NYU | LAW

Journal of Intellectual Property & Entertainment Law

VOLUME 15 NUMBER 2

Statement of Purpose

Consistent with its unique development, the New York University Journal of Intellectual Property & Entertainment Law (JIPEL) is a nonpartisan periodical specializing in the analysis of timely and cutting-edge topics in the world of intellectual property and entertainment law. As NYU's first online-only journal, JIPEL also provides an opportunity for discourse through comments from all of its readers. There are no subscriptions or subscription fees; in keeping with the open-access and free discourse goals of the students responsible for JIPEL's existence, the content is available for free to anyone interested in intellectual property and entertainment law.

The New York University Journal of Intellectual Property & Entertainment Law is published up to three times per year at the New York University School of Law, 139 MacDougal Street, New York, New York, 10012. In keeping with the Journal's open access and free discourse goals, subscriptions are free of charge and can be accessed via www.jipel.law.nyu.edu. Inquiries may be made via telephone (212-998-6101) or e-mail (submissions.jipel@gmail.com).

The Journal invites authors to submit pieces for publication consideration. Footnotes and citations should follow the rules set forth in the latest edition of The Bluebook: A Uniform System of Citation. All pieces submitted become the property of the Journal. We review submissions through Scholastica (scholasticahq.com) and through e-mail (submissions.jipel@gmail.com).

The Journal is also available on WESTLAW, LEXIS-NEXIS and HeinOnline.

NEW YORK UNIVERSITY

JOURNAL OF INTELLECTUAL PROPERTY AND ENTERTAINMENT LAW

VOLUME 15 SPRING 2026 NUMBER 2

TRAINING DATA GOVERNANCE

FRANK FAGAN*

As AI-generated summaries increasingly displace traditional search results, users are less likely to visit the underlying websites where content is published. This shift has sharply reduced traffic to those sites, threatening the economic viability of content creators and prompting a wave of paywalls, restrictions, and litigation. With referral-based revenue in decline, the continued supply of high-quality content faces mounting risk, precisely as generative AI has grown more dependent on such material. This tension, between innovation and sustainability, frames the central legal and policy inquiry of training data governance: how to preserve access to essential AI training inputs without undermining the incentives to produce them.

This Article examines licensing as a tool of training-data governance and focuses on the practical question of when content loss threatens model performance and reduces social welfare. The inquiry centers on whether withdrawal of high-value material is likely and whether voluntary bargaining can realistically prevent it. Where the risk of withdraw is low, additional protection is unnecessary; where the risk is substantial and bargaining fails, a narrowly tailored fallback, such as a standardized, non-exclusive license, can preserve access without disturbing fair-use doctrine or existing private arrangements. This welfare logic is implemented through a three-part test: licensing is warranted only when (1) the content has demonstrable value for AI training, (2) withdrawal is the rational market outcome absent remuneration, and (3) voluntary licensing fails due to transaction costs or bargaining frictions. Together, these conditions ensure that intervention occurs only where it improves overall welfare relative to the status quo.

* Professor of Law, South Texas College of Law Houston. Email: ffagan@stcl.edu.

INTRODUCTION

In June 2025, *The Wall Street Journal* reported a sharp and accelerating decline in referral traffic to online publishers.[1] Major outlets such as *HuffPost*, *Business Insider*, and *The Washington Post* had lost half or more of their search-driven visits over three years.[2] This erosion coincided with the rise of AI-generated summaries in place of traditional search results.[3] Where users once typed a query,

[1] Isabella Simonetti & Katherine Blunt, *News Sites Are Getting Crushed by Google's New AI Tools*, WALL ST. J., Jun. 10, 2025, https://www.wsj.com/tech/ai/google-ai-news-publishers-7e687141 [https://perma.cc/2L55-HR9V].

[2] *Id.*

[3] According to one study, adding AI-generated summaries to search results cuts the likelihood of users clicking through to a source by about half. *See* Athena Chapekis & Anna Lieb, *Google Users Are Less Likely to Click on Links When an AI Summary Appears in the Results*, PEW RSCH. CTR. (July 22, 2025),

scanned snippets, and clicked through to original sites, they are now often presented with an AI-composed answer that is condensed, polished, and complete enough to obviate the visit.[4]

For many publishers, this shift is not merely inconvenient; it is destabilizing. Referral traffic underwrites both advertising revenue and subscription pipelines.[5] A sustained collapse in that traffic forces difficult adjustments, such as reductions in staff, consolidation of coverage, migration behind paywalls, or a pivot to alternative products such as branded apps and events.[6] The pattern extends well beyond news. Independent bloggers, niche review sites, technical explainers, and specialized forums—many of which are run by small teams or individuals—face the same pressure.[7] In each case, the audience still wants the content, but increasingly consumes it through an intermediary that returns little or nothing to the original source.

The scale of this redirection has grown markedly. A decade ago, Google crawled roughly two pages for every visit it sent back to a publisher.[8] By late 2024, those ratios had become unrecognizable: Google was at 18:1, OpenAI at 1,500:1, and Anthropic at 60,000:1.[9] This means that for every user a large language model sends to a publisher, it may have already consumed multiple orders of magnitude more content from the same source. Today the web remains rich in high-quality material, but for many creators, the economics of keeping it open have changed materially.

Generative AI systems depend on that openness. Large language models (LLMs) are trained on vast and varied datasets including books, news articles,

https://www.pewresearch.org/short-reads/2025/07/22/google-users-are-less-likely-to-click-on-links-when-an-ai-summary-appears-in-the-results/ [https://perma.cc/M2DP-VFF6].

[4] *See id.* (observing that approximately sixty percent of respondents reported seeing an AI-generated summary in Google search results in March 2025).

[5] *See* Simonetti & Blunt, *supra* note 1.

[6] *Id.*

[7] *See* Alex Bocharov, *Declare Your AIndependence: Block AI Bots, Scrapers and Crawlers With A Single Click*, The Cloudfare Blog (July 3, 2024), https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click/ [https://perma.cc/AY57-LC5B].

[8] Christine Wang, *Publishers Facing Existential Threat from AI, Cloudflare CEO Says*, Axios (Jun. 19, 2025), https://www.axios.com/2025/06/19/ai-search-traffic-publishers [https://perma.cc/X6AX-DJCL].

[9] *Id.*

technical documentation, forum discussions, and more.[10] In the United States, this training has proceeded under a broad reading of the fair use doctrine, which permits copying and reuse for transformative purposes without prior authorization.[11]

This permissive environment has been central to U.S. leadership in AI, lowering entry costs and enabling both startups and incumbents to train powerful general-purpose and domain-specific systems at scale.[12]

Until recently, the principal constraint on training was legal uncertainty, especially regarding the scope of fair use.[13] These doctrinal issues have been

[10] Large language models reflect both the quality and quantity of their underlying data. *See* Jared Kaplan et al., *Scaling Laws for Neural Language Models*, ArXiv (2020) (noting the importance of quantity), https://arxiv.org/pdf/2001.08361 [https://perma.cc/LJ9V-CYYB]; Ming Li et al., *From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning*, ArXiv (2023) (noting the importance of quality), https://arxiv.org/pdf/2308.12032 [https://perma.cc/V6EJ-4W5V].

[11] *See* Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 Tex. L. Rev. 743, 745–46 (2021) (observing that a permissive fair use regime has facilitated the rapid development of large language models). Fair use turns on four statutory factors: "(1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes [and whether it is transformative]; (2) the nature of the copyrighted work; (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and (4) the effect of the use upon the potential market for or value of the work." 17 U.S.C. § 107. No single factor is dispositive. *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508 (2023). The doctrine's application is famously uncertain. Lemley and Casey quote Lawrence Lessig's remark that fair use is "the right to hire a lawyer." Lemley & Casey, *supra* 11, at 763 (quoting Lawrence Lessig, Free Culture 187 (2004)).

Courts evaluating LLM training could, for example, conclude under factor one that training serves a distinct, transformative purpose because the model learns from the articles rather than reproducing them to substitute for subscriptions. Under factor two, courts may characterize news articles as primarily factual and therefore more amenable to permissible use. Under factor three, courts may find that machine learning relies on textual features such as token frequencies or word co-location rather than full expressive works. Under factor four, courts could determine that publishers such as *The New York Times* do not market their content for LLM training and that such training does not impair core revenue streams or intended audiences. These possibilities illustrate the substantial flexibility courts retain to find that LLM training constitutes fair use.

[12] *See* Joshua Levine & Tim Hwang, Copyright, AI, and Great Power Competition 4 (2025) (noting that the development of American frontier models has relied heavily on free access to online data, which permissive copyright rules enable), https://cdn.sanity.io/files/d8lrla4f/staging/1e8b069431e03d9e8b23f03dede131ef5fdb16a9.pdf [https://perma.cc/YBN6-R2W9].

[13] Notable litigation is ongoing. *See, e.g.*, Complaint at 1, The New York Times Co. v. OpenAI, Inc., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023) (alleging unauthorized use and reproduction of NYT articles in ChatGPT outputs); In re OpenAI, Inc. Copyright Infringement Litig., 776 F. Supp. 3d 1352 (U.S. J.P.M.L. 2025) (coordinating author class actions alleging infringement in training data); Complaint at 1, Ziff Davis, Inc. v. OpenAI, Inc., No. 1:25-cv-00501 (D. Del. Apr. 24, 2025) (claiming unauthorized systematic scraping of journalism content); Complaint at 1, Authors Guild v. OpenAI, Inc., No. 1:23-cv-08292 (S.D.N.Y. Sept.

widely analyzed.[14] Far less attention has been paid to a parallel problem: practical gating, where high-value content is increasingly constrained for reasons unrelated to copyright and in ways that doctrine alone cannot resolve.

This erosion appears in the gradual loss of the abundant, openly accessible content that fueled the first wave of generative AI.[15] Facing declining engagement and limited return from AI-driven summaries, content producers are adopting defensive measures: blocking web crawlers, placing archives behind paywalls, throttling API access, or entering exclusive licensing deals.[16] These responses are economically rational. When the marginal return from keeping content public falls

19, 2023) (claiming copyright infringement by scraping plaintiffs' works of fiction); Complaint at 1, The Intercept Media, Inc. v. OpenAI, Inc., No. 1:24-cv-01515 (S.D.N.Y. Feb. 28, 2024) (alleging violations of the Digital Millennium Copyright Act of 1998 as a result of stripping author, title, copyright, and terms of use information from plaintiff's protected works); Complaint at 1, Raw Story Media, Inc. v. OpenAI, Inc., No. 1:24-cv-01514, (S.D.N.Y. Feb. 28, 2024) (same).

