimise the risk that the products they are contributing to substitute for their own
    - Attribution is a tricky issue: publishers want to be cited (and linked) where their journalism was used to generate outputs, but there is brand risk where AI content is attributed to a publisher which does not control the user environment, especially as LLMs continue to hallucinate or to mash together material based on different sources. Attribution even on accurate outputs could backfire, if users seeing that e.g. the BBC contributed to a generative search response reassures them about its accuracy and negates the need to visit the BBC's site
- What is the wider risk to pursuing licensing deals?
    - Publishers collectively hold many cards at this moment, which should not be given away in a hurry despite their exact value being difficult to estimate. At a time where the financial benefits of licensing deals are equivalent to a small revenue boost to most, careful thought should be given to terms that fit in with the prevailing commercial goal of sustaining quality journalism and content creation online
    - Early deals could legitimise a status quo largely driven by big tech companies, and bed in the wider perceived acceptance of the largest models' modus operandi. Seeking recourse later on could be complicated by having to show why prior terms are no longer acceptable or have been contravened, in spirit if not in letter. Arguably having an opt-out rather than an opt-in system for allowing data to be scraped to train models is already ground given up—concessions made in the realm of access to up-to-date information would be costlier
    - Publishers need to consider what licensing means in the long term: it will provide some incremental income in the short term, but if users adapt how they fulfil their informational needs and engage with journalistic content through new AI-powered intermediaries, this could ultimately cannibalise direct engagement with news titles. Of course, this could happen whether or not publishers engage with LLM-providers

[2] The Wall Street Journal, As Publishers Seek AI Payments, AP Gets a First-Mover Safeguard, 28 July 2023.

    - Search's longstanding position as a key source of traffic is by no means guaranteed in the long run given that, for many queries, a more useful chatbot will be one that saves the user the time of visiting websites
    - The risks are particularly strong in the case of affiliate marketing (see below), though a similar principle applies for wider news and other output and commercial models. There are many instances where AI products could substitute for original content, and where their widespread take-up by users would ultimately undermine the commercial model underlying the content that powers the LLM

## Publishers must act now to stake their claims on over-the-horizon outcomes

There is a wide scope for negotiating terms to govern how publisher information is used and rewarded by novel generative search products, and other as-yet unreleased products. The risks outlined in the previous section, along with the relatively small sums on offer for many publishers, suggest there is room to push for favourable terms in negotiations with a keen eye on the future, including specific products that rely on publishers' up-to-date content.

- Concessions that publishers should be pushing to secure include:
    - **Contractual rights**, such as mechanisms to opt out of deals or licensing arrangements at short notice if there is evidence of aspects of the deal being contravened
    - **A stronger shield against cloning of publisher content and made-for-advertising (MFA) sites**: this is an existing problem that is already being exacerbated by AI products
    - **Reliable minimum guarantees**, whether this is based on regular royalties (effectively a kind of annuity split across the industry) or minimum referral traffic
    - **Licensing that adapts to the market**, such as licensing to specific products powered by a model as they reach a certain size and revenue profile—with data supplied at an established cadence to track and audit this growth
    - **Transparency on performance**, for instance the proportion of queries that refer to a publisher's content, average position in the search results, click-through rates, topic segmentation, user segmentation (e.g. free tier, logged in, Pro), and comparisons to the average where this does not expose competitors' commercially-sensitive information
    - **Country-level summaries of publishers' share of queries**, broken down by relevant category such as informational and commercial, to allow the industry at large to audit its value and what it is getting back, much as the WGA (the Writers Guild of America) negotiated to instate a form of residuals on streaming revenues for top shows based on confidential data shared back to representatives of the Guild
- Other outcomes are more mixed from a publisher perspective:
    - **Direct product benefits from deals**, such as first-look at new AI tools, may seem attractive, but many publishers are uncomfortable with this sort of bundling, and with Google's development of a suite of newsroom tools
    - **The role of collective bargaining**, in the absence of a regulatory backstop, is unresolved. While it could be a powerful force for obtaining concessions, it could also lead to an unsatisfactory minimum where larger publishers rely on negotiating bespoke deals to augment their terms on the side, given an *exclusive* regime is undesirable. It could also require a carve-out from competition regimes
- **The ability of publishers to negotiate as one is limited in practice.** For one, a small publisher that makes most of its money from programmatic advertising has a different set of incentives to a large news site with a destination app and subscription model. Some of the terms would be difficult to organise and enforce for a smaller publisher given the absence of an equivalent to the Guild. We

consider some initiatives to bring together industry and make sure smaller publishers—and newer or smaller AI model owners—are not left out of the licensing process at the end of this report

## OpenAI acts early on publisher deals, while big tech companies struggle to fit AI into existing publisher relationships

Despite the high-profile NYT case (see below), **OpenAI** has been working to publicly position itself as a partner to publishers. Its response to the NYT lawsuit emphasised that it considers its training activities fair use, but that it has started to allow publishers to opt out of scraping for new models "because it's the right thing to do". It has launched into sectoral support schemes, including a deal with American Journalism Project to support local news—giving $5m in direct funding and the same again in OpenAI API credits—and a grant to NYU to fund a journalism ethics initiative.

