LIQVID

A LIQVID RESEARCH WHITE PAPER

Content Under Contract

The Emerging Architecture of AI Content Licensing in the Global Publishing Industry



About this report

This report maps the market for AI content licensing as it stands in 2026 — a market that, in barely three years, has grown from an informal free-for-all into a structured, multi-billion-dollar arena where publishers and AI developers negotiate the terms on which the written word may be used to train and power artificial intelligence.

It examines that market from four angles: its **scale and growth**, from the deals already signed to the trajectory ahead; the **five licensing models** now competing to become the standard, from bespoke direct deals to collective and marketplace arrangements; the **diverging legal frameworks** taking shape across the United States, the European Union, the United Kingdom, and India; and the **structural forces** that concentrate value in the hands of the largest publishers and the largest AI firms alike. The report is written for **publishers** weighing how to value and license their archives, **AI developers** navigating an increasingly regulated supply of training data, and **policymakers** shaping the rules of a market that is still being invented in real time.

About the Author(s)

Prerna Pardasani (prerna.pardasani@liqvid.com) is an Analyst in the Founder's Office at Liqvid, where she researches AI content licensing and publisher strategy. She is drawn to this subject because it sits at a rare inflection point: the moment an entire industry renegotiates the value of its own archive. AI has turned publishers' back catalogues into critical infrastructure for a new technology — and the terms of that exchange, still being written today, will determine who captures the value the AI era creates. This paper is her attempt to make sense of that shift, and to give publishers the market intelligence to act on it.

Vivek Agarwal is the Founder and CEO of Liqvid and a passionate ed-tech entrepreneur with over 25 years of experience. He is driven by a vision to deliver personalized learning through the innovative application of technology. An MBA from the IIM Calcutta, Vivek founded Liqvid (formerly English Edge) to advance that vision — building AI-enabled English language learning solutions used by organizations worldwide. Liqvid's programs have reached more than a million learners across 5,000 locations, delivering over 1.5 million hours of instruction. Earlier, in 2000, he founded eGurucool, a pioneering Indian e-learning platform for K-12 curriculum support and test preparation, which raised over USD 11 million from investors including News Corp and IFC before its acquisition by NIIT.

© 2026 LIQVID

Contents

Summary 5
01 Introduction 6
02 Literature Review 7
03 Sizing the Market: Scale, Growth, and Structure 8
04 A Typology of Licensing Models 10
05 Legal and Regulatory Frameworks 11
06 Competitive Dynamics and Market Concentration 14
07 How Publishers Are Responding 15
08 The Indian Case: A Natural Experiment in Regulatory Design 16
09 Future Outlook 17
10 Conclusion 18

List of Tables

Table 1: Selected AI content licensing deals, 2024–2026. 9
Table 2: Comparative regulatory frameworks for AI content licensing 13

List of Figures

Figure 1: Four Years That Rewrote the Rules of Licensing 6
Figure 2: Five Ways to License AI Content 10
Figure 3: The Regulatory Timeline 11

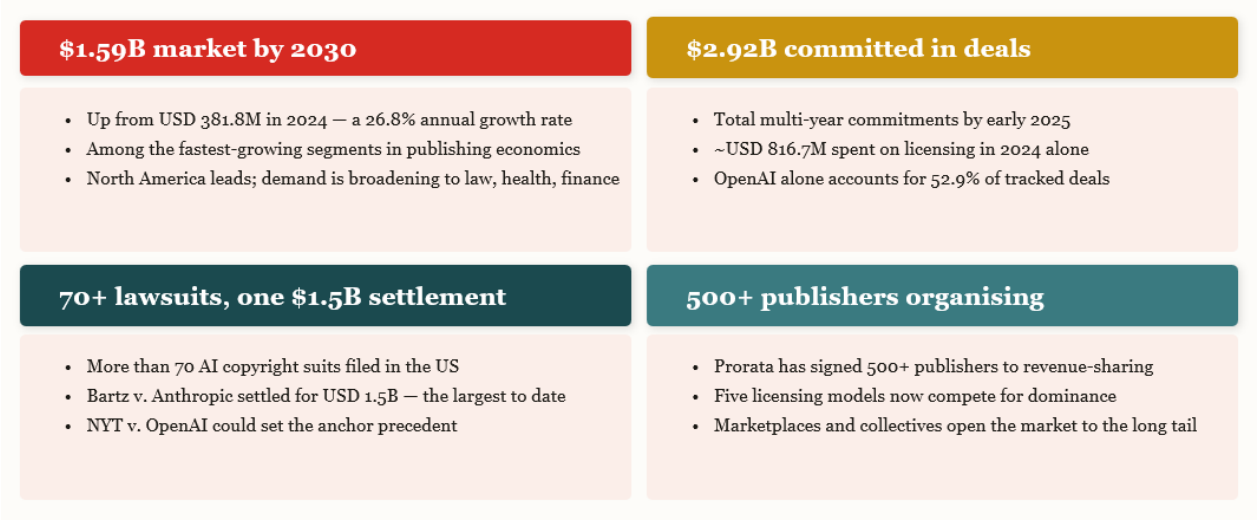
© 2026 LIQVID

Summary

In little more than three years, the relationship between AI developers and the publishing industry has been rebuilt from the ground up. What began as the quiet, unlicensed ingestion of newspapers, books, and scholarship into training data has become a negotiated market worth billions of dollars — contested in dozens of courtrooms and increasingly shaped by regulators on four continents.

This paper maps that market as it stands in 2026. Drawing on disclosed deal terms, litigation records, regulatory texts, and industry reporting, it sizes the market, develops a working typology of the five licensing models now competing for dominance, and examines the sharply divergent legal paths taken by the United States, the European Union, the United Kingdom, and India. Four tensions run throughout: publisher autonomy against the push for AI innovation; the difficulty of attributing value inside an opaque model; the trade-off between scalable and bespoke licensing; and the friction between national regulatory sovereignty and a global AI market. It closes with India's proposed mandatory blanket-licensing regime — a policy experiment no other major economy has yet attempted — and draws out the implications for publishers, developers, and regulators navigating a market still being built.