[14] There are many voices. *See, e.g.*, BJ Ard, *Copyright's Latent Space: Generative AI and the Limits of Fair Use*, 110 Cornell L. Rev. 509, 561 (2025) (noting that fair use should protect an author's chosen market); Olivia S. Hiltbrand, *Guarding the News Media's Intellectual Property in the Age of Generative AI*, 28 Stan. Tech. L. Rev. 35, 40 (2025) (suggesting that news outlets should be strongly protected because of their unique societal role); Frank Pasquale & Haochen Sun, *Consent and Compensation: Resolving Generative AI's Copyright Crisis*, 110 Va. L. Rev. Online 207, 207 (2024) (proposing an opt-out mechanism that permits copyright holders to prohibit non-consensual use of their work); Peter Henderson et al., *Foundation Models and Fair Use*, 24 J. Mach. Learning & Rsch. 1, 2 (2023) (suggesting that law could permit fair use when strong technical tools are used by AI developers to mitigate copyright infringement); David Atkinson, *Unfair Learning: GenAI Exceptionalism and Copyright Law*, ArXiv (2025) (arguing that consistent application of liberal fair use for Generative AI would require that no human ever pay for copyright work), https://arxiv.org/pdf/2504.00955 [https://perma.cc/PY33-U6EG]; Lemley & Casey, *supra* note 11, at 748–50 (noting that generative AI typically transforms the data that it ingests and that copyright law should permit its free use in most circumstances).

[15] *See* Bocharov, *supra* note 7 (reporting that of the bottom 10,000 websites in terms of internet traffic, more than 18% were blocking scrapers). The erosion of open access to content has not been confined to small producers. For example, Reddit, Stack Overflow, and X have gated their content. *See* Adam Levine, Tae Kim & Angela Palumbo, *Google Search Is Fading*, Barron's, June 16, 2025, at 16–17. And publishing executives are noting that "blocking [LLMs from scraping] should be publishers' first line of defense when planning for the expected decline in search traffic from the transition to generative search." Peter Brown & Klaudia Jaźwińska, *Journalism Zero: How Platforms and Publishers are Navigating AI*, Colum. Journalism Rev., May 15, 2025, at 42, https://towcenter.columbia.edu/sites/towcenter.columbia.edu/files/content/Journalism%20Zero_%20How%20Platforms%20and%20Publishers%20are%20Navigating%20AI_0.pdf [https://perma.cc/A2FJ-GLQ9].

[16] *Id.*

below the cost of production, withdrawal becomes a sustainable choice, so long as enough people are willing to pay for the content.[17]

This development represents a form of gating by other means. Fair use remains unchanged, but the incentives that once kept high-value material publicly accessible are weakening. The resulting constraints arise not from court orders or statutory amendments, but from private infrastructure control, platform design, and market leverage.[18] The practical availability of training data is shrinking in ways that doctrine alone cannot prevent.

Not all content is equally vulnerable. A substantial portion of LLM training material consists of *baseline content*: works produced under conditions that make withdrawal unlikely irrespective of AI reuse.[19] Examples include academic research, open-source documentation, government publications, and commercial journalism from well-capitalized outlets.[20] Institutional mandates, diversified revenue, reputational incentives, and legal obligations support their continued openness.[21]

The more immediate fragility lies with *at-risk content*: high-quality, domain-specific material produced by economically marginal creators without institutional support and whose continued publication depends on monetizing user attention or access.[22] These sources play an outsized role for model diversity, edge-case fluency, and temporal responsiveness,[23] yet they are the first to suffer when AI intermediates the relationship with their audience.[24]

---

[17] The standard law and economics copyright model recognizes as much. *See* William M. Landes & Richard A. Posner, *An Economic Analysis of Copyright*, 18 J. Legal Stud. 325, 326 (1989).

[18] *See infra* Part I.B.

[19] *See infra* Part II.A.

[20] *See* Rebecca Tushnet, *Copy This Essay: How Fair Use Doctrine Harms Free Speech and How Copying Serves It*, 114 Yale L.J. 535, 541 (2004).

[21] *Id.*

[22] *See infra* Part II.B.

[23] *Cf.* Edward Lee, *Fair Use and the Origin of AI Training*, 63 Hou. L. Rev. 105, 158–59 (2025) (observing that model performance depends on both the magnitude and the diversity of its training inputs).

[24] Consider the recent experience of one technology blog author who posted their dissatisfaction on Reddit:

> I have my own personal tech blog, and used to get a fairly large number of hits per month from search, but noticed it's been pretty steeply declining the past 6 months. I did some research recently using CloudFlare's AI audit tools, and found where it's going. It's almost all now traffic that comes through ChatGPT user queries or similar.

Between these poles lies *transitional content*: material still produced under stable-enough conditions, but increasingly exposed to the same pressures that drive at-risk creators toward withdrawal.[25] Transitional content functions as an early indicator, signaling where changes in platform economics or user behavior may soon convert stable producers into at-risk ones.

Against this backdrop, the policy question becomes how to preserve access to high-value but economically fragile content without overregulating sources that remain sufficiently stable. Existing academic proposals offer only partial answers. Licensing of training data remains a relatively young and evolving area of scholarship. Proposals by Pasquale, Sun, and others have brought needed attention to questions of fairness, labor value, and the distribution of benefits in AI development.[26] This Article builds on those contributions while taking a different analytic path. Rather than treating all training data as a single category, it emphasizes the heterogeneity of incentives across content types and grounds the case for licensing in situations where content loss poses a measurable risk to social welfare.

On this view, the central objective of training-data policy is to maximize overall joint welfare, which encompasses the interests of AI developers, content producers, and end users. Achieving that aim requires distinguishing among content types and targeting intervention only where content withdrawal is the predictable market outcome rather than relying on a uniform entitlement to remuneration. The licensing regime proposed here therefore activates only when three conditions are satisfied:

> Fairly clear measurable switch of traffic from Google landing on my site, to traffic instead going to ChatGPT. *It's gotten to the point where my user count on the site has dropped over 50%*, from people using the Google AI summary or ChatGPT search results instead, both of which just serve the information from my site.

Comment posted by u/Me4502, Reddit (r/technology) *New Sites Are Getting Crushed by Google's New AI Tools* (2025) (emphasis added), https://www.reddit.com/r/technology/comments/1l7xnaz/news_sites_are_getting_crushed_by_googles_new_ai/ [https://perma.cc/ZCX3-CE3L].

[25] *See infra* Part II.C.

[26] *See* Pasquale & Sun, *supra* note 14, at 231–36; Roy S. Kaufman, *Responsible AI Starts with* Licensing, 48 Colum. J.L. & Arts 403 (2025); Christophe Geiger & Vincenzo Iaia, *The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI*, Comput. L. & Sec. Rev., Apr. 2024, at 1, 6; Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 91–93 (2017).

(1) the content has demonstrable value for AI training;

(2) the producer operates at the economic margin such that withdrawal is the rational outcome absent remuneration; and

(3) voluntary licensing fails due to transaction costs, bargaining asymmetries, or collective action problems.

The framework is designed as a fallback rather than a foundation. Fair use remains the default rule for baseline content, and at-risk content producers are free to opt out and continue to post their content publicly. In this way, the licensing regime sets a price floor that at-risk producers can choose to accept.

The remainder of the Article proceeds in four Parts. Part I describes the shift from doctrinal uncertainty to practical gating and shows how litigation, platform enclosure, and deployment-layer exclusivity restrict access without changing the underlying fair use framework. Part II develops the trigger conditions for contingent licensing and the resultant three-tier structure of baseline, transitional, and at-risk content. Part III explains why licensing, if used, should remain narrowly tailored, drawing lessons from other compulsory regimes and avoiding premature intervention in baseline material. Part IV evaluates legal and institutional design options, including statutory, administrative, judicial, and hybrid approaches, and outlines safeguards to protect voluntary licensing and baseline access. The Article concludes that contingent licensing, applied sparingly and grounded in clear evidence of market failure, can help stabilize the content ecosystem while preserving the openness that has enabled AI innovation. Above all, the Article demonstrates that training data governance provides the conceptual tools needed to address problems that narrow doctrinal analysis alone cannot resolve.

I
Forms of Gating

Even without any change to fair use, the effective openness of AI training inputs is shrinking.[27] Content that was once easily accessible to large language models is increasingly subject to legal, technical, and strategic barriers.[28] This Article refers to these constraints as gating: a reduction in practical accessibility

[27] *Infra* Parts I.A–D.
[28] *Id.*

irrespective of formal copyright rule. Rather than altering doctrine, gating shifts the factual conditions under which doctrine operates, narrowing the pool of usable training data through upstream access decisions.

The risk is structural. Fair use may remain a robust defense, but it does not guarantee that socially valuable inputs will be available in the first instance. Rights holders can (and increasingly do) control access upstream, long before any fair use argument is invoked.[29] The result is an ecosystem in which training inputs are determined not only by what copyright law permits, but by who can access the data and on what terms.

This shift matters because the scale and breadth of training inputs directly influence the performance, competitiveness, and epistemic diversity of LLMs.[30] Consider, for instance, that the largest models rely on trillions of tokens drawn from across the public web,[31] yet recent analyses of Common Crawl datasets show that only a small fraction of the internet is consistently captured, with domain-level coverage varying markedly from month to month.[32] If major sources withdraw or gate content, those inputs may disappear from training corpora altogether, creating blind spots that can persist across successive model generations.

Gating now takes several distinct forms, often reinforcing one another:

1. *Litigation-based gating*—where legal action, or the credible threat of it, creates uncertainty that deters use regardless of ultimate merits.

[29] Examples are given below at Parts I.B–C.

[30] *See* Lee, *supra* note 23, at 158–59.

[31] For example, Google's PaLM 2 was trained with 3.6 trillion tokens. *See* Jennifer Elias, *Google's Newest A.I. Model Uses Nearly Five Times More Text Data for Training Than Its Predecessor*, CNBC (May 17, 2023, at 11:52 ET), https://www.cnbc.com/2023/05/16/googles-palm-2-uses-nearly-five-times-more-text-data-than-predecessor.html [https://perma.cc/8Q8Z-AN4V]. Even earlier models trained with large swaths of the internet. For instance, GPT-3 was trained on data sourced from Common Crawl (a large-scale snapshot of the web), WebText2 (a filtered collection of Reddit posts), Books1 and Books2 (collections of publicly available books), and the English-language version of Wikipedia. *See* Tom B. Brown et al., *Language Models Are Few-Shot Learners*, ArXiv, at 9 (2020), https://arxiv.org/pdf/2005.14165 [https://perma.cc/7GF8-236U].

[32] Common Crawl emphasizes that it does not crawl the entire web, "nor even a representative sample of it." Stefan Baack, A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl 1 (2024), https://facctconference.org/static/papers24/facct24-148.pdf [https://perma.cc/ZL7B-Q76V]. While inclusion in a crawl depends on a score assigned to a webpage, and that the score depends in part upon an internal monthly quota established by Common Crawl, persistent and patterned gating deprives the population of data that can be selected for crawling. *Id.* at 5.

2. *Platform enclosure*—where intermediaries controlling major data streams impose technical or contractual barriers to access.
3. *Deployment-layer exclusivity*—where the most valuable outputs of LLMs depend on proprietary datasets unavailable to competitors, effectively limiting access through downstream integration.
4. *Market-wide fragmentation*—where the cumulative effect of dispersed and inconsistent access restrictions distorts competition and reduces the diversity of training data.

Each of these forms can operate independently, but their combined effect is to erode the predictability and openness that characterized the early generative AI ecosystem. Understanding how they arise and interact is essential for designing policy interventions that preserve the advantages of open training while respecting legitimate incentives for content creation. Each is addressed in turn below.