On the commercial side, OpenAI made an early deal with the AP, and subsequently struck deals with European publishers Axel Springer, Le Monde and Prisa Media (Figure 1)—having a spread of international publishers allows its products to serve a range of territories, languages and cultures. It is in talks with a range of international publishers and UK nationals. More deals are likely soon: News Corp's Robert Thomson publicly hinted at such an imminent deal in the latest earnings call, directly praising OpenAI and Sam Altman and contrasting his company's strategy with the NYT's: "Courtship is preferable to courtrooms. We are wooing, not suing". Some reporting has placed offers to publishers at $1-5 million per year.[3]

### Figure 1: OpenAI publisher deals

**July 2023:** OpenAI licenses "part of AP's text archive", while AP uses OpenAI "technology and product expertise". Reportedly includes a most-favoured nation clause

**December 2023:** OpenAI deal with Axel Springer worth "tens of millions of dollars over several years" according to execs interviewed anonymously by The Information

Before the deal was announced, Axel Springer CEO Mathias Döpfner said his first choice would be a quantitative model like that used in music industry; an annual agreement for general use would be second best

**March 2024:** Deals signed with Le Monde and Prisa Media (owner of titles including El País) "to bring French and Spanish news content to ChatGPT". Covers model training and access to ongoing coverage for ChatGPT

[Source: Enders Analysis, company press releases, media reports]

Part of the issue for both sides is that the pressure to conclude licensing deals has come before it is clear what the AI products will ultimately look like: in the case of OpenAI, its first consumer product (ChatGPT) is more of a public beta test of the company's tech. OpenAI has entered the field relatively recently, but

[3] The Information, OpenAI Offers Publishers as Little as $1 Million a Year, 4 January 2024.

incumbent tech companies (like Google and Apple) are now having to work AI into their existing publisher relationships at a time when some (though not all) of those relationships are showing signs of strain.

**Google** has the longest standing publisher relationships, with bespoke deals in many territories that *could* be tweaked to include AI licensing provisions. Google will be keen to bundle AI licensing with wider deals involving publishers' use of Google tech, rather than having to contend with higher publisher payouts than it is already facing under various regulatory regimes for the 'use' of publisher content. Google has been working with smaller publishers to develop AI tools as part of the Google News Initiative[4], and has trialled Gemini newsroom tech with larger publishers; some prominent publisher deals (like the NYT's) involve the use of a lot of Google AI tools.[5] However, some publishers will be reluctant to accept this bundling, which they worry could take the place of material licensing revenues—and many are uncomfortable with Google's development of a suite of newsroom tools.

In the case of both Google and Microsoft/OpenAI, publishers are rightly apprehensive about trials of generative AI-enhanced search products: anything that reduces the prominence of links and makes search more of a destination in itself, rather than a discovery portal based on click-through traffic. For Microsoft, there is more upside to disrupting the existing search model, because any share it can gain at Google's expense is worth a great deal, even if it is less monetisable or lower-margin than the existing model. Google has more of an incentive to preserve something closer to the link economy status quo. As the far more significant search player, Google's decisions around generative search will make much more of a difference to publishers: it has launched this in a number of territories including the US, and is testing it in the UK.

**Meta** will not be focusing on news deals—it is pulling back from news more than ever, declining to extend existing commercial news deals and closing down Facebook's News tab in more and more countries. News is never going to be a central part of its AI consumer products, the first examples of which have been conversational chatbots trialled in its social apps, whose appeal is not informational. Meta will also rely much more on its own massive bank of user-generated content in training, and has so far made more of the idea of 'publicly available' training datasets. Social media rival **TikTok** is reported to have begun testing AI-powered virtual influencers.

**Apple** appears to be behind on developing its own LLMs. There are recent reports that Apple has been in talks with potential partners including Google to use their cloud-delivered generative AI tech on iPhones (which would mean less direct publisher relationships on this issue). Apple's own AI would be limited to on-device functionality at this stage—more will be revealed at WWDC in June. In the long run, expectations are high for Apple's ability to exploit AI across its devices. NYT reporting on Apple's direct offers to publishers included terms that publishers considered "too expansive"—leaving the potential uses of publisher content too open, and asking publishers to assume wider content liability. Nonetheless, it was reportedly nearing a number of "multiyear deals worth at least $50 million".[6]

Though publishers are looking to these larger companies for licensing opportunities now, there is scope for **new entrants** to emerge.

These negotiations will interact with several open questions around how the LLM market develops:

- Will licensing costs be a barrier to entry to the LLM market?
- Will guardrails to prevent problematic output (including misattribution and copyright) also make products less generally useful, for example through LLMs refusing to answer questions they otherwise would?

[4] Adweek, Google Is Paying Publishers to Test an Unreleased Gen AI Platform, 27 February 2024.
[5] The New York Times Company and Google Expand Agreement on News and Innovation, 6 February 2023.
[6] NYT, Apple Explores A.I. Deals With News Publishers, 22 December 2023.

- Will smaller models with more contained inputs become more important? This could include "ethical LLMs" whose models are trained exclusively on licensed content. If so, it would be an opportunity for select publishers who position themselves correctly
- Will regulators such as the CMA and the European Commission encourage a plurality of model providers, such as by obliging designated gatekeepers to give users a choice of LLM at critical junctions?

# A legislative backstop for licensing negotiations is not currently on the horizon in the UK

The UK government does not have any concrete plans to introduce horizontal AI legislation (like the EU's AI Act).[7] It has more or less maintained the initial 'pro-innovation' stance it set out in early 2023, designed to encourage AI investment, though it has shifted its messaging numerous times, especially on the issue of copyright (see Figure 2). This included proposing an extension of the existing text and data mining copyright exception to allow commercial uses, which was then dropped. This would have ensured that use of copyrighted content to train LLMs is not copyright infringement. After dropping this, the government held talks between key stakeholders in the hopes of establishing a voluntary code on copyright and AI. These talks failed, so it is now exploring other mechanisms for enabling licensing deals.