Four forces define that market — its scale, its money, its legal battles, and the response taking shape. In brief:



Keywords: AI content licensing; generative AI; large language models; copyright; publisher rights; collective licensing; EU AI Act; data governance; publishing industry

© 2026 LIQVID

01 Introduction

The arrival of large language models — GPT, Claude, Gemini, Llama, and their many successors — has quietly rewired the economics of the written word. These systems learn from corpora measured in hundreds of billions, sometimes trillions, of tokens, and a disproportionate share of what makes them fluent and reliable comes from professionally produced text: reported journalism, peer-reviewed research, monographs, legal analysis, and fiction. That places the publishing industry in an awkward position. The institutions that have spent centuries building the machinery to produce trustworthy writing are now the essential suppliers to a technology that may, in the same motion, undercut the way that writing reaches readers and earns revenue.

In barely three years, the relationship moved from unlicensed extraction to structured and fiercely contested dealmaking. The milestones below trace that shift: a public tool that made the stakes visible, the first wave of litigation, the first nine-figure agreements, a landmark settlement, and the arrival of a marketplace model. The chapters that follow ask not when each step happened but why the old arrangement could not hold, why publishers and developers alike were driven to the negotiating table, and who gains and loses as the terms are written.

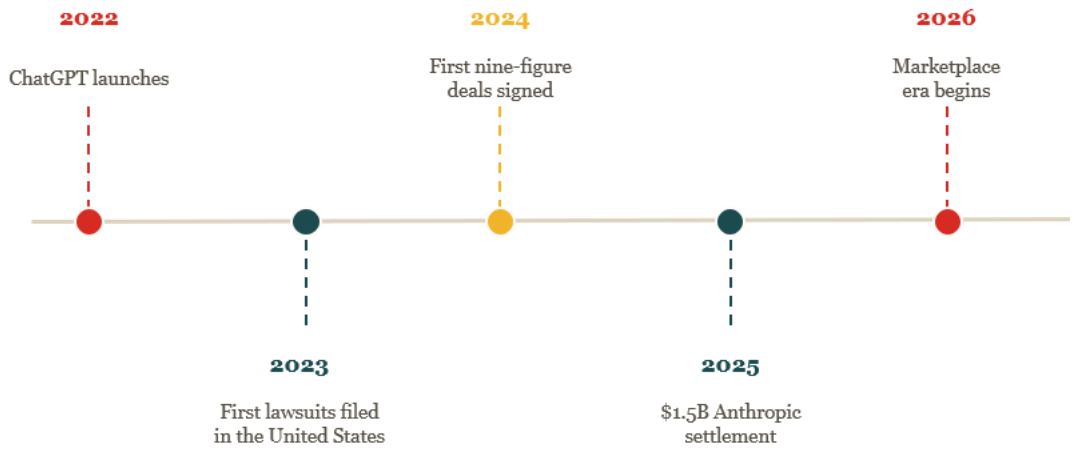


Figure 1: Four Years That Rewrote the Rules of Licensing

This paper is an attempt to describe that transformation while it is still in motion. It asks four questions. What licensing models have actually emerged, and which publishers do they serve? How are the divergent legal regimes shaping commercial outcomes? What structural dynamics reward the largest players on both sides of the table at the expense of everyone else? And what does India's unusually interventionist proposal reveal about the wider menu of choices open to governments?

The contribution here is threefold. It synthesises the current state of the AI content licensing market from the scattered evidence available — disclosed terms, litigation filings, regulatory documents, and trade reporting. It offers a typology of licensing models that situates each within its economic and legal logic. And it identifies the structural tensions likely to define the market's next decade, with recommendations pitched at publishers, developers, and regulators in turn. A note on evidence is warranted at the outset: because deal terms are overwhelmingly confidential, much of what follows

© 2026 LIQVID

rests on partial disclosure and industry estimates rather than audited figures, and should be read as a considered reading of an opaque market rather than a definitive accounting.

The remainder of the paper proceeds in nine further sections. Section 2 reviews the relevant scholarship. Section 3 sizes the market and describes its deal architecture. Section 4 develops the typology of licensing models. Section 5 works through the legal frameworks in the key jurisdictions. Section 6 turns to competition and market concentration. Section 7 examines how publishers are actually responding. Section 8 presents the Indian case in depth. Section 9 considers what comes next, and Section 10 concludes.

02 Literature Review

Scholarship on AI content licensing sits at the meeting point of three traditions: copyright law and information economics, the political economy of platform markets, and the newer field of data governance and AI policy. Because the phenomenon is so recent, rigorous empirical work remains thin, and much of the sharpest analysis still lives in working papers, policy briefs, and trade reporting rather than in settled journal literature. The review that follows draws these strands together to frame the market analysis in the sections that follow.

2.1 Copyright, Fair Use, and the Information Economy

The tension at the centre of AI licensing is doctrinal. Copyright protects original expression fixed in a tangible medium against unauthorised reproduction, adaptation, and distribution (17 U.S.C. § 102). Training a model on protected text introduces a use the statute never contemplated, raising a genuinely open question: does ingesting a work to extract statistical patterns count as reproduction in the copyright sense at all?

The fair use doctrine, codified at 17 U.S.C. § 107, has long served as the safety valve for transformative uses. Courts weigh four factors — the purpose and character of the use, the nature of the work, the amount taken, and the effect on the market for the original. Two Supreme Court decisions anchor the modern analysis. *Campbell v. Acuff-Rose Music* (1994) made transformativeness central and refused to treat commercial purpose as automatically disqualifying, while the earlier *Sony Corp. of America v. Universal City Studios* (1984) protected home video recording precisely because it did not substitute for the market it touched.

Applied to model training, these precedents pull in opposite directions. A trained model is plainly not a copy of any single book — yet it can, on occasion, reproduce a newspaper article closely enough to raise the spectre of market substitution. The scholarship reflects that ambivalence. Samuelson (2023) argued that outcomes will turn on how courts characterise the purpose of training, a question then unresolved at the appellate level. Lemley and Casey (2024) contended that training is unlikely to clear the fair use bar once the commercial stakes and substitution effects are weighed, while Grimmelmann (2024) took the opposing view, reasoning that learning from the non-expressive, statistical features of text is doctrinally distinct from copying its expression. As Section 5 shows, the courts have since begun — inconsistently — to test these positions in practice.