A. Litigation-Based Gating

Litigation has become one of the most visible mechanisms through which content producers challenge AI training on their works.[33] Over the past two years, authors, visual artists, news organizations, and others have filed a wave of suits alleging copyright infringement, digital piracy, or related claims against developers of large language models and image generators.[34] Some plaintiffs seek damages; others aim to secure declaratory judgments or injunctions aimed at limiting or reshaping how training occurs.[35]

The legal merits of these cases remain uncertain. No court has yet issued a definitive ruling on whether using copyrighted works to train a generative model constitutes infringement or falls within fair use. Many claims face substantial doctrinal hurdles, particularly given the longstanding acceptance of intermediate copying for transformative uses in other contexts.[36] But the immediate effects

[33] For a non-exhaustive sample of ongoing suits, see *supra* note 13.

[34] *See id.* (collecting complaints that include such claims).

[35] *Id.* In some cases, damages may be marginally preferred to injunctions. Consider that the toleration of scraping can serve multiple strategic aims. It can reinforce reputation and agenda-setting power, increase referral traffic through citations, build bargaining leverage by making models reliant on their content, and signal openness to future licensing or partnership opportunities.

[36] *See* Perfect 10, Inc. v. Amazon.com, Inc., 508 F.3d 1146, 1165 (9th Cir. 2007) (holding that Google's use of thumbnail images in an image search engine was transformative because it served a different function than

extend beyond the courtroom. For many rights holders, filing suit functions as a form of practical gating. It is a method for signaling objection to AI training, deterring prospective users, and occasionally resulting in settlements that restrict use or impose licensing obligations.[37]

This use of litigation as a gating mechanism reflects a broader dynamic in intellectual property enforcement. Even when plaintiffs cannot secure a clear judicial victory, the prospect of prolonged litigation that entails discovery burdens, legal expense, and reputational risk can make negotiated resolution the rational choice for defendants.[38] In the generative AI context, several disputes have reportedly ended in licensing agreements or the removal of contested datasets from training corpora.[39] Such resolutions may resolve the controversy between the parties, but they also reshape the broader training ecosystem: the contested material exits the open pool, and other rights holders may be encouraged to bring similar claims in anticipation of comparable concessions.[40]

---

the original photographs); Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 572, 579–82 (1994) (holding that a commercial parody of Roy Orbison's song "Oh, Pretty Woman" was a transformative use, even though it borrowed heavily from the original); *compare* Princeton Univ. Press v. Michigan Document Servs., Inc., 99 F.3d 1381, 1389–91 (6th Cir. 1996) (holding that the commercial reproduction of coursepack materials for sale to students was not transformative and thus not fair use); Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 521–26 (2023) (holding that Warhol's licensing of a silkscreen image derived from Goldsmith's photograph for use in a commercial magazine was not sufficiently transformative to qualify for fair use).

[37] *See* Mark A. Lemley & Carl Shapiro, *Patent Holdup and Royalty Stacking*, 85 Tex. L. Rev. 1991, 1991 (2007) (describing the strategic use of litigation in patent licensing negotiations); Shyamkrishna Balganesh, *The Uneasy Case Against Copyright Trolls*, 86 S. Cal. L. Rev. 723, 723–24 (2013) (describing the strategic use of litigation in the copyright context).

[38] *Cf.* Ben Depoorter, *Technology and Uncertainty: The Shaping Effect of Copyright Law*, 157 U. Pa. L. Rev. 1831, 1837, 1842–43 (2009). Depoorter documents how litigation by prominent rights holders in response to technological disruptions—including the audio cassette, VCR, CD, DAT, DVD, MP3, and file sharing—often took years to resolve and created substantial uncertainty. Within a rational choice framework, such uncertainty translates into expected costs from discovery, legal fees, and reputational harm, and can encourage defendants to settle as a precaution.

[39] *See* Tori Noble, *Copyright and AI: Cases and Consequences*, Electronic Frontier Foundation (Feb. 19, 2025), https://www.eff.org/deeplinks/2025/02/copyright-and-ai-cases-and-consequences [https://perma.cc/ET5X-LDNW] (describing pending suits as a struggle for leverage in settlement negotiations).

[40] This is because other rights holders' expectations of success are influenced by the outcomes of prior litigation. While they may not know a particular defendant's reservation price, i.e., the maximum they are willing to pay to settle, observing similarly situated plaintiffs can provide useful signals. On the dynamics of strike suits, *see generally* Lucian Bebchuk, *Suing Solely to Extract a Settlement Offer* (Nat'l Bureau of Econ. Rsch., Working Paper No. 2161, 1987).

Litigation also performs a signaling function for regulators and the broader market. By characterizing unlicensed training as an appropriation of creative labor, plaintiffs can shift public opinion toward stricter controls.[41] In this sense, litigation operates as both a legal and political act that is capable of shaping legislative debates, informing administrative rulemaking, and influencing emerging industry norms. As additional suits are filed, the perceived legal risk of unlicensed training increases, even if the underlying doctrine remains unchanged.

B. Platform Enclosure: The Grok Model

A second, and more immediate, source of gating arises from platform owners that control large, continuously updated streams of user-generated content. These platforms, such as social media networks, discussion forums, and other high-traffic intermediaries, occupy a distinctive position in the data supply chain. With a change in terms of service or access permissions, they can shift material from an open or semi-open environment into a closed, proprietary one.

The launch of X's "Grok" assistant illustrates this shift. In connection with Grok's debut, X announced that the assistant would be trained on the platform's full corpus of user posts, but that this data would no longer be broadly available to developers via the API.[42] A resource that had functioned as a shared, public-facing input thus became a proprietary training asset reserved for the platform's own systems. Reddit's decision to restrict API access to third-party developers reflects a similar move, effectively enclosing user-generated content that had long been available for external analysis and AI training.[43] Although permissible under

[41] *Cf.* Pasquale & Sun, *supra* note 14, at 231–36 (considering the value of content in terms of labor and its unlicensed use as a form of unjust enrichment).

[42] *See* Emma Roth, *X's New Policy Prevents Companies from Using Posts to 'Fine-tune or Train' AI Models*, The Verge (Jun. 5, 2025), https://www.theverge.com/news/680626/x-ai-training-ban-posts [https://perma.cc/P6MR-ALRX]; *see also* Edge8, *Elon Musk Twitter Data Strategy: The Real AI Play Behind X Acquisition*, LinkedIn (June 2025), https://www.linkedin.com/posts/edge8ai_elon-musk-twitter-data-strategy-the-real-activity-7333704827778048000-X-5W/ [https://perma.cc/8VFS-YY8Y] (asserting that Musk secured a human communication dataset that includes 400 million daily conversations in order to develop proprietary AI that comprehends customer sentiment, market dynamics, and cultural trends with accuracy).

[43] Reddit currently earns about nine percent of its revenue from licensing its data to OpenAI. Levine, Kim & Palumbo, *supra* note 15, at 16–17. Reddit hosts human discussions of many topics. Its CEO believes that "human intelligence [as reflected in discussion] is still going to be worth a lot, and it's going to go up and up in value." *Id.* at 17.

existing copyright and contract rules, these actions exemplify enclosure in the generative AI context, i.e., the withdrawal of a once-common input from the open pool, conferring competitive advantage on the actor controlling the chokepoint.

Enclosure has at least three important consequences. First, it concentrates training advantages in the hands of incumbent platforms, which may already benefit from network effects and other sources of market power. Second, it reduces the diversity of training inputs available to other developers, potentially skewing model behavior toward the linguistic and cultural patterns that are prevalent on whatever platforms remain accessible.[44] Third, it may accelerate the adoption of vertically integrated AI strategies, in which content capture, model development, and product deployment occur within a single corporate structure.[45] Such integration not only forecloses rivals' access to key inputs but also narrows public visibility into how those inputs are used.[46]

Although the motives for enclosure vary, from monetizing data through proprietary AI products to avoiding the costs of serving high-volume API traffic, the strategic logic is consistent. By converting publicly visible streams into proprietary reservoirs, platforms insulate themselves from competitive training, slow the diffusion of advanced capabilities, and ensure that any residual licensing occurs on terms they dictate.

C. Technical Gating: The Microsoft Copilot Model

Not all gating strategies rely on legal or contractual barriers; some operate at the technical layer, embedding access controls directly into the systems that deliver content. Microsoft's Copilot products illustrate this approach. Although marketed as an AI assistant within the Microsoft 365 suite, Copilot's architecture can be configured to mediate, and selectively restrict, how its models access corporate

[44] *See* Tshilidzi Marwala, *The Commercial Importance of Diversity in Generative AI*, Daily Maverick, Aug. 21, 2024, https://www.dailymaverick.co.za/opinionista/2024-08-21-the-commercial-importance-of-diversity-in-generative-ai/ [https://perma.cc/Q9UW-ZWD3] (emphasizing the commercial risks of uniformity).

[45] Marwala observes that developers can respond to vertical integration by assembling cross-disciplinary teams and incorporating underrepresented voices. *Id.* Yet vertical integration may offer a stronger competitive position, creating a moat that deters entry and confers an advantage over existing competitors.

[46] *Cf.*, Recent Proposed Legislation, *Platform Accountability and Transparency Act, S. 1876, 118th Cong. (2023)*, 137 Harv. L. Rev. 2104, 2105 (2024) (oting that transparency can play an important role in supporting accountability for how platforms use data).

or institutional data.[47] These controls are framed as mechanisms for enforcing privacy, security, and compliance obligations,[48] but they also generate a form of technical enclosure. Content that could, in principle, support general-purpose training is rendered accessible only within a tightly controlled environment, with usage parameters determined unilaterally by the platform provider.

The gating effect arises from the interplay between proprietary infrastructure and closed data pathways. Copilot's integration into Microsoft's cloud ecosystem ensures that access to documents, emails, or other materials passes through Azure's permissioning framework.[49] That framework can block export to non-Microsoft environments or prevent materials from being ingested into models outside the sanctioned stack. While these design choices serve legitimate security purposes, they also restrict opportunities for independent developers to work with the same content, thereby consolidating training advantages within Microsoft's AI ecosystem.

Similar patterns have emerged in other domains. In legal services, for example, the AI startup Harvey has partnered with major law firms to deliver generative AI capabilities trained on firm-specific knowledge. Harvey's deployment model keeps that knowledge within closed technical and contractual boundaries, ensuring that it benefits only participating firms and the vendor's own systems.[50] Other providers, such as Bloomberg with BloombergGPT and Google with its internal model sandboxing environments, use comparable strategies to retain exclusive access to high-value, domain-specific datasets.[51]

---

[47] *See* Microsoft, *Data, Privacy, and Security for Microsoft 365 Copilot* (Jan. 7, 2026), https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-privacy [https://perma.cc/6SU5-W4K6] (noting that while prompts, responses, and data accessed through Microsoft Graph are not used to train foundation LLMs, Copilot conditions outputs on user-specific organizational data).