### Figure 2: Timeline of UK government policy proposals on copyright/AI

| | |
|---|---|
| 2014 | Text and data mining (TDM) exception for non-commercial research introduced |
| 2022 | Government Intellectual Property Office (IPO) proposes extending this to commercial uses to promote AI innovation |
| February 2023 | Government drops these proposals |
| June 2023 | IPO begins consulting stakeholders on a voluntary code on copyright and AI |
| February 2024 | IPO talks on a voluntary code break down<br><br>Government publishes outcome of AI consultation signalling no immediate intention to legislate AI specifically<br><br>States it will "now lead a period of engagement with the AI and rights holder sectors", with a focus on "transparency" for inputs and "attribution" for outputs |

[Source: Enders Analysis]

As a starting point, the government has indicated it will prioritise transparency and attribution in developing proposals. This would give publishers more certainty and evidence around the use of their content to train models, which could strengthen their position in any negotiations or litigation.

Copyright law clarification of some kind is possible, though neither the current government nor the Labour Party have indicated this is likely any time soon. On the important question of allowing continued access to content for LLM products, potential legislative intervention is not straightforward, but

[7] Recent reports suggest the government has begun to craft new legislation, with a focus on the largest model owners sharing their algorithms and showing evidence of safety testing, though plans are said to still be at an early stage—see: Financial Times, UK rethinks AI legislation as alarm grows over potential risks, 15 April 2024.

publishers will at least want to secure controls over scraping, including separation by purpose, beyond the limited ones Google and OpenAI have made available.

Some publishers have called for the CMA's Digital Markets Unit (DMU) to review these issues as part of the new competition regime, which will establish a bargaining regime for commercial deals on publisher content appearing on social platforms and in search. The overarching structure of the legislation is that it will designate firms with 'Strategic Market Status' in various digital activities. This makes it a tricky fit for AI issues. If the relevant digital activity is search, for example, it's not even clear whether Microsoft's Bing is big enough to fall under the legislation's scope (it doesn't fall under the equivalent European regulation). The DMU might at least consider AI training and grounding in the context of ensuring good faith negotiations between publishers and platforms (as the French competition regulator has, see below).

The CMA would also need to consider the effects of any intervention on the emerging market for LLMs and foundation models, where its initial review emphasised the importance of access to the market. Making extensive licensing payments a condition of entry to this market could serve to entrench the advantage of existing big tech players. Additionally, publishers will be keen to avoid any system that reduces their freedom to negotiate terms—they won't want any legal backstop that would in effect make licensing compulsory. Still, some regulatory backstop could be necessary to facilitate collective licensing, if smaller publishers are less able to secure deals.

# The New York Times battles with Open AI

**The NYT lawsuit against OpenAI and Microsoft is an early high-profile case, and a special case.** From a legal perspective, it is *relatively* contained compared to other publisher/developer cases: the same providers (OpenAI, in partnership with Microsoft) trained, own and operate the model, in the US. The NYT ranks above other publishers in key training sets, making it easier to point to its role in training. A court ruling in the NYT's favour might be a helpful precedent only to a handful of similarly well-positioned publishers. In any case, the fair use defence is still a powerful obstacle.