© 2026 LIQVID

2.2 Platform Economics and the Value of Data

A second literature reads AI licensing through the lens of platform economics. Rochet and Tirole's (2003) foundational account of two-sided markets helps explain the incentives at work: a platform serving two groups at once will often subsidise the side it most needs to attract. For years, that logic encouraged AI developers to treat content as a free input and to monetise the finished model downstream — a stance that only became commercially untenable once litigation and regulation raised the cost of unlicensed ingestion.

The way data works as an asset adds another twist. Jones and Tonetti (2020) pointed out that data is "non-rival" — one company using it doesn't use it up for anyone else. In a perfect market, that would push its price close to zero. But in practice, quality and exclusivity create scarcity anyway. AI developers clearly value clean, fact-checked, professionally edited text far more than the larger but messier open web — which means high-quality content commands a premium that the "price near zero" logic misses. Arrieta-Ibarra et al. (2018) took the argument further, suggesting that supplying data should be treated like labour, deserving ongoing payment rather than a one-time sale — an idea that foreshadows the collective and revenue-sharing models now spreading through publishing.

2.3 Data Governance, Competition, and the Empirical Gap

The newest research tackles AI licensing directly. Epstein, Haber, and Levendowski (2023) examined the legal questions raised by training at scale and proposed a tiered licensing system, with terms set according to how the content is used and how much money it generates. The European Parliament Research Service (2025) gave a close reading of the EU's text-and-data-mining exceptions and how they fit with the AI Act. On the competition side, a report to the European Commission by Crémer, de Montjoye, and Schweitzer (2019) — written before the licensing debate began — offered useful tools for understanding how control of a unique data asset can become a competitive chokepoint. ProMarket (2025) applied those tools directly, warning that without standardised access, large AI firms will use exclusive deals to shut out rivals — effectively fencing off the data market.

The obvious gap across all of this work is hard evidence. Deal structures, pricing, and market outcomes stay mostly hidden behind commercial confidentiality. This paper pushes against that gap by piecing together what can be inferred from disclosed terms, regulatory filings, and trade reporting — while acknowledging that a fuller picture will have to wait for either greater transparency or dedicated survey work.

03 Sizing the Market: Scale, Growth, and Structure

3.1 A Market Built at Speed

Few segments of the publishing economy have grown as quickly as this one. The market for AI datasets and licensing in academic research and publishing was valued at roughly USD 381.8 million in 2024 and is projected to reach USD 1.59 billion by 2030, a compound annual growth rate of 26.8% (Grand View Research, 2024). North America leads, reflecting the concentration of frontier developers in the United States, and demand is dominated by the model builders themselves — though it is broadening

© 2026 LIQVID

as AI application firms in healthcare, law, and finance seek out specialised content libraries of their own.

The actual flow of money looks even larger. Tracking 35 disclosed deals through early 2025, one widely cited industry analysis put AI firms' content-licensing spend at roughly USD 816.7 million for 2024 alone, at an average deal size near USD 24 million, with total multi-year commitments approaching USD 2.92 billion (MediaAndTheMachine, 2025). Because these figures capture only disclosed transactions, the true total is almost certainly higher.

3.2 A Concentrated Demand Side

Demand is strikingly concentrated. OpenAI has been the dominant buyer, accounting for roughly 52.9% of tracked deals as of early 2025, followed at a distance by Google (11.8%), Microsoft (9%), and Meta (6%) (MediaAndTheMachine, 2025). That concentration mirrors the capital intensity of frontier model development and the tight link between model scale and appetite for high-quality data. Perplexity and a widening field of smaller developers make up a growing but still-minor share of demand.

Table 1 lists the landmark deals disclosed through early 2026. Taken together, they show a market that has not settled on a single formula: the deals range from large upfront training licences to narrower inference-only arrangements, and from big-name news brands to specialist academic publishers.

Publisher / Content Owner	AI Company	Reported Value	Key Terms / Scope
News Corp (excl. Factiva, HarperCollins)	OpenAI	~USD 250M, multi-year	Training & inference; excludes database and book assets
The New York Times	Amazon	Undisclosed	Training + AI assistant integration
The Washington Post	OpenAI	Undisclosed	Inference-only: summaries & links; no training use
Associated Press	OpenAI	~USD 12.5M (est.)	Training & citation in ChatGPT responses
Associated Press	Google	Undisclosed	Real-time news in Gemini; first Google–AP deal
Springer Nature	Google	~USD 23M	Backlist academic content for LLM training
Wiley	Multiple AI companies	~USD 23M	Scientific and academic journal content
Taylor & Francis	Undisclosed buyer	USD 10M upfront + recurring to 2027	STEM academic content
Agence France-Presse	Mistral AI	Undisclosed, multi-year	~2,300 daily stories across six languages
CNN, Fox News, People, USA Today	Meta	Undisclosed, multi-year	Meta's first systematic publisher programme

Table 1: Selected AI content licensing deals, 2024–2026. Sources: Press Gazette (2025–2026); Digiday (2026a, 2026b); Scott (2025); Media and The Machine (2025).

3.3 From Bilateral Deals to Marketplace Infrastructure

Through 2023 and 2024, the market grew one bilateral announcement at a time. Late 2025 marked a structural shift. In September 2025, Microsoft unveiled its Publisher Content Marketplace — the first attempt by a major technology firm to build a two-sided platform on which publishers could list content for AI developers to license at scale (Digiday, 2026b). If such intermediaries can cut transaction costs, standardise terms, and support competitive pricing, they may begin to address the long-tail problem that bilateral dealmaking has so far left unsolved.

Alongside the marketplace model, buyers and sellers are experimenting with compensation tied to demonstrated use rather than blanket archive access. Amazon's arrangements with Condé Nast and

© 2026 LIQVID

Hearst for its Rufus shopping assistant illustrate the pattern: payment is linked to a specific product application rather than a general training licence. A parallel development, especially visible in markets such as India, is the rise of specialist agencies that aggregate mid-sized publishers' catalogues and broker access to global AI buyers on their behalf — extending the reach of the market to houses that would never secure a direct negotiation, while introducing a new layer of intermediation whose terms and take-rates remain largely opaque.