[48] *Id.*

[49] *Id.*

[50] *See* Bob Ambrogi, *Harvey AI To Move Out Of Early Access Phase, Release More Affordable Versions Of Its Custom AI Models*, LawSites (May 1, 2024), https://www.lawnext.com/2024/05/harvey-ai-to-move-out-of-early-access-phase-release-more-affordable-versions-of-its-custom-ai-models.html [https://perma.cc/PKM8-7PVK] (explaining that Harvey is developing a platform enabling firms to securely train generative AI systems on their private data, integrate them with existing legal technology and workflows, and continually refine them through feedback from their legal workforce). This product strategy of technical gating mirrors that employed by Microsoft's Copilot.

[51] While roughly half of its training data (48.73%) comes from public datasets, BloombergGPT relies heavily on Bloomberg's proprietary corpus of financial documents. *See* Shijie Wu et al., *Bloomberg-GPT: A*

Like litigation and platform enclosure, technical gating also has a signaling effect. By embedding access limits into the architecture of widely adopted enterprise tools, Microsoft, Harvey, and similar providers normalize the expectation that high-value content will remain within closed technical boundaries. Over time, this can shift industry norms toward data localization and platform-specific exclusivity, reinforcing the broader trend toward fragmentation of the training commons.

D. Data Degradation and Crawl Decline

A final form of gating is less visible than litigation or explicit platform enclosure but may prove equally consequential: the steady contraction in the breadth and depth of material captured by large-scale web crawls. Public datasets such as Common Crawl, which has long been regarded as a proxy for the "open web," are increasingly constrained by both technical barriers to crawling and the strategic decisions of major content hosts.[52] As sites deploy bot detection, rate limiting, paywalls, and other access restrictions, the share of the web that remains open to automated collection continues to shrink.

The decline is measurable. Analyses of Common Crawl snapshots show a steady reduction in coverage of certain high-value domains, accompanied by an increase in duplicate or low-quality material. The ratio of unique, high-quality pages to total pages crawled has fallen over time, reflecting both the withdrawal of valuable sources and the proliferation of SEO-driven or spam content.[53] As one study observes, large-scale crawls now risk overrepresenting "synthetic" pages, i.e.,

---

Large Language Model for Finance, ArXiv, at 9 (2023), https://arxiv.org/pdf/2303.17564 [https://perma.cc/G8XM-5XTD]. By keeping this proprietary material inaccessible to outsiders, Bloomberg employs a form of technical gating.

Google uses internal sandboxing systems—such as the GDC Sandbox, to emulate *air-gapped* environments, where systems are physically or logically isolated from external networks. Rohan Grover & TJ Banasik, *Emulating the Air-Gapped Experience: GDC Sandbox Is Now Generally Available*, Google Cloud (June 3, 2025), https://cloud.google.com/blog/topics/hybrid-cloud/using-gdc-sandbox-to-emulate-air-gapped-environments [https://perma.cc/X2JY-NX7X]. This setup addresses major barriers to adopting generative AI, including strict regulatory mandates, sovereignty requirements, low-latency processing needs, and the challenge of handling large volumes of on-premises data. By limiting access to training pipelines and infrastructure to approved users, sandboxing also operates as a form of technical gating.

[52] *Infra* Parts I.B–C.

[53] *See* Baack, *supra* note 32, at 6 (explaining that "big and important domains . . . now block Common Crawl" and that its sampling process is designed to filter out spam). Even before recent copyright litigation and removals by major publishers like *The New York Times*, Common Crawl systematically excluded

content produced or amplified by automated systems, including AI-generated text optimized for search ranking, while underrepresenting authoritative and curated sources.[54]

For AI developers, this form of gating is particularly difficult to counteract. Unlike platform enclosure, where the loss is tied to a discrete event or policy change, crawl decline unfolds gradually and diffusely.[55] Each individual barrier may be minor, but their cumulative effect can be substantial. Models trained on newer datasets may exhibit subtle but persistent erosion in coverage of specialized domains, niche expertise, and historically open archives.

This degradation has significant feedback effects. As high-quality open content disappears from training pools, models may be forced to rely more heavily on older copies of that material, risking staleness and reduced accuracy in domains where facts evolve rapidly.[56] In parallel, the proliferation of synthetic content online increases the risk of recursive training loops, in which models learn from data originally generated by other models, leading to compounding errors, artifacts, and bias.[57] The result is a gradual shift in the composition of "public" training data, one that can impair model performance without any formal change to copyright law or platform access policies. In this respect, crawl decline illustrates a broader governance problem. Access can collapse even as doctrine remains permissive, because the practical composition of the training corpus is determined by upstream infrastructure rather than legal rules.

---

platforms such as Facebook, despite their centrality to online activity. *Id.* Thus, Common Crawl cannot be understood as a comprehensive "snapshot" of the web.

[54] *Cf.* Jesse Dodge et al., Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus 1292 (2021) (reporting that crawl-based corpora contain substantial amounts of machine-generated text). With the spread of gating, crawls are more likely to contain a higher fraction of synthetic pages.

[55] This is because crawl decline results from many small and decentralized actions, such as publishers blocking bots, changing link structures, or removing pages, which accumulate gradually. Unlike a discrete enclosure event, there is no single trigger or counterparty, making the loss harder to detect and address.

[56] This pattern is reinforced by inclusion practices that favor domains from past successful crawls, increasing the likelihood that previously collected material will be recaptured. *See* Baack, *supra* note 32, at 6 (explaining that Common Crawl employs this practice).

[57] *See* Ilia Shumailov et al., *AI Models Collapse When Trained on Recursively Generated Data*, 631 Nature 755, 755 (2024) (describing the phenomenon).

II
A Tiered Content Framework

Part I documented how legal, technical, and platform-based gating is already reshaping the content environment for large language model (LLM) training. These shifts do not affect all producers uniformly.[58] Some operate well above the margin, with strong economic and institutional incentives to keep their content publicly accessible even absent direct licensing revenue.[59] Others sit much closer to the margin, where the economic returns on continued openness are weaker, and the incentives to withdraw, restrict, or license become correspondingly stronger.[60]

Without a way to distinguish between these positions, proposals for content regulation risk being either overbroad, imposing unnecessary obligations on stable producers, or underinclusive, leaving vulnerable sources unprotected.[61]

This Part develops a three-tier classification framework that enables more precise targeting of policy interventions. The tiers—baseline, at-risk, and transitional content—are defined by both economic resilience and the likelihood of gating under current and foreseeable market conditions. Baseline content comprises inframarginal producers whose economic position and complementary benefits make content restriction unlikely, regardless of whether they receive licensing revenue. At-risk content refers to marginal producers whose works are socially valuable but economically precarious, making withdrawal or gating likely absent compensation. Transitional content occupies the middle ground, where market changes or policy shifts have begun to erode accessibility without yet producing full withdrawal; this tier includes both inframarginal and marginal producers, though the latter are more vulnerable.

The framework is designed to be dynamic: content can move between tiers as market conditions, legal rules, or technical architectures evolve. This adaptability is essential, as the current AI training ecosystem remains fluid, with expansion and contraction occurring at different points in the content supply chain.[62]

---

[58] *Infra* Parts II.A–C.

[59] *Infra* Part II.A.

[60] *Infra* Part II.B.

[61] *Infra* Part III.A.

[62] As noted earlier, this pattern is visible in the push toward gating discussed in Part I and reflected in studies of Common Crawl, where certain high-value websites have been withdrawn by content producers

By distinguishing among these categories, the framework operationalizes the Article's three-part test for contingent licensing developed in Part III. All licensing justifications begin with the same threshold: the content must have demonstrable value for training. The remaining two factors, whether withdrawal is likely and whether voluntary bargaining has failed, determine placement within the tiers. Content that satisfies both conditions is at-risk; content that satisfies neither is baseline; and content for which the conditions are emerging but not yet decisive is transitional. This mapping allows regulatory tools such as contingent licensing to be deployed only where warranted, preserving access to high value but vulnerable inputs without imposing unnecessary obligations on stable producers and thereby maintaining both the openness and diversity of the training pool.[63]

A. Baseline Content

Baseline content refers to works produced by inframarginal creators, producers whose incentives to keep content publicly accessible remain strong even without direct licensing income from AI developers.[64] In economic terms, an inframarginal producer operates well above the point at which the marginal benefit of openness equals the marginal cost. Because they derive substantial complementary benefits from accessibility, they are unlikely to gate content under current market conditions.

Inframarginality can arise for several reasons. Some content, especially user-generated material such as posts on X or Reddit, is motivated by reputation incentives or a desire to contribute information.[65] Many baseline producers

themselves. *See* BAACK, *supra* note 32. Other studies have shown that low-traffic sites have also disappeared. *See* Bocharov, *supra* note 7.

[63] Again, as noted *supra* note 26, the goal advanced by this Article is to maximize social welfare in a broad sense, encompassing the welfare of LLM developers, content producers, and consumers of both LLMs and content. Achieving that goal, as will be shown, requires a recognition that not all content is identical.

[64] *See* Hal R. Varian, What Use Is Economic Theory? 7–8 (Aug. 1989) (unpublished conference paper), https://people.ischool.berkeley.edu/~hal/Papers/theory.pdf [https://perma.cc/Y2DL-BEAX] (noting that inframarginal decisions are ones for which people are insensitive to price changes). By analogy, inframarginal content consists of works whose producers already obtain sufficient non-licensing benefits, e.g., institutional funding, advertising revenue, reputational returns, such that, at prevailing conditions, the marginal cost of supplying the content is below the marginal benefit from those non-licensing sources. Because those external benefits, and not LLM licensing income, determine output, any licensing income does not materially change the equilibrium quantity supplied.

[65] See Roland Bénabou & Jean Tirole, *Intrinsic and Extrinsic Motivation*, 70 REV. ECON. STUD. 489, 489 (2003), which distinguishes between extrinsic incentives, such as reputation and social approval, and intrinsic

are also embedded within institutions, including publicly funded universities, large commercial publishers with diversified revenue streams, and mission-driven nonprofits, whose business models do not depend on per-use monetization.[66] Others rely on indirect strategies, such as advertising, brand visibility, or cross-selling, making openness a driver of value elsewhere in the enterprise. Government agencies releasing data pursuant to transparency mandates,[67] and news organizations maintaining limited open access for credibility and reach,[68] reflect the same underlying logic.

---

incentives, such as an internal desire to help others. Both forms can sufficiently encourage content creation independent of any licensing income.

[66] Consider law professors: compensation is typically salary-based and does not vary substantially with the number or quality of publications produced. Accordingly, the decision to publish a law review article is largely independent of compensation. *Cf.* Pamela Samuelson, *Google Book Search and the Future of Books in Cyberspace*, 94 Minn. L. Rev. 1308, 1332 (2010) (observing that academic authors often write for reputational returns).

Similarly, large commercial publishers such as *The New York Times* have diversified revenue streams that include subscriptions, advertising, and other lines of business; licensing fees from LLM developers are not necessary for their profitability. As a result, prospective licensing revenue is unlikely to affect its publication decisions. Those decisions are largely insensitive to such income.