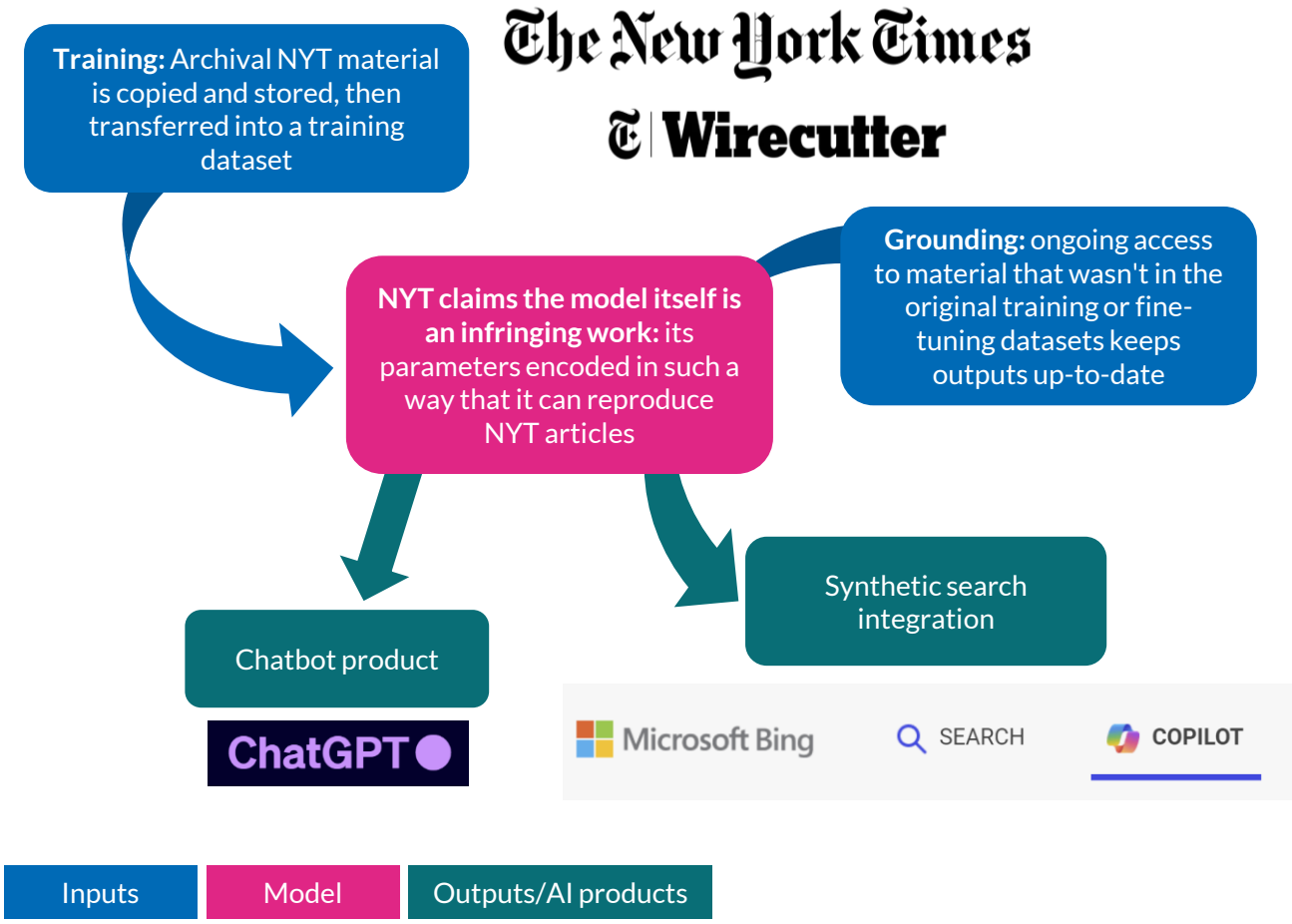
The NYT was in commercial negotiations until just before filing this case, and in filing the case the NYT has activated a threat point as a way of maximising its commercial deal (as well as taking a strong stand against the possibility of damage to its brand). The stakes are high on both sides, increasing the likelihood of a settlement (meaning a time-bound commercial licensing agreement rather than a perpetual legal ruling). On the one hand, a defeated OpenAI would face more pressure to pay out to rightsholders, and at the extreme end it could be ordered to destroy its models; on the other hand, the NYT would find its leverage in licensing negotiations vastly diminished if a judge accepted a fair use defence at any of the key stages of alleged infringement.

The NYT alleges various infringements at the levels of *inputs* and *model training*, the *model itself*, and *outputs* via Bing search and ChatGPT (see Figure 3):

- On inputs, the NYT alleges that OpenAI removed copyright-management information when copying NYT content (otherwise, it claims, such material would be reflected in outputs)
- The NYT points to the use of web scrapers that were enabled for search products to feed up-to-date material to AI products
- The NYT calls the model an "embodiment" of "many unauthorized copies or derivatives" of copyrighted work. The copyrighted works' "expressive contents have been substantially encoded within the parameters of the GPT series of LLMs". This cannot be verified by inspecting the model itself—instead, the *outputs* are the evidence
- The NYT states that outputs include **"unauthorised copies"** of NYT articles (generated text that is near identical to NYT material, outputted in response to user prompts, including those requesting specifically to get behind the paywall)
    - It claims this has a **"substitutive use"**, not justified by "transformative purpose"
    - Links are less prominent than traditional search, and less likely to be used
- Outputs also include **"inaccurate forgeries"** (i.e. citing non-existent articles) that damage its brand

The questions of output and model are interlinked. On outputs, the NYT filing contains numerous examples of verbatim copying—including cases where large portions of articles are shown in response to prompts asking to get behind the paywall. But the *scale* of distribution is what determines how damaging this is, and it seems a relatively easy fix to reduce that scale: OpenAI points out that these sorts of outputs *cannot* be easily reproduced. The infringing outputs the NYT submitted might then be viewed more as evidence that the model itself is infringing. But the scale of outputs also affects consideration of the model: how reliably does the model actually reproduce this content? See below for more detail on the copyright question.

## Figure 3: The New York Times copyright infringment claims against OpenAI and Microsoft




[Source: Enders Analysis, NYT legal filing. Images: NYT, Microsoft, OpenAI]

### Killing affiliate: AI's commercial threat to Wirecutter could extend to UK publishers

The commercial threat of AI-generated search responses that summarise publisher content is even greater for the NYT's product recommender title Wirecutter, whose revenues are mostly from affiliate, where readers click through to buy a recommended product, and the publisher takes a percentage of the sale. If users can see immediately what Wirecutter recommends, the need for that user to visit the original article and use the affiliate link there is massively reduced—meaning even clear attribution and a prominent link in the search interface are of little value.