04 A Typology of Licensing Models

4.1 Five Models in Competition

No single deal structure has won out. Instead, several models coexist, each suited to a different combination of publisher scale, content type, and developer strategy. Drawing on disclosed terms and trade analysis (Digital Content Next, 2025; Encypher AI, 2026; Digiday, 2026a, 2026b), five broad categories can be distinguished.



Figure 2: Five Ways to License AI Content

4.2 The Attribution Problem

A single difficulty cuts across every one of these models: the opacity of the training process itself. Models do not draw on individual works in any traceable, linear way. They distil statistical regularities from vast quantities of text, and the resulting capabilities cannot be cleanly attributed to any one input. Establishing how much a given publisher contributed to a model's performance is therefore both technically hard and commercially sensitive, which leaves developers holding far better information than the publishers they negotiate with.

The consequences run in three directions. Collective and revenue-sharing schemes struggle to allocate royalties fairly without new auditing tools. Developers enjoy an informational edge in bilateral talks, knowing better than the publisher how much a given archive actually improved the model. And high-quality content risks being systematically under-priced relative to commodity text, because quality differences wash out in aggregate performance metrics.

© 2026 LIQVID

Emerging technical work offers a partial answer. Data-valuation methods adapted from cooperative game theory — the Shapley value and its faster approximations — attempt to assign each contributor a defensible share of a model's output, and influence-function techniques have recently been scaled to LLM-sized training sets to estimate how much specific data shaped specific behaviour (Choe et al., 2024). None of these is yet robust or cheap enough for routine commercial use, but they mark the beginning of the auditing infrastructure that fair compensation at scale will ultimately require. The EU AI Act's training-data disclosure mandate, discussed in Section 5.2, is the first serious regulatory push in the same direction.

4.3 Deal Structure by Publisher Segment

Which model fits depends heavily on where a publisher sits. The largest houses — established news brands, major academic publishers, big trade imprints — typically command substantial upfront payments plus recurring fees, and can afford to carve out their most valuable assets. News Corp's exclusion of Factiva and HarperCollins from its OpenAI deal is the textbook example of a publisher structuring a licence to protect its crown jewels (Press Gazette, 2025).

Mid-tier publishers more often accept bundled partnerships with a revenue-sharing component, or route their catalogues through aggregators and agencies of the kind now emerging in several markets. The long tail — independent journalists, local outlets, boutique digital media, and freelancers — remains the least served. Their content may be collectively valuable but individually too small to justify a direct deal, leaving collective licensing and marketplaces as their only realistic routes to compensation. Because the attribution problem bites hardest at this end of the market, there is a real risk that what these publishers receive will bear little relation to what their work actually contributes.

05 Legal and Regulatory Frameworks

5.1 United States: Litigation as the Battleground

Nowhere has the legal contest been more active than in the United States, where copyright owners had filed more than seventy infringement suits against AI firms by early 2026 (Copyright Alliance, 2026).

The plaintiffs span authors, news organisations, visual artists, and online creators; the defendants include OpenAI, Anthropic, Meta, Google, and Microsoft. Fair use is the central battleground, with developers arguing that training is transformative and non-substitutive, and rightsholders arguing that outputs can and do displace the originals.

The courts have answered inconsistently, which is itself telling. In *Kadrey v. Meta Platforms* (2025), the court granted partial summary judgment for Meta, finding that training on the plaintiffs' books was sufficiently transformative to attract fair use protection (Norton Rose Fulbright, 2026) — a result widely read as favourable to developers. By contrast, in *Thomson Reuters v. Ross Intelligence* (2025), the court sided with the rightsholder on a narrower but commercially pointed ground: copying legal headnotes to build a directly competing product was not fair use. The through-line is that market substitution, more than the act of copying itself, is emerging as the decisive factor.

The case that matters most is still The New York Times Co. v. OpenAI and Microsoft, which puts the fair use defence to its hardest test yet: the mass reproduction of news reporting. The Times argues that

© 2026 LIQVID

OpenAI's models can reproduce its articles almost in full, functioning as an unlicensed stand-in that chips away at both subscription and licensing revenue (Harvard Law School, 2024). Even before any verdict, the case has changed how these disputes are fought. In January 2026, Judge Sidney Stein upheld a magistrate's order requiring OpenAI to hand over a sample of 20 million anonymised ChatGPT logs, brushing aside the company's privacy objections — a discovery ruling that will likely shape how courts handle AI firms' internal data for years to come (Bloomberg Law, 2026). There is an irony here too, one the Harvard Law Review (2024) was quick to point out: in New York Times Co. v. Tasini (2001), the Times leaned on the argument of technological necessity to prevail over freelance writers. Now the same argument is being turned back on it by OpenAI.

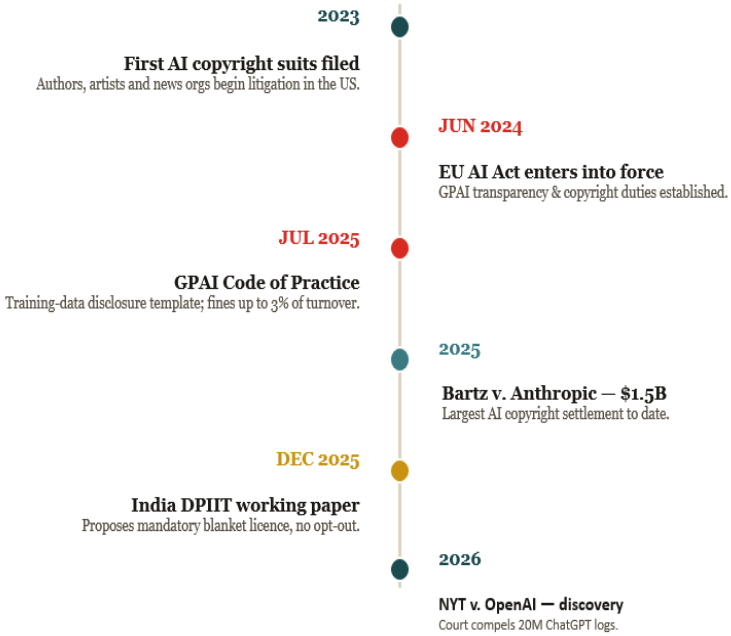


Figure 3: The Regulatory Timeline

The consolidation of sixteen suits into a single multidistrict proceeding in the Southern District of New York means the eventual fair use ruling there is likely to become the anchor precedent for the entire wave of litigation (Copyright Alliance, 2026). The largest resolution to date, however, came by settlement. In *Bartz v. Anthropic* (2025), Anthropic agreed to pay at least USD 1.5 billion — roughly USD 3,000 for each of some 500,000 works — to resolve claims that it had trained on pirated copies of books. Crucially, Judge William Alsup's June 2025 summary-judgment ruling had already split the question: training on lawfully acquired books could qualify as fair use, but downloading pirated copies did not. The settlement therefore turned on how the books were obtained rather than on whether training is lawful in principle, and it left the deeper doctrinal question unresolved even as it signalled the scale of exposure developers now face (Authors Guild, 2026).