Mission-driven nonprofits exhibit the same inframarginal decision to publish. Their output is justified by mission goals and financed by donations, grants, and endowment income, often alongside open-access policies, so prospective LLM-licensing revenue is ancillary. Once marginal costs are covered, additional licensing income does not meaningfully alter the quantity or timing of publication. Furthermore, many non-profits can rely on volunteer labor, which keeps operating costs at low. *See, e.g.*, Maryana Iskander, *Love Wikipedia? Get to Know the Nonprofit Behind It*, Wikimedia Foundation (Nov. 9, 2024), https://wikimediafoundation.org/news/2025/11/09/love-wikipedia-get-to-know-the-nonprofit-behind-it/ [https://perma.cc/48Y5-8UZ4] (noting that Wikipedia is maintained by almost 250,000 volunteers and funded through reader donations); Benjamin Good et al., *Microtask Crowdsourcing for Disease Mention Annotation in PubMed Abstracts*, ArXiv (2014), https://arxiv.org/pdf/1408.1928 [https://perma.cc/4DR7-8NKJ] (describing how biomedical databases like PubMed have been partially annotated using crowdsourced labor paid as little as $0.06 per abstract).

[67] Federal agencies, for instance, are required to make data open by default, maintain comprehensive public data inventories, and designate Chief Data Officers to ensure compliance. *See* 44 U.S.C. § 3506(b) (requiring agencies to ensure open access to data assets); *id.* § 3511(a) (mandating development and publication of comprehensive data inventories); *id.* § 3520(a) (establishing Chief Data Officers to oversee data governance). These initiatives can (and are) supported by non-profits that make the data more accessible. See for example, CourtListener.com, which archives legal data and provides excellent search tools. Court Listener, https://www.courtlistener.com/opinion/ [https://perma.cc/R365-TH6F] (last visited Jan. 25, 2026).

[68] *See* Lesley Chiou & Catherine Tucker, *Paywalls and the Demand for News*, 25 Info. Econ. & Pol'y 61, 61 (2013) (observing a reduced readership in response to gating).

Although baseline producers are stable under present conditions, stability should not be mistaken for permanence. Their inframarginal status depends on the persistence of the economic and institutional arrangements that make openness valuable. If those conditions shift, through declining advertising markets, reduced institutional funding, or sustained erosion of audience engagement, baseline producers may migrate toward transitional or even at-risk status. Monitoring is therefore essential, even within this ostensibly secure tier.

Because baseline content is not presently endangered, it does not warrant regulatory intervention. Imposing licensing obligations on this tier would be overbroad, adding unnecessary transaction costs and potentially discouraging openness by converting a nonrival public good[69] into a monetized asset. Policy should instead treat baseline content as the anchor of the AI training ecosystem, reinforcing the forces that sustain openness through measures such as open-data initiatives, public funding, and competition policy that limits overconcentration in distribution channels. Only if systemic gating were to emerge, meaning a widespread shift from openness to restriction across a substantial share of baseline producers, would targeted intervention be justified, particularly where the transaction costs of bilateral negotiation would be prohibitive.[70]

[69] Baseline content does exhibit textbook positive externalities: LLM developers benefit from its existence without compensating the creators, which do not internalize that value. In theory, this should lead to underproduction. *See* Paul A. Samuelson, *The Pure Theory of Public Expenditure*, 36 Rev. Econ. & Stat. 387, 387–89 (1954) (explaining that public goods tend to be underproduced when their benefits cannot be fully captured by individual producers). But for now, the shortfall appears minimal. The institutions and communities that support this content, such as universities, governments, open-source projects, continue to produce it for reasons that are largely independent of potential licensing income from AI developers.

[70] Specifically, "systemic gating" refers to a pattern of access restrictions by a critical mass of baseline content producers, such that the training ecosystem as a whole suffers a measurable and enduring reduction in quality or breadth. Under this approach, systemic gating would need to be shown to: (1) affect a significant share of relevant content within a category; (2) materially impair the ability of developers to train models that meet public-interest benchmarks; and (3) occur in a market environment where transaction costs or bargaining asymmetries make voluntary licensing infeasible.

Systemic gating analogizes, imperfectly, to Michael Heller's anticommons. The anticommons describes underuse that arises when multiple parties each hold rights to exclude, making it hard to assemble a workable bundle of permissions at reasonable cost. *See* Michael A. Heller, *The Tragedy of the Anticommons: Property in the Transition from Marx to Markets*, 111 Harv. L. Rev. 621, 622–23 (1998). The oft-invoked "Soviet grocery store" image (bare shelves despite upstream supply) captures the intuition of serial veto points, but it stems from hierarchical bottlenecks and price controls within a single owner (the state), not from fragmented private exclusion rights (of numerous independent content creators).

B. At-Risk Content

At-risk content consists of work that is economically *marginal*.[71] In economic terms, its producers operate close to the break-even point, such that modest shifts in costs or revenue can determine whether their content remains publicly accessible.[72] As noted in the Introduction, the rise of generative AI heightens these pressures by enabling broad substitution, where users satisfy their information needs through AI outputs rather than through original sources, and by creating new bargaining asymmetries between large AI developers and smaller producers. When these producers experience reduced traffic, diminished advertising revenue, or declining licensing income, the rational response is to gate or withdraw their works unless compensated.[73]

The economic fragility of marginal producers leaves them with weak incentives to preserve open access in the absence of direct payment.[74] Their

By contrast, systemic gating of content more closely resembles a patent thicket: complementary inputs are dispersed across many rights holders and intermediaries (paywalls, robots.txt and API/TOS restrictions, DRM/database rights, platform gatekeeping), so that each actor holds a veto, raising search and bargaining costs, creating royalty-stacking risk, and inviting holdup even when downstream value is clear. *See* Michael A. Heller & Rebecca S. Eisenberg, *Can Patents Deter Innovation? The Anticommons in Biomedical Research*, 280 Science 698, 698 (1998); Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, 1 Innovation Pol'y & Econ. 119, 120 (2000); Mark A. Lemley & Carl Shapiro, *Patent Holdup and Royalty Stacking*, 85 Tex. L. Rev. 1991, 1991 (2007).

The analogy of a patent thicket to systemic gating is not perfect since many "gates" are contractual or technical in nature rather than represented by the exercise of traditional property rights. Moreover, fair use and similar defenses mitigate exclusion, and access policies change over time. Nonetheless, the analogy is still useful. The mechanism (many veto points lead to underuse) and the remedies (coordination tools like clearinghouses, pools, or default/blanket licenses) align.

[71] *See* Varian, *supra* note 64, at 8 ("The marginal decisions are the ones that you agonize over .... If the price were a little higher or a little lower, the results of your agonizing might be different").

[72] At-risk content captures production decisions at the margin, where small changes in expected remuneration can flip whether the work is created. This maps onto the classic access–incentives tradeoff: charging for access reduces use ex-post but can raise ex-ante incentives to create where they are weakest, i.e., at the margin. *See* William M. Landes & Richard A. Posner, The Economic Structure of Intellectual Property Law 20–21 (Harv. Univ. Press 2003) ("Unless there is power to exclude, the incentive to create intellectual property [in this case, content for LLM training] in the first place may be impaired .... [T]he result is the 'access versus incentives' tradeoff: charging a price for a public good reduces access to it (a social cost), making it artificially scarce ... but increases the incentive to create it in the first place, which is a possibly offsetting social benefit.").

[73] This basic tradeoff is broadly acknowledged. *See* Shyamkrishna Balganesh, *Foreseeability and Copyright Incentives*, 122 Harv. L. Rev. 1569, 1578 n.27 (2009) (collecting citations).

[74] *See* Landes & Posner, *supra* note 72, at 20–21.

revenue models, which are often dependent on advertising, subscriptions, or donations, are highly sensitive to even modest declines in audience engagement.[75] Niche journalism outlets, specialized academic repositories, and independent creative archives, for example, may find that AI-driven substitution reduces their readership enough to threaten their viability.[76] Without compensation, continued openness would effectively require these producers to subsidize AI development at their own expense.

From a policy perspective, the at-risk category is where a contingent licensing mechanism is most justified, not because the set of producers is small or easily enumerated, but because the expected social loss from withdrawal is high and the likelihood of withdrawal is elevated by substitution effects and weak outside options.[77] Transaction costs and bargaining frictions affect both baseline and at-risk producers, but in the at-risk segment they can be especially acute because parties are numerous and heterogeneous, valuation and attribution are difficult, and bargaining power is often asymmetric. These frictions, rather than any principled unwillingness to license, can explain why voluntary agreements may fail to

[75] *See*, *e.g.*, Rebecca Tushnet, *Attention Must Be Paid: Commercial Speech, User-Generated Ads, and the Challenge of Regulation*, 58 Buff. L. Rev. 721, 721–23 (2010) (describing the "attention economy" and the difficulty of attracting audience eyeballs for ad-supported media).

[76] *See* Chapekis & Lieb, *supra* note 3 (noting that Google AI summaries reduced click-through traffic to underlying websites by about half).

[77] In addition, vulnerable content once removed from general accessibility is rarely restored. Panagiotis Papadopoulos et al., Keeping out the Masses: Understanding the Popularity and Implications of Internet Paywalls 5 (2020), https://dl.acm.org/doi/epdf/10.1145/3366423.3380217 [https://perma.cc/5W7W-6RF6] ("We observed no instances of sites using paywalls, removing the paywalls, and then re-establishing it."); Mike Ananny & Leila Bighash, *Why Drop a Paywall? Mapping Industry Accounts of Online News Decommodification*, 10 Int'l J. Comm. 3359, 3359 (2016) (documenting temporary drops and later reinstatements for emergencies, promotions, and experiments). Legal and technical controls further lock in closure (e.g., DMCA § 1201–style anticircumvention applied to paywalls). Theresa M. Troupson, *Yes, It's Illegal to Cheat a Paywall: Access Rights and the DMCA's Anticircumvention Provision*, 90 N.Y.U. L. Rev. 325, 325 (2015). To be sure, reversals do occur, and many outlets temporarily dropped paywalls for COVID-19 coverage before restoring them, but these are exceptions. Once closed, content generally remains closed absent a compelling shock. Mark Jacob, *Local News Outlets Drop Paywalls for Pandemic Stories, but Gain Digital Subscribers Anyway*, Nw. U. Loc. News Initiative (Mar. 25, 2020), https://localnewsinitiative.northwestern.edu/posts/2020/03/25/digital-subscriptions-virus/ [https://perma.cc/R2X6-RNCR].

materialize.[78] Accordingly, the at-risk tier marks the point at which the three-part test for contingent licensing is most consistently satisfied.

Importantly, the at-risk designation is dynamic. Producers may enter or exit the category as market conditions evolve. A formerly marginal outlet that secures stable funding or develops a reliable licensing market may transition to baseline status. Conversely, baseline producers facing sharp revenue declines due to changes in search algorithms, platform policies, or audience behavior may become newly at risk. Any policy framework must therefore be capable of updating classifications regularly and with minimal friction.[79]

C. Transitional Content

Transitional content occupies the middle ground between stable baseline sources and genuinely at-risk material. It consists of works or datasets whose economic position, legal posture, or distribution model is beginning to shift in ways that could, over time, affect their availability for AI training.[80] Such changes may stem from new monetization strategies, alterations in ownership, or early moves toward gating, i.e. developments that signal potential vulnerability but do not, at present, make withdrawal likely.