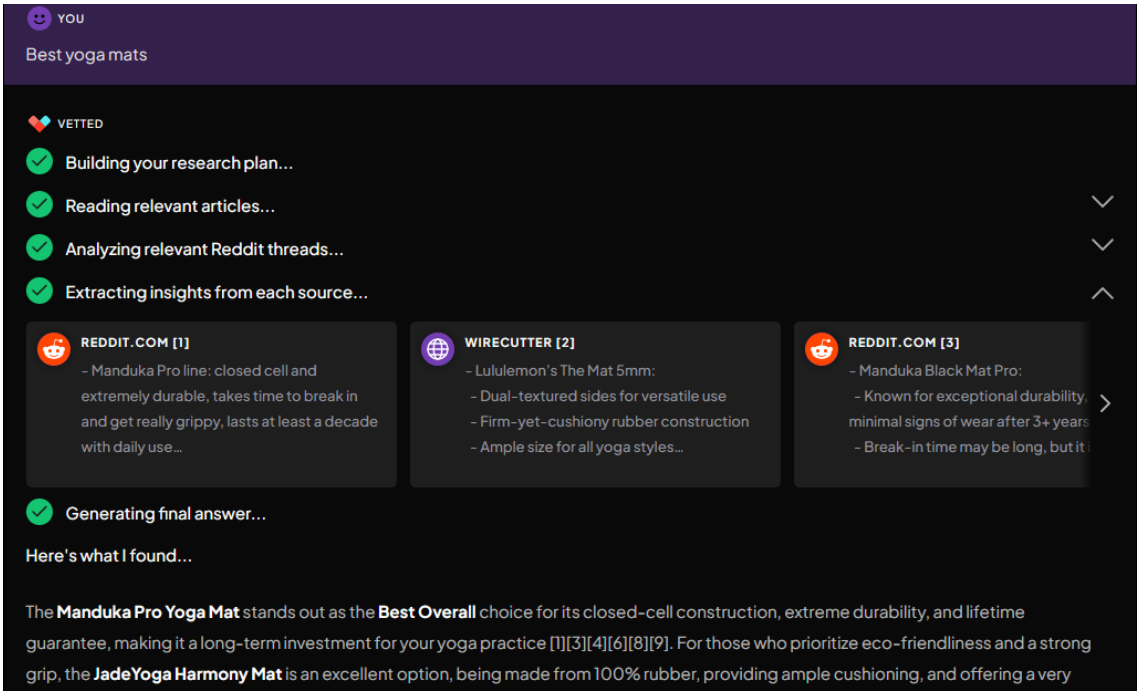
There are material revenues at risk here for UK publishers too, where affiliate can make up a sizeable chunk of income—Future plc generated 34% of its revenues from affiliate in 2023. Accreditation revenues, which have very high upfront capital investment, are also at risk (e.g. Good Housekeeping's quality assessment tests, which require testing facilities, office workspace, video production facilities).

The degree of exposure is dependent on the verticals that publishers operate in. Those active in highly immersive interest and 'passion' verticals, such as tech, cycling, photography, or music, have audiences who tend to value community and trusted recommendations, and so may be better insulated. In contrast, when consumers search out reviews for categories like white goods, the relationship is likely to be more transactional—users read reviews, buy the product and leave, with no guarantee they will use the site again at a later date. The balance of archive/ongoing access is different to news content, with more value lying in the archive as some categories of article can remain relevant (and therefore monetisable) for years after they are written.

Generative search isn't the only threat: shopping focused AI products that aggregate reviews (like Vetted, which trawls 'reliable' sites including Wirecutter, and Reddit—see Figure 4) could take off independently of licensing arrangements with the big search providers.

This is an obvious case of where the AI snake could eventually eat its own tail. Users might appreciate a product that can synthesise review content found in disparate parts of the internet, but even if it can generate or summarise text very well, AI fundamentally can't test products (like it can't do most journalism). If they want to be synthesising more than just Reddit posts, these AI products are ultimately relying on the business model that underlies product reviews—whose sustainability they directly endanger.

## Figure 4: Vetted provides AI-generated summaries of product reviews




[Source: Vetted, Enders Analysis]

# Unprecedented legal and practical questions cloud publishers' leverage in licensing negotiations

## Publisher leverage hinges on whether and where LLM training and deployment is considered copyright infringement

The creation and deployment of an LLM involves several steps that *could* be considered copyright-infringing, but aren't *necessarily* infringing: this involves genuinely new uses of copyrighted works that will have to be clarified either through new case law, or legislative/policy clarification. Outcomes could vary by jurisdiction; this section considers primarily the US and UK. In the EU, the copyright status of training at least is a bit clearer following a post-Brexit update to copyright law that allows for commercial text and data mining, with an opt-out option for publishers—see the end of this report.

### How LLMs work in brief:

LLMs are trained on a huge corpus of text, made up of copyrighted works and other internet content (like Wikipedia). The 'parameters' of the LLM are set when it is trained on these datasets: it learns complex patterns, and then generates text output by predicting what terms are likely to follow others in a sequence. After a process of pre-training, the LLM is fine-tuned on task-specific datasets, and its parameters are modified in a number of ways: this can be done to make it more reliable at producing a given kind of output. Higher quality training data can be more heavily 'weighted' in the fine-tuning process to ensure higher-quality outputs, or outputs that conform to a certain style. The LLM can be continually adjusted using this sort of fine-tuning procedure. Once deployed, a model might also be 'grounded' with material that wasn't part of its original training set, like a search LLM having ongoing access to websites at the point of user query, so that its output is relevant and up-to-date. Guardrails can be added to make it more or less likely to output certain kinds of content.

Even before the training process begins, the act of **creating a training database** involves copying and storing large amounts of copyrighted works, some of which has been extracted from behind a paywall. Then there is the **training process**, and the creation of the model: this involves turning text into tokens, and feeding them into the model. **Fine-tuning** might involve high quality works being weighted particularly heavily. It's relevant *where* these parts of the process happen: AI developers could shop around for a friendly jurisdiction in which to train (though in some cases this could restrict where they can deploy—see the EU case at the end of this report). It is also relevant *when* this occurred: copyright claims can be time limited.