5.2 European Union: Regulation over Litigation

The European Union has taken the opposite tack, leading with regulation rather than litigation. The AI Act (Regulation (EU) 2024/1689) is the centrepiece. Its provisions on general-purpose AI require model providers to adopt a copyright-compliance policy and to publish a sufficiently detailed summary of their training data, so that rightsholders can assess potential infringement and exercise the opt-out

© 2026 LIQVID

available under Article 4 of the Digital Single Market Directive (European Parliament Research Service, 2025).

Two 2025 developments gave those obligations teeth. The GPAI Code of Practice, published in July 2025, set out implementation guidance on copyright, transparency, and safety, and the Commission issued a mandatory template requiring providers to disclose the sources and nature of their training data — public datasets, private datasets, scraped web content, user data, and synthetic data alike — with non-compliance exposing firms to fines of up to 3% of global annual turnover (WilmerHale, 2025). The core disclosure duties took effect on 2 August 2025. Layered on top of the pre-existing text-and-data-mining exception, under which mining is permitted unless rightsholders reserve their rights in machine-readable form, the result is a compliance regime of real complexity (Clifford Chance, 2025).

Bruegel (2025) captured the bind neatly: enforce copyright too strictly and Europe risks cutting itself off from frontier models built elsewhere; permit mining too freely and European creators are under-compensated while non-EU developers face fewer constraints at home. The Code of Practice is an attempt to thread that needle, but its practical effect will only become clear through enforcement.

5.3 United Kingdom: A Regulatory Vacuum

The United Kingdom charted an initial course toward a broad copyright exception for AI training, then reversed after sustained opposition from publishers, authors, and creative-industry bodies. As of early 2026 it has enacted no AI-specific copyright legislation, leaving developers and publishers to operate under a copyright framework that predates generative AI. With more than forty publishers having issued formal legal notices to AI firms (TechJack Solutions, 2026), the UK now combines high near-term litigation risk with unusual regulatory uncertainty — arguably the least settled position among the major markets.

5.4 Divergence Across Jurisdictions

Table 2 sets the frameworks side by side. The same underlying question — does training require a licence? — has produced strikingly different answers, and that divergence is itself a source of cost.

Jurisdiction	Primary Mechanism	Fair Use / TDM Scope	Opt-Out Rights	Key Risk for Publishers
United States	Judicial (case law)	Broad; doctrine unsettled	Not codified; market-based	Adverse fair use ruling forecloses claims
European Union	Regulatory (AI Act + DSM Directive)	Narrower TDM exception	Yes; machine-readable opt-out	Compliance complexity; opt-out administration
United Kingdom	Uncertain; exemption withdrawn	Traditional fair dealing (narrow)	No statutory opt-out at present	High litigation risk; regulatory vacuum
India	Proposed mandatory blanket licence	Limited fair dealing (§ 52)	None proposed; mandatory access	Loss of editorial and commercial autonomy
Australia / Canada	Under review	Varies by jurisdiction	Under consultation	Regulatory uncertainty; bilateral exposure

Table 2: Comparative regulatory frameworks for AI content licensing. Sources: Copyright Alliance (2026); European Parliament Research Service (2025); Tech Jack Solutions (2026); DPIIT (2025).

© 2026 LIQVID

For developers operating globally, this fragmentation is expensive. A structure that satisfies EU transparency rules may be commercially suboptimal under US law; a practice protected by US fair use may infringe in the UK or EU. The compliance burden falls hardest on smaller firms, handing yet another structural advantage to the largest developers with the legal infrastructure to operate across regimes at once.

06 Competitive Dynamics and Market Concentration

6.1 Content as Competitive Infrastructure

Licensed content has become, in effect, infrastructure for the AI economy. Access to large, curated, high-quality archives confers an advantage in model training that is hard to replicate, and the link between data quality and model performance is well established (Hoffmann et al., 2022). Exclusive access to premium content therefore functions as a genuine competitive moat.

ProMarket (2025) characterised the resulting dynamic as competition for the market rather than within it: heavy upfront investment, scale economies in data acquisition, and reinforcing feedback loops between data quality and performance. Once a developer trains on a superior corpus and pulls ahead, the combination of user network effects and privileged data access makes that lead increasingly difficult to challenge. Licensing decisions taken today may thus shape the competitive structure of the industry a decade out.

6.2 The Long-Tail Exclusion Problem

The market's architecture systematically favours large players on both sides. Only well-capitalised developers can execute major licences; only large publishers have the leverage and archive depth to negotiate directly. The result is a two-tier structure — a well-compensated tier of premium publishers with bilateral deals, and an under-compensated long tail whose content is nonetheless swept into training sets.

The problem has both an equity and a competition dimension. On equity, it under-pays the journalists, bloggers, academics, and independent publishers whose collective contribution may be substantial even where each individual share is small. On competition, it entrenches incumbent media conglomerates, which can monetise archive depth, against newer entrants who have no comparable back catalogue to license. ProMarket (2025) argues that a statutory licensing regime, setting minimum rates and mandating non-discriminatory access, could address both by treating premium content as regulated infrastructure rather than a source of exclusive advantage — akin to essential-facilities doctrine. The counter-argument is equally real: guaranteed access at administered rates could blunt the incentive to invest in original, high-quality work in the first place.