Because transitional content can move in either direction, toward stability or toward risk, it warrants more active monitoring. Timely observation enables policymakers and market actors to identify when a source's trajectory has shifted enough to require reclassification. This tier is distinct from at-risk content, where barriers to licensing or policy intervention are more pronounced by definition.

---

[78] *See* Jacob Alhadeff, Cooper Cuene & Max Del Real, *Limits of Algorithmic Fair Use*, 19 Wash. J. L. Tech. & Arts 1, 49–51 (2024) (observing that payment of licensing fees presents a substantial barrier to small, start-up model developers and that concentration of large developers, and potentially less innovation, would result). Because they operate at the economic margin, they often lack dedicated legal or licensing staff and cannot justify the upfront expense of contract negotiation for a single developer. Developers, for their part, cannot feasibly pursue individual agreements with large numbers of such producers, especially when the value of any one source is uncertain. *Cf.* Jacob Victor, *Reconceptualizing Compulsory Licenses*, 72 Stan. L. Rev. 915, 915–16 (2020) (noting the transaction costs problem within the context of music licensing with many musical artists).

[79] *Infra* Part IV.A.

[80] *Cf.* James Boyle, *The Second Enclosure Movement and the Construction of the Public Domain*, 66 L. & Contemp. Probs. 33, 33–34 (2003) (conceptualizing progressive enclosure of formerly open resources as legal, technological, and social conditions evolve). By analogy, publicly accessible inframarginal content may migrate behind paywalls as the environment changes; the flow can also run in reverse if conditions loosen.

Because transitional content is identified early, before the economic pressure to gate becomes acute, transitional content is generally easier to license or address through targeted agreements, which reduces the need for heavy-handed intervention at the outset.

One might object that all content is transitional, given a long enough horizon. But that view is too broad to be operationally useful. The framework draws lines based on observable changes that matter for near-term training access, not on hypothetical long-run shifts. In this sense, "transitional" identifies a definable subset of content whose future availability is meaningfully in flux. Monitoring is appropriate across all tiers, including baseline content, but the intensity of monitoring should vary with proximity to a tipping point. Transitional content warrants closer attention because small changes in incentives can more readily shift it toward at-risk status where the timing pressure to remunerate and avoid gating is theoretically immanent.

At-risk content remains the more pressing policy challenge because it involves producers facing the highest transaction costs and bargaining frictions, making voluntary agreements difficult to secure. This is why the proposed contingent licensing mechanism is tailored to the at-risk tier. It provides genuinely endangered producers with a means to maintain viability without disrupting the property rights or market positions of more stable sources. Transitional content could migrate into the at-risk category if conditions deteriorate, but that movement is neither automatic nor presumed. The value of the transitional designation is that it flags content for closer observation while preserving space for market solutions before any regulatory tools are triggered.[81]

III
Why Contingent Licensing?

Parts I and II showed that the primary pressure point in the training-data ecosystem lies with at-risk producers, sources whose works remain socially valuable but are increasingly vulnerable to withdrawal as substitution, gating, and transaction costs accumulate. When these producers exit, the resulting contraction cannot be offset through fair use or voluntary bargaining alone. Part III therefore

[81] For examples of baseline and at-risk content, *see infra* Part III.D.

turns to the institutional question, i.e., how to preserve access to high-value, economically fragile content without disrupting markets that already function well.

A narrowly tailored contingent licensing framework offers one approach. It supplies a predictable fallback that preserves access to socially valuable content when bilateral negotiation is impracticable, while minimizing intrusion into markets that function well on their own.[82] The analysis that follows considers how such a fallback operates, the conditions under which it should apply, and why extending licensing beyond those limits would be premature.

A. Contingent Licensing as a Bargaining Floor

A contingent license functions as a bargaining floor. It reduces the transaction costs that prevent mutually beneficial exchanges by standardizing terms and aggregating demand.[83] This mechanism is aimed at cases where producers are willing to license but cannot feasibly negotiate separately with multiple developers, and where developers cannot identify or contract efficiently with a large number of small or heterogeneous sources.[84] By offering a low-friction alternative, the fallback preserves access to content that might otherwise disappear from the training pool.

A central feature of this floor is its opt-out structure. Producers retain full discretion to withdraw their works from the licensing pool and pursue bespoke arrangements, ensuring that the fallback does not operate as a back-door compulsory license. Developers gain a reliable access channel when bilateral negotiation is impractical, while producers maintain control over participation and terms.

---

[82] For a historical example of legal intervention used to address large-scale coordination failures in content licensing in early radio broadcasting, *see* Victor, *supra* note 78, at 915–16. As a compulsory regime, the analogy is imperfect, but it illustrates how fallback mechanisms can stabilize access when transaction costs render bilateral negotiation infeasible.

[83] *See* Jane C. Ginsburg, *Fair Use for Free, or Permitted-But-Paid?*, 29 Berkeley Tech. L. J. 1383, 1386 (2014) (noting that licensing schemes reduce the transaction costs associated with negotiating numerous small deals, while still allowing for voluntary agreements, which offers flexibility).

[84] Victor notes that the original goal of promoting broad distribution of protected works through compulsory licensing has largely been supplanted by a focus on the minimization of transaction costs. Victor, *supra* note 78, at 915–16.

Effective design depends on accurate price calibration.[85] If the fallback price is set too low, content producers will opt out, and the mechanism will fail to prevent the withdrawal of socially valuable content. If it is set too high, AI developers will avoid the regime and attempt individual negotiation or forego access altogether, recreating the very transaction-cost barriers the fallback is meant to address. When calibrated correctly, however, the fallback operates as a form of insurance. It reduces uncertainty for marginal producers whose continued openness depends on securing some predictable return, and it preserves room for higher-value deals for producers with greater bargaining leverage. A well-set price stabilizes supply at the margin without distorting markets that already function effectively.

Yet the same feature carries a countervailing risk. When the licensing floor sits close to a content producer's opportunity cost, the incentive to seek higher-value arrangements may diminish. Producers with limited brand recognition or a narrow market reach may see little benefit in negotiating individually if the floor offers a guaranteed, low-friction payout. Over time, this dynamic can shift the composition of licensing activity. Large, well-positioned producers continue to negotiate bespoke terms, while smaller ones default to the fallback and seldom pursue direct bargaining. This bifurcation may be efficient if the fallback price approximates what marginal producers could obtain through negotiation, but it can also dampen competitive dynamics if the fallback becomes the dominant mode of exchange rather than a safeguard against market failure.

Designing the floor level is therefore a delicate exercise.[86] If set too low, it will fail to prevent the withdrawal of socially valuable content; if set too high, it risks entrenching a one-size-fits-all market that displaces differentiated bargains. The appropriate calibration should preserve the fallback's insurance function for producers near the margin while maintaining meaningful upside for those able to negotiate above it. Striking that balance is not merely a matter of economic modeling. It will require iterative adjustment as market conditions evolve and as

[85] *See* Pasquale & Hun, *supra* note 14, at 236–42 (discussing benchmark pricing of AI training data); Sobel, *supra* note 26, at 92-93 (noting the difficult of ascertaining price). The challenge for any licensing scheme, including the one proposed in this Article, is to identify an efficient price in the absence of clear market signals that arise from bargaining. *See* Armen A. Alchian, *Information Costs, Pricing, and Resource Unemployment*, 7 Econ. Inquiry 109, 109 (1969) (emphasizing the need for information that arises from real-world exchange to accurately discover prices).

[86] Real-world bargains should be approximated where possible. *See* Alchian, *supra* note 85, at 109.

producers and developers respond to the regime in ways that may not be fully predictable ex ante.

It should be clear that accuracy in this dimension is central to the framework's success.[87] The purpose of the floor is not to dictate prices for all transactions, but to provide a credible baseline that lowers search and negotiation costs and thereby makes voluntary agreements more likely.[88] If a reliable price cannot be set, either because the underlying information is thin or because costs vary too widely across producers, then a contingent framework should not be applied; poorly calibrated intervention can be more distortive than no intervention at all.[89]

B. Excluding Baseline and Transitional Content

It may seem benign to extend the contingent licensing scheme to all producers, including baseline and transitional ones, on an optional basis.[90] After all, these producers could accept the statutory floor or decline it in favor of private arrangements. But even a voluntary floor is not costless. Its availability can shift expectations and distort incentives, alter bargaining strategies, and reshape markets

[87] Note that administrative costs, including the costs of acquiring information to set efficient prices, must be considered as well. If these prove to be prohibitive the contingent licensing scheme fails on welfare grounds. *See infra* Part IV.

[88] A related way to preserve participation incentives is to allocate pooled licensing revenues probabilistically rather than proportionally by means of lottery-style distributions when individualized accounting would exhaust the very funds it seeks to allocate. Online creator economies already function this way. Most YouTube or Substack contributors receive little or nothing, yet participation persists because the prospect of an outsized return sustains aggregate supply. A contingent-licensing pool could emulate this logic by distributing a small share of total receipts through weighted draws among verified at-risk contributors, thereby maintaining expected value while avoiding high administrative cost. Such probabilistic rewards would not track precise contributions but would achieve statistical sufficiency by providing enough of an expected payoff to keep marginal producers engaged. The idea echoes classic law-and-economics treatments of uncertain recovery. *See* Saul Levmore, *Probabilistic Recoveries, Restitution, and Recurring Wrongs*, 19 J. L. Studies 691, 691–92 (1990). Applied here, the same logic justifies modest randomness in payout design: when the goal is to sustain socially valuable production rather than to achieve distributive precision, probabilistic mechanisms can preserve incentives at lower cost than universal micropayments.

[89] This point will be further elaborated *infra* Part IV.

[90] As mentioned *supra* note 26, existing scholarship proposes blanket licenses. This approach is largely taken by the Copyright Act of 1976 as well. *See* 17 U.S.C. §§ 111, 114, 115, 118, 119, 122, 511 (providing for compulsory mechanical licenses for the creation of new sound recordings of non-dramatic musical compositions; the distribution of phonorecords to the public; satellite and cable retransmissions; and non-interactive digital performances of sound recordings, i.e., streaming); *see also* Ginsburg, *supra* note 83, at 1432–33 (cataloging the various forms).

in ways that are counterproductive where voluntary exchange already functions well.

Consider incentive distortion. If a guaranteed licensing payment were available to all producers, some who currently keep their content openly accessible might instead gate it to qualify for compensation, reducing the supply of freely available training data.[91] That outcome would run directly counter to the framework's core aim of preserving openness wherever market incentives already support it.

A further concern is administrative. Licensing requires resources to set terms, monitor compliance, and recalibrate prices as conditions change.[92] Extending the scheme to producers who are not at risk of exit would expand the volume of covered works without a corresponding benefit, making the system more complex and resource-intensive to operate. Broader coverage would also increase the frequency and difficulty of price-setting and eligibility determinations, raising the likelihood of error and inefficiency.