Does this process make the **model itself** an infringing work? If its parameters merely capture the usual relationship between words in a statistical model, maybe not. Copyright protects expression, not underlying ideas, facts, grammatical structure, logic or the relationship between words. Still, if a model can reliably *reproduce* copyrighted content, a court could plausibly decide that this amounts to the model 'containing' or 'constituting' a copy of the training data.

Then there is the **deployment** of the model itself, and its **grounding** with up-to-date access to scraped material. The *context of its use* is important—whether it is used for internal research, or in the form of a consumer or enterprise product (like ChatGPT). The **generative output** itself could be infringing. The context and practicalities of model deployment (such as whether a model is *in fact* used to generate copies of copyrighted work, and at what scale, or whether this is commercial activity) may affect whether copyright has been infringed, and would certainly impact the scale of any damages.

Some generative outputs will be clearly infringing: if I get an LLM to generate a novel featuring copyrighted Harry Potter characters, which I then commercially publish, that output would obviously be a case of infringement. But in many cases, it is difficult to determine *where* the infringement happened, and *who* did the copying or infringing (me as the user, the model owner or deployer, whoever copied the works for the original training set).

This might have to be decided on a case-by-case basis, which makes it an uphill battle for rightsholders pursuing litigation. In the example of my novel, if the model used was not trained on Harry Potter books, and I managed to make it generate that output through careful prompting, then it may be like me writing the infringing work myself without AI assistance. If the model was trained almost exclusively on Harry Potter books, then the model developer might also be liable, and there would be more of a case that the model itself constitutes an infringing work.

What if Harry Potter was only one of thousands or millions of works that were used to train the model that outputs the infringing work? In that case, is the model still a copy? Is it a copy of millions of different works—of every Wikipedia page and social media post that went into its training? Early decisions on some of the lawsuits filed by authors in the US suggest that courts there will be unlikely make a blanket judgment on models without considering the similarity of outputs. We don't expect to see courts ordering the destruction of LLMs on the wider principle that they are infringement *as such* because they have ingested copyrighted material.

The strongest claims of copyright owners concern where models can output verbatim copies of copyrighted material. AI developers consider reproduction of material in outputs a bug that they have already begun to fix, and user prompting of this kind of verbatim output is likely against terms of service of the product—the scale of distribution of the infringing outputs, and how easy it is to get the model to produce them, will both matter. Our example of *commercially publishing* an infringing work is a lot clearer than a chatbot occasionally spitting out phrases that are similar to the text of Harry Potter, as part of a conversation with an individual user on a chatbot interface—even in the NYT case, the allegedly infringing examples are not *full* articles produced verbatim. If products have guardrails to reduce the likelihood of these kinds of outputs, then this may provide some legal protection.

The different stages of the AI supply chain makes matters even less straightforward. How is liability assigned for an open-source model if the deployer of the model had no role in its pre-training, and the company that trained it had no role in deployment? Many popular training datasets were scraped for permissible non-commercial purposes.

Jurisdictional differences will have an impact. In the US a fair use defence can involve a concession that a defendant *has* 'copied' copyrighted work, but it is made permissible through factors like 'transformative purpose': the model itself could be such a different kind of product to news content that it is not an infringing work. Generative search results might be transformative too: particularly if guardrails make verbatim copying much less likely. There is a lack of clear precedent in UK case law too, but here 'fair dealing' is a more restrictive condition than US fair use, with specifically outlined exceptions: commercial uses will be harder to justify.

## News publisher content is most important for LLM deployments that require ongoing grounding

While the copyright status of the different stages is undetermined, how *important* a given segment of training data is to the performance of a model is unknown outside of the tech companies testing the technology (and even internally to some extent). The answer will vary by kind of publisher, and whether we are talking about historical training on archival material; continued access to archives (for training *new* models); or ongoing access to up-to-date news content at the point of query. AI developers will be experimenting with LLMs that *don't* use news publisher content in training, or that only use licensed

material: LLMs might need *some* news publishers for quality output, but not *all* of them. In the absence of collective bargaining, this could allow AI companies to divide and conquer news publishers with smaller deals, especially if AI provision is concentrated.

**On the training side**: a defining characteristic of LLMs is the *scale* of their training sets: together, the copyrighted content that has fed LLMs is hugely important, but the marginal contribution of any given publisher's content is likely to be reasonably small to practically non-existent. The bargaining power of an entity like the NYT, whose domain features heavily in some key training sets, will be much higher than that of a small local publisher, or a blog owner, or the author of an individual social media post.

Snapshots of key training sets used by OpenAI in its earlier models show the NYT domain ranking highly, followed by the LA Times and the Guardian. Forbes, the Huffington Post and The Washington Post also rank highly. Some UK publishers can point to similar prominence. In its submission to a House of Lords Communication and Digital Committee report on LLMs, DMG Media submitted historical snapshots showing MailOnline's prominence at one point in the key OpenAI dataset WebText (making up 0.3% of its content), and in Common Crawl (540k webpages, 0.007% of content in a period in 2018).

Many of the biggest LLM providers aren't transparent about the contents of their training sets (OpenAI and Meta, for example, stopped releasing that information publicly after their first few models). For more transparent earlier models, we know that their training content included a number of datasets compiled using large-scale web-crawlers, a process that is not itself copyright-infringing when it is not commercial, under an existing text and data mining exception. This includes Common Crawl—a non-profit that "maintains a free, open repository of web crawl data that can be used by anyone". Such datasets may be filtered before being used to train models.