6.3 The Search-and-Training Bundling Problem

A specific asymmetry attaches to Google. Because its search crawler and its AI-training crawler have not been fully separated, publishers who opt out of AI training risk losing search visibility as well (Ansari, 2025; Times Network, 2025). For many digital publishers, organic search is a primary revenue driver, so the effective price of refusing AI ingestion is punishingly high. Microsoft has separated the two functions, allowing publishers to decline training while retaining indexing — a distinction

© 2026 LIQVID

publishers have noticed and rewarded (Digiday, 2025b). Google's bundling may eventually draw competition scrutiny, particularly in the EU, where it could be characterised as leveraging search dominance to undercut the opt-out rights the DSM Directive is meant to protect.

07 How Publishers Are Responding

7.1 Litigation and Partnership, Often at Once

Publishers have spread themselves across a spectrum from aggressive litigation to close partnership — and many occupy several points on it simultaneously. The New York Times sued OpenAI while licensing content to Amazon; the Washington Post struck an inference-only deal with OpenAI while reserving the right to litigate over training use. This dual posture reflects a real tension between the near-term revenue a licence delivers and the longer-term value of a favourable precedent that would strengthen every publisher's hand.

The litigation route carries its own hazards. Copyright cases are slow, costly, and uncertain, and fair use gives developers a credible defence even where a publisher feels plainly wronged. Settlement pressure can push publishers to accept terms below what a favourable judgment might have secured. And licensing at low rates while litigation is pending can hand the defence evidence that the market has already set an acceptable price — quietly weakening the damages case.

7.2 The Traffic Substitution Threat

The deeper worry for publishers is not training but substitution. AI-generated answers can answer a reader’s question without a click, eroding the advertising and subscription revenue that fund original work. Digital Content Next (2025) reports that Google’s AI Overviews have reduced referral traffic by an average of 25%, with some publishers experiencing a 50 to 90% drop in click-through rates when an AI summary was available (Digiday, 2026b). Licensing revenue must therefore compensate not only for training use but for an ongoing diversion of readers away from publisher platforms — which is why the distinction between training and inference licensing, visible in the Washington Post's approach, is becoming commercially central. Publishers increasingly have reason to price inference use, where the traffic loss actually occurs, well above training use.

7.3 New Sources of Leverage

Publishers have also begun building leverage into the infrastructure itself. In July 2025, Cloudflare introduced tools letting publishers selectively block AI crawlers or charge per-crawl fees (Digiday, 2026b). A credible ability to block raises the value of a licence, since developers must then either negotiate or go without — strengthening the publisher's hand even if the block is never actually used.

7.4 Collective Action and Its Limits

Coordinated approaches are spreading. The News/Media Alliance's arrangement with Prorata, offering members a 50% royalty on a revenue-sharing basis, drew more than 500 participating publishers (Digiday, 2025a). The Copyright Clearance Center's 2025 usage study found that only 62% of employees at organisations using AI tools understood that their copyright policies extended to third-party content used with AI — pointing to substantial latent demand for the institutional licensing that

© 2026 LIQVID

collective bodies are positioned to supply. The broader challenge is one of pace: the market is moving faster than any single publisher or association can respond, and the most durable strategies pair near-term revenue from licensing with longer-term positioning through litigation, regulatory engagement, and the development of AI-native products that let publishers participate in the value chain rather than merely feed it.

08 The Indian Case: A Natural Experiment in Regulatory Design

8.1 A Distinctive Regulatory Philosophy

India occupies an unusual position in this debate, at once a major content-producing nation and an ambitious aspirant in AI development. The Department for Promotion of Industry and Internal Trade (DPIIT) convened a committee on publisher rights in relation to AI, and in December 2025 it released Part One of a working paper whose recommendations carry significance well beyond India's borders. The case is worth close attention not only because India hosts one of the world's largest media industries, but because the regulatory philosophy on display departs sharply from both the American and the European models.

The DPIIT paper reportedly evaluated five licensing models and set aside voluntary licensing on grounds of scalability, recommending instead a hybrid that blends collective and mandatory elements — captured in the slogan ‘one nation, one licence, one payment’ (DPIIT, 2025; Times Network, 2025). The proposal has four defining features:

i. a mandatory blanket licence covering the use of protected content for model training;
ii. a centralised, government-designated non-profit to collect payments from developers and distribute royalties to rightsholders;
iii. government-appointed committees to set royalty rates; and
iv. most controversially, no opt-out — rightsholders would not be permitted to withhold their works from AI training.

8.2 Reading the Proposal Critically

The design is distinctive on two counts. First, removing the opt-out asserts a public interest over private property rights more forcefully than any comparable regime. It effectively treats content as quasi-public infrastructure for national AI development rather than as property whose owner may simply decline to license. That framing fits India's positioning of AI as a development priority, but it sits uneasily with the foundational idea of copyright as a bundle of exclusive rights that includes the right to say no.

Second, having government committees set rates introduces a form of tariff regulation more familiar from broadcasting and utilities than from content markets. Rate-setting across content as heterogeneous as investigative journalism and commodity wire copy is genuinely difficult, and in a sector as politically sensitive as news, administered pricing raises the further hazard of political influence over what publishers are paid. Indian media organisations have objected on both autonomy and economic grounds, arguing that a mandatory regime erodes the editorial and commercial

© 2026 LIQVID

independence that lets a newsroom decide who uses its work and on what terms (Times Network, 2025). The attribution problem discussed in Section 4.2 is especially acute here: fair proportional distribution presupposes a granular measure of each publisher's contribution that current technical methods cannot reliably deliver.

8.3 The ANI Litigation and Domestic Legal Dynamics

ANI Media v. OpenAI, filed on 1 April 2025, is being watched closely as the first real test of Indian copyright law against AI training. The case arises under the Copyright Act 1957, whose 'fair dealing' exception in Section 52 is considerably narrower than US fair use, reaching mainly research, private study, criticism, and review. Whether training could shelter under Section 52 absent a licence — and how a mandatory blanket licence, if enacted, would sit alongside India's obligations under TRIPS and the Berne Convention — are live questions the case may help resolve. Should the court find that training infringes under existing law, pressure to legislate a licensing regime of some form would intensify.

8.4 India in Comparative Perspective

India's model occupies its own point on the spectrum. Where the United States relies on judicial resolution and bilateral bargaining, and the EU on transparency and opt-out within a market frame, India leans toward mandatory access and centralised rate-setting — closer, structurally, to the statutory licences that govern cable retransmission in some jurisdictions, with an administered tariff standing in for negotiation.