Extending coverage could also entrench large baseline producers in ways that disadvantage smaller or newer creators.[93] Even with an opt-out, a guaranteed floor may function as a de facto revenue stream for incumbents, reinforcing their market position and narrowing diversity in the content ecosystem. At the same time, broad inclusion would dilute the resources and administrative attention available to

[91] *See* Cass R. Sunstein, *Deciding by Default*, 162 U. Pa. L. Rev. 1, 17–24 (2013) (observing that defaults alter behavior via inertia, endorsement effects, and loss aversion, thereby shifting market expectations); Mark A. Lemley, *Intellectual Property Rights and Standard-Setting Organizations*, 90 Calif. L. Rev. 1889, 1960 (2002) (noting that default rules promulgated by standard setting organizations can create false expectations).

Monetary incentives, such as licensing income, can crowd out norm-based compliance and reframe conduct as a priced option. *See* Uri Gneezy & Aldo Rustichini, *A Fine Is a Price*, 29 J. Legal Stud. 1, 1 (2000) (observing that a small fine for late pickups of children from daycare increased lateness of parents consistent with substitution of fines for non-monetary penalties such as shame).

[92] In the context of administering collective management systems, "owners of rights authorize collective management organizations to monitor the use of their works, negotiate with prospective users, give them licenses against appropriate remuneration on the basis of a tariff system and under appropriate conditions, collect such remuneration, and distribute it among the owners of rights." Mihaly Ficsor, Collective Management of Copyright and Related Rights 17 (2d ed. 2002).

[93] *See* Herbert Hovenkamp, *Antitrust and the Patent System: A Reexamination*, 76 Ohio St. L.J. 467, 469–70 (2015) (discussing the ways that patent and related intellectual property protections can reinforce incumbency and impede competition).

support genuinely at-risk creators, whose contributions are most likely to disappear without targeted intervention.

Absent systemic gating,[94] extending the scheme more broadly could also weaken its legitimacy. A mechanism narrowly focused on preventing the loss of socially valuable content is easier to justify, both politically and legally, than one that extends to well-resourced producers whose incentives already support continued openness.[95] The narrower and more targeted the intervention, the stronger the case that it is necessary and proportionate.

C. Why Not Compulsory Licensing?

A universal compulsory licensing scheme could, in theory, offer several advantages. It would ensure continuity of access to a broad spectrum of training inputs, reducing the risk that valuable datasets disappear due to shifting incentives or individual producer decisions.[96] By guaranteeing that baseline, transitional, and at-risk content remains available on uniform and predictable terms, such a scheme could blunt concentration effects that arise when only a few developers secure exclusive rights to key resources. Broader access might also preserve diversity in model development, easing the tendency toward a small number of dominant architectures trained on idiosyncratically gated corpora. Finally, compulsory licensing could simplify contracting and reduce transaction frictions, particularly where numerous small producers are otherwise difficult to reach,

[94] For a definition, *see supra* note 70. While it is important to anticipate the possibility of systemic gating, the evidence to date does not suggest it has reached a level that justifies immediate, across-the-board licensing. The current landscape is marked more by isolated acts of gating, often by high-profile platforms or publishers, than by the kind of widespread, coordinated withdrawal that would threaten the overall quality and diversity of training data. Treating systemic gating as a live but premature risk keeps the policy response proportionate, preserving flexibility to escalate only if access losses become persistent, pervasive, and demonstrably harmful.

[95] The legitimacy of exceptions in international copyright law has long been tied to proportionality. Article 9(2) of the Berne Convention requires that limitations on exclusive rights be confined to "certain special cases," avoid conflict with the "normal exploitation" of a work, and not "unreasonably prejudice" the rights-holder. By insisting that exceptions be narrowly tailored and justified, the three-step test reflects the intuition that interventions are more defensible, politically and legally, when directed at specific risks rather than applied broadly. *See* Berne Convention for the Protection of Literary and Artistic Works art. 9(2), Sept. 9, 1886, as revised at Paris, July 24, 1971, S. Treaty Doc. No. 99-27.

[96] *See* Jacob Victor, *Reconceptualizing Compulsory Copyright Licenses*, 72 Stan. L. Rev. 915, 915–16 (2020).

allowing developers to plan longer training cycles with greater confidence in data availability.[97]

However, across-the-board compulsory licensing carries significant drawbacks. Foremost is the risk of eroding property rights on a broad scale. By requiring all producers to make their works available for AI training on fixed terms, such a regime would deprive creators of meaningful control over how their works are used.[98] That loss is not merely symbolic. Control over licensing terms can shape how producers position themselves in the marketplace, cultivate audience relationships, and capture downstream value.

Moreover, compulsory licensing carries a nontrivial risk of chilling innovation in both content creation and AI development. If license terms are miscalibrated, that is, if they are too favorable to developers or too burdensome for creators, then either side may face reduced incentives to invest.[99] In particular, a regime in which creators cannot withhold access, even strategically, may dampen the incentive to produce distinctive, high-value works, narrowing the diversity of inputs available for future models.[100]

From an administrative perspective, a universal compulsory licensing regime could, in some respects, be simpler to administer than a contingent, at-risk-only scheme. Because it would apply across all content categories, there would be no need to identify, monitor, and periodically reassess which producers qualify as at-risk or whether triggering conditions have been met. Eliminating those eligibility determinations could reduce complexity and lower the potential for disputes, allowing the licensing authority to focus on setting and enforcing uniform terms. At the same time, administering royalty collection and distribution across a much larger pool of participants could generate its own burdens, and the net effect on administrative costs would depend heavily on implementation.[101] Even if the balance ultimately favored universal licensing on administrative grounds, the efficiency concerns outlined above would remain decisive.

---

[97] *Id.*

[98] *See* Berne Convention, *supra* note 95.

[99] *See* Victor, *supra* note 96, at 944, 988.

[100] In comparison, a well-calibrated licensing regime that provides an opt-out, discussed in Part III.C, *supra*, captures broader benefits insofar as information and administration costs are low.

[101] *See infra* Part IV.

D. Contemporary Illustrations: Gearspace *and* The New York Times

To this point, the analysis has shown that the case for intervention turns on differentiating content that is genuinely vulnerable to withdrawal from content produced by stable actors with independent incentives to remain accessible. That distinction becomes clearer when applied to concrete contexts. The examples that follow, one near the at-risk margin and one firmly within the baseline tier, demonstrate why a uniform licensing regime would be overbroad and that a targeted intervention must focus on sources for which withdrawal is a predictable market outcome rather than a strategic choice.

Gearspace is a long-running forum dedicated to professional audio engineering and music production, and exemplifies content that sits near the at-risk end of the spectrum.[102] Its contributions are narrow in scope, highly technical, and the product of decades of cumulative community knowledge.[103] Although the forum's revenue model is stable enough for now, it lacks the diversified income streams of a major media organization. If site operators were to restrict access to web crawlers due to concerns over uncompensated AI use, the impact on AI training would be disproportionate. Models tuned for audio production assistance could lose a key corpus of domain-specific expertise, degrading niche capabilities in ways that are difficult to restore once the data is gone. Under a contingent licensing regime, *Gearspace* could benefit from standardized licensing terms that preserve access while providing modest compensation, ensuring that its value as a training input is not lost through market exit or prohibitive negotiation costs.

By contrast, *The New York Times* lies at the opposite end of the risk spectrum. It is a baseline producer of content that is widely read, highly trusted, and supported by stable, diversified revenue streams. Its recent decision to gate content from AI crawlers, enforced through technical measures and litigation,[104] does not reflect a risk of market failure, but rather a strategic effort to enhance bargaining leverage. Unlike *Gearspace*, the *Times* has strong, independent incentives to continue producing and distributing content, reinforced by its subscription model, institutional reputation, and established market position. Extending a compulsory

[102] *See* Gearspace, https://gearspace.com/ [https://perma.cc/4LUM-TBT8] (last visited Feb. 16, 2026).

[103] *See* Gearspace, https://gearspace.com/board/ [https://perma.cc/K7PK-8UBU] (last visited Feb. 16, 2026). A simple search of the forum displays rich and detailed content that goes back at least a decade.

[104] *See supra* note 13.

or contingent license to such a producer would not preserve endangered supply; it would instead interfere with a functioning market in which the producer is capable of negotiating access on its own terms.[105]

The contrast between these two cases underscores why a uniform approach to AI training access would be both overbroad and inefficient. *Gearspace's* vulnerability reflects structural fragility. It possesses limited resources, relies on volunteer contributions, and cannot pursue alternative monetization paths readily. The *Times*, by contrast, has the resources and institutional stability to negotiate licensing terms without risking the viability of its output. A contingent licensing framework is therefore suited to preserving access in the former case while respecting property rights and bargaining autonomy in the latter.

Taken together, these examples show that effective policy must track underlying economic conditions. Intervention is warranted only where withdrawal is a predictable market outcome, not where producers already possess the capacity and incentives to negotiate access on their own terms. A contingent licensing framework reflects that distinction, directing support to vulnerable sources without disrupting stable markets or altering incentives where they already function.

IV
Legal Design and Implementation

A licensing regime, no matter how well-justified in principle, should be implemented only if it can operate efficiently. Its benefits must outweigh both the informational demands of implementation and the administrative burdens of ongoing operation. This constraint is especially important for a contingent or fallback framework, which depends on accurate classification of content types, timely monitoring of availability, efficient price-setting, and enforcement mechanisms that preserve rather than distort existing market incentives. As Part III explained, the case for intervention arises primarily when transaction costs make voluntary bargaining infeasible. The design choices that follow proceed from that assumption and evaluate how a contingent licensing system could function in practice under those conditions.

[105] The *Times* recently closed its first licensing deal with Amazon. *See* Jaspreet Singh, *New York Times Partners with Amazon for Its First Licensing Deal*, Reuters, May 29, 2025, https://www.reuters.com/business/retail-consumer/new-york-times-amazon-sign-ai-licensing-deal-2025-05-29/ [https://perma.cc/67F2-2T27].

A. Information and Classification Requirements

A contingent licensing framework relies on accurate, up-to-date information about the content it is meant to support. At a minimum, the system must be able to distinguish at-risk sources from those that can negotiate effectively on their own, and it must do so in a way that is transparent, predictable, and resistant to strategic manipulation. This requires clear criteria for classification and mechanisms for updating those classifications as market conditions change.

Classification cannot rest solely on producer declarations or static indicators. Some sources may appear stable but face declining viability over time; others may present themselves as fragile to obtain favorable terms. A workable system therefore must rely on observable signals such as patterns of withdrawal, shifts in access policies, or documented failures in voluntary negotiations, rather than subjective assessments of vulnerability. The goal is to identify circumstances in which access to a socially valuable input is at genuine risk of disappearing.

Price-setting presents a related informational challenge, but one that is manageable in practice. The objective is not to determine the intrinsic value of each individual work, but to approximate the level of compensation necessary to keep marginal producers from withdrawing access. For many at-risk sources, this information is accessible: operating costs for community forums, technical archives, and niche publishers, such as hosting, storage, moderation, and basic staffing, can be estimated with reasonable accuracy using public data, industry benchmarks, and comparable-market analyses of the kind antitrust and regulatory agencies routinely perform. Where continued availability depends on covering these relatively stable minimum-viability costs, a fallback price can be set to ensure that participation remains economically feasible without displacing higher-value negotiations for producers with stronger market positions.[106] The calibration exercise therefore does not require precision in any microeconomic sense; it

[106] Demand-side indicators can also inform calibration. Antitrust and regulatory agencies routinely approximate the value of inputs by examining downstream firms' revenues, scale, or competitive incentives, and similar methods could be used here to estimate developers' willingness to pay for continued access without requiring precise attribution of a dataset's marginal contribution to model performance.