With open-source models historically trained on copyrighted material out in the world, the toothpaste is out of the tube to some extent: publishers won't be able to completely undo historical access to regain control over the development of models. Still, where big tech companies are training the next generation of models—a process that involves new training sets, rather than just building on top of earlier models—publishers have more leverage now that they are opting out of scraping for this purpose.

We do not know how much worse an LLM would perform if publisher content were excluded from training—this determines how effective a bargaining chip withdrawing access is. Wikipedia's strong presence in key training sets will provide a base of informative content, but more sophisticated sources could provide stylistic, reasoning and other benefits.

Publishers have pointed to their use in **fine-tuning**, with their content weighted more heavily in that process. DMG Media notes that MailOnline content made up 70% of a key dataset (along with CNN content) used in the fine-tuning process by Google DeepMind and OpenAI on some earlier models. The content was useful for training because of its uniform structure of headline, bullet point summary, and full article text. That structure is of course not unique, nor is it copyrightable, but if it was *in fact* used in this way, a case could be made that MailOnline content played a special role in the development of the tech. In practice there might not be many other sources with the scale of useful examples around a range of output, but as with the wider issue of training inputs, MailOnline could potentially be substituted here. There is no guarantee that fine-tuning methods haven't already moved on to make this particular format of text less critical, or that model makers couldn't point to a host of finetuning sequences where the publisher content was not used at all.

**Grounding models with ongoing access to publisher material once deployed is where publishers have the strongest case that they are important**. Access to reliable journalistic content at the point of query makes LLM responses to factual questions far more accurate than relying on model weights, determined by a much larger bank of older training data. Here, incentives might align, allowing some publishers to come to productive licensing arrangements.

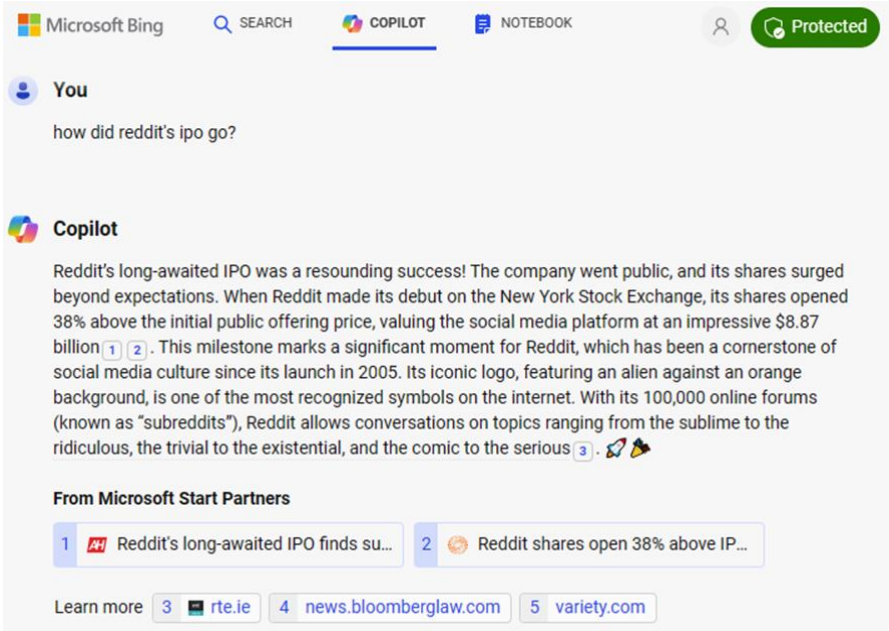
At present publishers don't have full control over the use of their content for grounding as opposed to training, particularly in the case of generative search, making it difficult for them to establish leverage in negotiations in practice. Google and OpenAI now allow publishers to opt out of ongoing crawling without turning off search crawling through the robots.txt protocol (Figure 5), but it is not clear that they can opt out of their content contributing to search generative answers where that activity is classed by the tech companies as a search functionality, and is instead bundled with search crawlers (Figure 6).

## Figure 5: UK nationals using robots.txt to opt out of OpenAI and Google AI crawlers

| | OpenAI | Google AI |
|---|---|---|
| The Times | ✗ | ✗ |
| The Telegraph | ✗ | ✗ |
| The Guardian | ✗ | ✗ |
| Daily Mail | ✗ | - |
| BBC | ✗ | ✗ |
| FT Financial Times | ✗ | - |
| Mirror | - | - |
| The Sun | ✗ | ✗ |
| Independent | - | - |

'X' signifies publishers have opted out. Publishers have been able to block Google Extended AI crawler separately to search crawlers since late 2023. Blocking crawlers by other means can require significant investment. Retrieved 17 April 2024.
[Source: Enders Analysis, publisher sites]

## Figure 6: Microsoft generative search integration




[Source: Enders Analysis, Bing]

## Industry licensing bodies and startups take tentative steps forward

One way for industry to begin to work around the lack of certainty is to avoid an overreliance on slow-moving legislation and the cases brought by some of the biggest publishers, which have their own incentives—and who might be taken out of the game once they settle. This could also help smaller or newer model makers license content without deploying the same resources as established tech.