Whether that is wise depends on how one weighs the competing interests: developers' desire for scalable, affordable access; large publishers' interest in autonomy and fair value; small publishers' interest in compensation they could never negotiate alone; and the state's interest in a competitive domestic AI industry. A mandatory blanket licence serves all of these to some degree, but it does so by spending publisher autonomy — a price Indian media organisations have signalled they are unwilling to pay without significant changes. As one participant in a widely covered industry panel observed, if India's model succeeds it could well become a template for others (Times Network, 2025). Its large market and content base give it real standard-setting potential; but the no-opt-out provision in particular is likely to face sustained resistance from domestic publishers and international developers alike, and may require substantial revision before it can be implemented in a form stakeholders will accept.

09 Future Outlook

9.1 Consolidation and the Coming Renewal Cliff

The marketplace model may prove the most consequential development of the next three to five years, provided it attracts both supply and demand. A liquid, standardised venue would cut transaction costs, open the market to smaller publishers, and introduce genuine price competition. Its fate depends on network effects — and on whether the largest developers choose to participate or to keep pursuing bilateral deals on more favourable terms. Meanwhile, many landmark agreements signed in 2023–2025 run only two or three years, so a wave of renewals is due in 2026–2027. Whether renewal prices

© 2026 LIQVID

rise, as publishers grasp their leverage, or fall, as synthetic data and web-scraped alternatives mature, will depend heavily on how the US multidistrict litigation and EU enforcement play out.

9.2 The Shift from Training to Inference

The most serious long-run threat to publisher economics may lie not in training but in deployment. If AI assistants become the primary interface through which people get information — answering questions directly rather than routing readers to source sites — the advertising-funded model faces disruption that no training licence can fully offset. Several major publishers are already building AI-native products, from AI-powered search to interactive access to their archives, in an effort to sit inside the value chain of AI-mediated delivery rather than merely supply its raw material. For licensing, the implication is that inference-time compensation, paid for each AI response drawing on a publisher's content, may come to matter more than training licences — leaving publishers who signed training-only deals to renegotiate for inference use as the distinction sharpens.

9.3 Internationalisation and Fragmentation

The market is going global. Brazil, Australia, Canada, South Korea, and Singapore have all opened regulatory or legislative processes, and the rise of significant non-English models — Mistral in France, Sakana AI in Japan, Cohere in Canada — is generating fresh demand for non-Anglophone content and, with it, opportunities for regional-language publishers largely absent from the English-language market so far. The open question is whether this proliferation hardens into fragmentation, with every jurisdiction devising its own bespoke rules, or whether some minimum international standard emerges. A framework administered through a body such as WIPO, setting a floor for training-data transparency and rightsholder compensation while leaving room for national variation above it, would reduce cross-border complexity and disproportionately help the smaller publishers and developers least able to absorb it.

10 Conclusion

This paper has traced the emergence of AI content licensing in the global publishing industry — a market that moved from unlicensed extraction to structured commercial relationships in a compressed and turbulent span of three years. Four conclusions stand out.

First, the market is real, substantial, and growing fast, with disclosed deals in the hundreds of millions and projections approaching USD 1.6 billion by 2030. But it is structurally incomplete: it serves large, well-leveraged publishers effectively and the long tail of creators poorly, reflecting power asymmetries that markets alone are unlikely to correct without collective infrastructure or regulatory intervention.

Second, no single licensing model is universally right. Direct, collective, revenue-sharing, marketplace, and value-in-kind arrangements each carry distinct strengths and limits keyed to publisher scale, content type, and leverage. The most promising path is probably tiered: bilateral deals for the largest archives, marketplaces or collectives for the middle and the long tail, and statutory floors for content types where the market persistently fails.

Third, the legal landscape is fragmented and fast-moving. The US is resolving the question through litigation with still-unsettled doctrine; the EU has built a transparency-first regime with real

© 2026 LIQVID

enforcement power; the UK sits in a regulatory vacuum; and India has proposed a mandatory model without opt-out. This patchwork raises costs and uncertainty, again to the disadvantage of smaller players. An international framework, however hard to negotiate, would benefit the ecosystem as a whole.

Fourth, the underlying tension — between the property rights of creators and the innovation interests of developers — will not yield to market forces or regulation acting alone. A durable settlement will need clear legal rules, workable collective infrastructure, credible training-data transparency, and competitive market structures that stop any single party from dominating either content access or model deployment.

India's proposal is instructive less as a likely blueprint than as a mirror: it surfaces the full range of choices governments face — publisher autonomy against national ambition, market pricing against administered rates, opt-in against mandatory access. These are political-economy choices, not merely technical ones, and how different societies resolve them will shape not only the economics of publishing but the information architecture of the AI era. For researchers, the most urgent need remains empirical: rigorous work on actual deal economics — pricing, scope, attribution, and market effects — that is currently obscured by confidentiality. Qualitative study of publisher decision-making, comparative legal analysis of how regimes interact, and formal modelling of long-run equilibria under different licensing rules would all repay the effort.

© 2026 LIQVID

References

Ansari, M. (2025). *Publishers versus AI: The licensing debate* [Panel remarks]. Times Network. New Delhi, India.

Arrieta-Ibarra, I., Goff, L., Jiménez-Hernández, D., Lanier, J., & Weyl, E. G. (2018). Should we treat data as labor? Moving beyond "free." *AEA Papers and Proceedings, 108*, 38–42. https://doi.org/10.1257/pandp.20181003

Authors Guild. (2026, April). What authors need to know about the $1.5 billion Anthropic settlement. https://authorsguild.org

Bloomberg Law. (2026, January 5). *OpenAI must turn over 20 million ChatGPT logs, judge affirms*. https://news.bloomberglaw.com

Bruegel. (2025). *The European Union is still caught in an AI copyright bind* [Policy brief]. Bruegel. https://www.bruegel.org

Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).

Choe, S. K., Ahn, K., Panda, A., Kang, M., Xu, S., Wu, X., … Xing, E. P. (2024). *What is your data worth to GPT? LLM-scale data valuation with influence functions*. arXiv. https://doi.org/10.48550/arXiv.2405.13954

Clifford Chance. (2025, October). *Copyright compliance under the EU AI Act for GPAI model providers* [Client briefing]. https://www.cliffordchance.com

Copyright Alliance. (2026, January). *AI copyright lawsuit developments in 2025: A year in review*. https://copyrightalliance.org

Copyright Clearance Center. (2025). *2025 content usage trends*. Copyright Clearance Center.

Crémer, J., de Montjoye, Y.-A., & Schweitzer, H. (2019). *Competition policy for the digital era*. Report for the European Commission, Directorate-General for Competition. Publications Office of the European Union.

Department for Promotion of Industry and Internal Trade. (2025, December). *Working paper on AI content licensing: Part one*. Ministry of Commerce and Industry, Government of India.

Digital Content Next. (2025, March). AI content licensing lessons from Factiva and TIME: Emerging licensing models. https://digitalcontentnext.org

Digiday. (2025a, July). Perplexity launches publisher revenue-sharing program. https://digiday.com

Digiday. (2025b, November). AI royalties for small and midsize publishers: Collective licensing's next big play. https://digiday.com

Digiday. (2026a, January). A timeline of major deals between publishers and AI tech companies in 2025. https://digiday.com

Digiday. (2026b, January). Here are the biggest moments in AI for publishers in 2025. https://digiday.com

Encypher AI. (2026, January). *The publisher's guide to AI content licensing in 2026*. https://encypherai.com

Epstein, A. S., Haber, E., & Levendowski, A. (2023). *Copyright, AI, and a tiered licensing proposal* [Working paper]. SSRN.

European Parliament Research Service. (2025). *AI and copyright: The training of general-purpose AI* (PE 754.584). European Parliament. https://europarl.europa.eu

Grand View Research. (2024). AI datasets & licensing for academic research and publishing market size, share & trends analysis report, 2024–2030. Grand View Research.

© 2026 LIQVID

Grimmelmann, J. (2024). Training data, copyright, and fair use: What makes AI training transformative? *Columbia Journal of Law & the Arts, 47*(2), 145–179.
Harvard Law Review. (2024). NYT v. OpenAI: The Times's about-face. *Harvard Law Review Forum*. https://harvardlawreview.org
Harvard Law School. (2024). *Does ChatGPT violate New York Times' copyrights?* Harvard Law School News. https://hls.harvard.edu
Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskiy, E., Cai, T., Rutherford, E., … Sifre, L. (2022). *Training compute-optimal large language models*. arXiv. https://doi.org/10.48550/arXiv.2203.15556
Jones, C. I., & Tonetti, C. (2020). Nonrivalry and the economics of data. *American Economic Review, 110*(9), 2819–2858. https://doi.org/10.1257/aer.20191330
Kadrey v. Meta Platforms, Inc., No. 23-cv-03417 (N.D. Cal. 2025).
Kluwer Copyright Blog. (2025, October). *EU copyright law roundup Q2–Q3 2025*. https://legalblogs.wolterskluwer.com
Lemley, M. A., & Casey, B. (2024). Fair learning. *Texas Law Review, 102*(3), 700–756.
MediaAndTheMachine. (2025, January). *5 takeaways from the AI content licensing deals I'm tracking* [Substack newsletter]. https://mediaandthemachine.substack.com
New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. 2023).
New York Times Co. v. Tasini, 533 U.S. 483 (2001).
Norton Rose Fulbright. (2026). *AI in litigation series: Update on AI copyright cases in 2026*. https://www.nortonrosefulbright.com
Press Gazette. (2025–2026). *Who's suing AI and who's signing* [Rolling tracker]. https://pressgazette.co.uk
ProMarket. (2025, November). *Content licensing agreements will concentrate markets without standardized access*. Stigler Center, University of Chicago. https://promarket.org
Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
Rochet, J.-C., & Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association, 1*(4), 990–1029. https://doi.org/10.1162/154247603322493212
Samuelson, P. (2023). Generative AI meets copyright. *Science, 381*(6654), 158–161. https://doi.org/10.1126/science.adi0656
Scott, W. (2025, October). How AI licensing deals determine search visibility in 2025: Academic publishers and the marketplace model. https://willscott.me
Sony Corp. of America v. Universal City Studios, Inc., 464 U.S. 417 (1984).
TechJack Solutions. (2026, April). Three copyright regimes, three diverging answers: US, EU, UK divergence in AI licensing law. https://techjacksolutions.com
Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc., No. 1:20-cv-00613 (D. Del. 2025).
Times Network. (2025). *Publishers vs. AI: The licensing debate* [Television panel discussion]. Times Network, India.

© 2026 LIQVID

About Liqvid

AI-Powered Learning Solutions for the Modern Learner & Educator

Liqvid is a leading EdTech company revolutionizing the way people learn with a particular emphasis on English and language learning. With over 20 years of experience, we empower publishers, educational institutions and government organizations with AI-powered learning solutions that deliver engaging, personalized learning experiences.

OUR OFFERINGS

BOOK PLUS

AI Learning LMS

Publisher content delivered through an interactive LMS with online & offline access.

INTERACTIVE EBOOK

Rich Digital Content

Interactive PDFs with multimedia, assessments, and AI-powered test generation.

SMART EPUBS

Enhanced Reading

EPUBs with audio, quizzes, summaries, and multilingual learning support.

TESTPREP

Exam Preparation

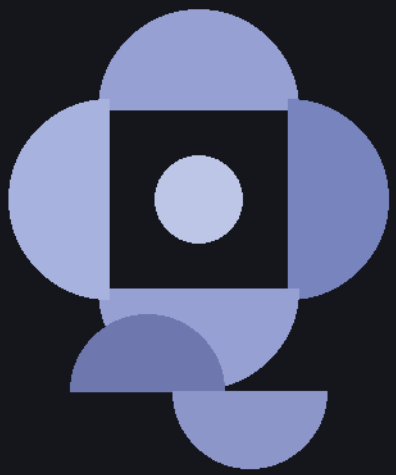
Adaptive test prep powered by Wiley content, analytics, and personalized practice.

India Liqvid House, F-3, Sector 8,
Noida, U.P. 201301

WWW.LIQVID.COM
info@liqvid.com

LIQVID

“Publishers' archives are no longer a record of the past. They are the raw material of artificial intelligence.”



Corporate Office

Liqvid House, F-3, Sector 8,
Noida, U.P. 201301

info@liqvid.com
+91 8448360148
www.liqvid.com