Assuming any competition hurdles can be overcome, any kind of collective bargaining is still fraught given the diversity of the publishers and their business models:

- How would royalties be distributed? Traffic size? Circulation? Average time spent? Some publishers will argue their content is worth more, based on quality or commercial relevance, but this is thorny to prove and that is before considering that many publishers have a vast range of type of content—and that, in any process, the smallest publishers will struggle to have a voice
- What form would royalties or other revenue guarantees take? Some publishers would prefer an annual licensing fee divided by an agreed formula, while others would opt for guaranteed referrals they could monetise via programmatic advertising or otherwise
- What special terms would be acceptable, for instance, to help protect affiliate revenues? The more special terms granted, the more likely the system becomes unwieldy, but one-size-fits-all may favour the largest brands
- How would such a body fit in with larger publishers who strike their own deals on the side? Clearly the appetite for a *compulsory* and *exclusive* regime will be slim for the most part
- Finally, the most pressing issue to resolve in the near term—who would be the proper governing body to administer such as system? We consider some options to close off this report

One Boston startup, TollBit, hired Meta veteran Campbell Brown last year as it seeks to create a marketplace matching AI scrapers to publisher content with a dynamic fee set by both sides. TollBit is particularly focused on the supply of up-to-date information to power new AI products, and provides data back to source companies about the traffic generated. The disadvantage of such a product is it may not capture the nuance of different publisher demands or allow for industry-wide terms to be set. TollBit's own site suggests it is aimed to be a complement to individual licensing deals, to provide reporting infrastructure, rate limits, and authentication.

Given many publishers are anchored within a core home market, negotiating through a national copyright agency—such as the NLA in the UK and CCC in the US—is another option. Last July, the NLA reported it had 700,000 copyright-infringing articles taken down in the year so far, compared to 30,000 in 2022. These agencies are likely to lack the resources to tackle the growing scale of enforcement and breadth of AI-specific issues outlined, and some partnership with industry may be required—the NLA already works closely with the IPO and Google's copyright counsel team.[8]

Given the scale of possible disruption from generative AI, supervision by a body such as the CMA, as is ongoing for Google's Privacy Sandbox, may provide a forum for stakeholders to come to consensus terms. Nonetheless, this kind of process must be underpinned by a formal investigation, slowing the timeframe, and entrenching the competitive advantages of the big players chosen to be supervised.[9] As mentioned above, some publishers have called for the DMU to review these issues as part of the new competition regime, but many are wary of being bounced into a *compulsory* licensing regime.

[8] Press Gazette, News copyright theft 'through the roof' as 700,000 articles taken down in six months, 31 July 2023.
[9] The CMA published an updated paper on Foundational Models on 11 April 2024, identifying risks to "fair, effective, and open competition" centred on an "interconnected web" of over 90 partnerships and strategic investments involving key firms: Google, Apple, Microsoft, Meta, Amazon, and Nvidia.

## The EU's AI Act: transparency over model inputs, right to opt out of commercial scraping for model training

The EU's AI Act doesn't change European copyright law—which includes a text data mining (TDM) copyright exception that covers commercial LLMs—but it contains two crucial clarifications. A year after it passes (likely in the next few months), providers of general-purpose models will need to:

- Provide a summary of their training datasets
    - This provision is not just about copyright, but it should mean that rightsholders would *know* when their content is being used. There has been some disagreement around how detailed this summary would need to be
- Put in place a policy to respect rightsholders' right to opt out of commercial TDM
    - This should make it easier for publishers to opt out of scraping for model training

The EU's TDM copyright exception (introduced in the 2019 copyright directive, subsequently adopted by most Member States) allows for web scraping for both research and commercial purposes. For commercial TDM, rightsholders have a right to *opt out* of the TDM exception. This is why the likes of OpenAI has claimed that European copyright law has been on its side up to now. Rightsholders complain that they did not have an effective means of opting out until very recently, and that they didn't always know when they were being scraped, or when that activity changed from research to commercial.

The AI Act clarifications should ensure that publishers know when they are being scraped to train LLMs, and have the clear ability to opt-out. (The issue of non-commercial bodies scraping and then making that data available to commercial players might still cause problems under this system.) This doesn't quite amount to a statutory licensing regime, but it provides a legal backstop: rightsholders should have an easier time exercising their right to opt out, and can make licensing a condition of not opting out.

These provisions are supposed to apply no matter *where* AI training happened: so even if training is deemed fair use in the US, US providers could be restricted in what models they can deploy in the EU without licensing. This could end up being hugely restrictive for future models, depending on how enforcement works in practice.

In the meantime, Google has separately been hit with a €250 million fine by the French competition authority, in part for its failure to notify publishers that their work was being scraped to train AI models, which constituted a breach of its commitment to deal with publishers in good faith. This commitment came from a case around licensing content to appear in Extended News Previews in search, which are protected by a neighbouring copyright under the 2019 directive. Google had originally refused to pay out in France on the grounds that publishers had the right to opt out of being featured in this way in search results (a right most publishers did not want to exercise—not a problem they will necessarily have in the emerging AI licensing system).

### Ongoing access for grounding queries is still less straightforward

The issue of scraping for grounding models is more slippery and less clearly covered by the AI Act (or existing rights) than model training. The French regulator noted that it is not yet determined whether using press publications in an AI service falls under the protection of neighbouring rights. If generative search responses were to replace Extended News Previews, this could strain the usefulness of the narrowly defined neighbouring copyright.

# About Enders Analysis

Enders Analysis is a research and advisory firm based in London. We specialise in media, entertainment, mobile and fixed telecoms, with a special focus on new technologies and media, covering all sides of the market, from consumers and leading companies to regulation. For more information go to www.endersanalysis.com or contact us at info@endersanalysis.com.

If your company is an Enders Analysis subscriber and you would like to receive our research directly to your inbox, let us know at www.endersanalysis.com/subscribers