As already noted, lottery-style payouts to content creators could be deployed to encourage higher levels of participation. *See supra* note 88.

requires a credible floor capable of preventing supply collapse among structurally fragile sources while preserving room for market-driven bargaining above it.

Accurate price-setting is only part of the institutional design. A viable framework must also incorporate monitoring and enforcement mechanisms to ensure that the terms of the license are observed in practice. A contingent license must ensure that participating producers actually make their content available under the specified terms and that developers comply with the conditions attached to use. These obligations need not be onerous, but they must be clear enough to prevent opportunistic behavior and light enough to avoid deterring participation. Enforcement mechanisms should focus on detectable breaches, including improper withholding, misuse of licensed material, or failures to report use, rather than on intrusive oversight of day-to-day activity.

A licensing system that cannot efficiently meet these informational and administrative requirements should not be implemented. Its value depends on the ability to function reliably, adapting to changes in the content ecosystem without imposing excessive burdens on producers, developers, or the institutions responsible for administering it.

B. Institutional Options

Multiple institutional pathways could support a contingent licensing system, each with distinct advantages in scope, flexibility, and enforceability. These models are not mutually exclusive, and any implementation should remain modular so that only the elements needed to address demonstrated market failures are activated.

A statutory pathway offers the clearest legal foundation. Congress could enact a narrowly tailored contingent license, drawing structural guidance from existing regimes such as the mechanical license for musical compositions, but limiting its scope to content that meets strict eligibility criteria and clearly defined rate-setting rules. Such a framework would provide legal certainty and lower transaction costs, while still requiring careful calibration to ensure that it does not extend beyond the market failures it is designed to address.[107]

[107] *See* Victor, *supra* note 96, at 927–30 (explaining that compulsory licensing for musical compositions was justified by the impracticability of individualized negotiation and the public value of broad access to protected works). The Copyright Act's rate-setting criteria similarly emphasize balancing access, fair return, and minimal disruption to existing industry structures. *See* 17 U.S.C. § 801(b)(1)(A)–(D). By

Alternatively, a limited administrative body or collective rights organization could manage the licensing process. Institutions such as ASCAP, BMI, or the Copyright Clearance Center provide models for aggregating rights and distributing royalties, though at a much larger scale and for different types of works.[108] A more modest analogue could identify eligible at-risk content, collect fees, and distribute compensation using standardized criteria rather than granular usage tracking. This approach can scale and adapt to changing conditions but would require ongoing coordination across diverse content categories and careful constraints to ensure that it remains narrow, modular, and limited to contexts where voluntary licensing has failed.

Judicial mechanisms could also play a limited transitional role. Courts have at times responded to recurring copyright disputes by articulating presumptive rules or structured remedies that reduce uncertainty across similar cases.[109] In a training-data context, courts could recognize a narrow presumptive license for gated content that meets defined criteria, such as limited substitution risk, demonstrated public value, and unsuccessful attempts at voluntary negotiation. Although courts are not well suited to designing or administering a licensing system, they can help stabilize expectations and reduce litigation pressure while legislative or administrative solutions develop.

Each institutional pathway carries trade-offs: statutory approaches risk rigidity, administrative models require sustained resources, and judicial mechanisms can yield uneven results. Hybrid options, such as limited legislative authorization for a registry paired with judicial presumptions in recurring

---

contrast, compulsory licensing for noncommercial broadcasters rested primarily on reducing transaction costs rather than on dissemination. Victor, *supra* note 96, at 945. These rationales illustrate the economic logic behind targeted intervention when voluntary licensing becomes infeasible, though the contingent framework proposed here is narrower in scope and focused solely on at-risk content.

[108] *See, e.g.*, Kristelia A. García, *Facilitating Competition by Remedial Regulation*, 31 Berkeley L. & Tech. J. 183, 194 (2016) (describing the ASCAP blanket license).

[109] For instance, courts engage in rate setting when they attempt to approximate the fair market value of a license for performance rights organizations. *See* In re Pandora Media, Inc., 6 F. Supp. 3d 317, 353 (S.D.N.Y. 2014) (quoting Am. Soc'y of Composers, Authors, & Publishers v. MobiTV, Inc., 681 F.3d 76, 82 (2d Cir. 2012), *aff'd*, 785 F.3d 73 (2d Cir. 2015)). In other instances, courts can compel copyright owners to license but leave rate-setting to the parties subject to judicial verification that licenses "are easily accessible [and] reasonably priced." Cambridge University Press v. Becker, 863 F. Supp. 2d 1190, 1237 (N.D. Ga. 2012), *rev'd sub nom*, Cambridge University Press v. Patton, 769 F.3d 1232 (11th Cir. 2014). *See also* Ginsburg, *supra* note 83, at 1397–98 (describing judicial compulsory licensing).

disputes, are also plausible. Analogous coordination tools exist in other domains, including patent pools[110] and fair, reasonable, and non-discriminatory (FRAND) licensing,[111] which address fragmentation by facilitating access to essential inputs while preserving private incentives to innovate. Although the legal structures differ, the underlying challenge of enabling scalable access without collapsing bargaining autonomy is similar. Whatever form is adopted, the central aim should be to maintain contingency, minimize market distortion, and provide clear legal footing for the narrow circumstances in which intervention is warranted.

C. Global Considerations

Although grounded in U.S. law, any contingent licensing must remain compatible with international obligations.[112] Under the Berne Convention and TRIPS, limitations and exceptions must be confined to special cases, avoid conflict with normal exploitation of works, and refrain from causing unreasonable prejudice to the legitimate interests of rights-holders.[113] A narrowly tailored,

[110] *See* Josh Lerner & Jean Tirole, *Efficient Patent Pools*, 94 Am. Econ. Rev. 691, 700–02 (2004) (explaining that patent pools bundle complementary patents, reduce transaction costs, and avoiding blocking rights by resolving overlapping claims).

[111] *See* Carl Shapiro, *Navigating the Patent Thicket: Cross Licenses, Patent Pools, and Standard Setting*, 1 Innovation Pol'y & Econ. 119, 122 (2001) (observing that standard-setting organizations commonly require patent-holders to license essential technologies on FRAND terms); *see also Intellectual Property Rights*, Eur. Telecomms. Standards Inst. (ETSI), https://www.etsi.org/intellectual-property-rights [https://perma.cc/YC7A-L6VW] (last visited Feb. 16, 2026). ETSI requires Standard Essential Patent holders to commit to licensing on FRAND terms to ensure interoperability in global technology standards.

[112] *See, e.g.*, Mattias Rättzén, *Location Is All You Need: Copyright Extraterritoriality and Where to Train Your AI*, 26 Colum. Sci. & Tech. L. Rev. 175, 175 (2024) (noting divergence of copyright rules across the United States, the EU, the U.K., Japan, Singapore, Australia, India, Israel, "and many more countries" and that "there is therefore a high likelihood that the same type of training activity would be considered copyright infringement in some countries but not in others."); Matthew Sag & Peter K. Yu, *The Globalization of Copyright Exceptions for AI Training*, 74 Emory L. J. 1163, 1168, 1212 (2025) (observing a lack of harmonization but signs of an emerging global equilibrium that permits, in some instances, unauthorized use of copyrighted works for training).

[113] *See* Berne Convention, *supra* note 95 (limiting exceptions to cases that are limited in scope, do not conflict with normal exploitation of the work and, do not unreasonably prejudice the rights-holder); Agreement on Trade-Related Aspects of Intellectual Property Rights art. 13, Apr. 15, 1994, 1869 U.N.T.S. 299 (incorporating a similar "three-step test" for limitations and exceptions to exclusive rights).

market-failure-triggered mechanism limited to at-risk content can plausibly satisfy these standards.[114]

Global approaches vary. The EU allows text and data mining for research while allowing commercial rightsholders to opt-out[115] whereas China often relies on state-backed mechanisms to facilitate access.[116] The United States need not harmonize with either model, but should avoid creating obstacles to cross-border training or inconsistencies that complicate international deployment. A transparent, nondiscriminatory, U.S.-focused system would minimize friction and reinforce the legitimacy of any contingent licensing system.

Over time, reciprocal or federated licensing arrangements could emerge across jurisdictions, whether through mutual recognition of rights or coordinated registries for at-risk content. Although such developments are speculative, they illustrate that a narrowly tailored contingent framework is compatible with international law and adaptable to different legal environments.

Conclusion

The shift toward AI-mediated access has reconfigured the economics of online content in ways that copyright doctrine alone cannot address. To date, fair use continues to permit large-scale training, yet its permissive scope does little to sustain the underlying supply of high-value inputs when the incentives that once supported openness begin to erode. As Parts I and II showed, the resulting pressures fall unevenly across the content ecosystem. Baseline producers remain viable under a wide range of conditions, while at-risk creators face growing incentives to withdraw. Transitional sources fall between these poles, signaling where market stress may soon become acute. Any governance response must therefore distinguish

[114] *Cf.* Kaigeng Li, Hong Wu & Yupeng Dong, *Copyright Protection During the Training Stage of Generative AI*, Computer L. & Security Rev., Nov. 2024, at 1, 18 (suggesting that remuneration rights are consistent with international intellectual property treaties).

[115] *See* Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, 2019 O.J. (L 130) 92. Article 3 permits text and data mining for research; Article 4 provides for owners of data to opt-out.

[116] *See China Accelerating Development of Large Language Models*, Xinhua (Jun. 14, 2023), https://english.news.cn/20230614/52ef051d78414df9a5f2f1d94b605c55/c.html [https://perma.cc/Y4KW-36G3] (reporting on large-scale state funding for LLM infrastructure and models).

among these very different positions rather than treating "training data" as a homogeneous resource.

Contingent licensing offers one tool for doing so. Properly designed, it can operate as a narrow, fallback mechanism that activates only when three conditions converge: the content is demonstrably valuable for training, withdrawal is the rational outcome absent remuneration, and bilateral negotiation is blocked by transaction costs or collective-action frictions. A system meeting these criteria does not displace fair use, impose compulsory terms on stable producers, or convert open access into a default licensing market. It preserves what already works, intervenes only where failure is demonstrable, and respects the basic structure of incentives that support continued content production.

As Part IV emphasized, any licensing mechanism must be able to classify content reliably, set prices at a level that stabilizes supply without distorting markets, and operate with administrative efficiency. Where those informational and operational demands cannot be efficiently met, the case for intervention collapses; the welfare-maximizing outcome is non-intervention. But where the conditions are satisfied, a contingent license can prevent the disappearance of socially valuable inputs that models cannot easily substitute for or reconstruct once withdrawn.

Training data governance thus reframes the debate. Rather than focusing on the limits of fair use or the propriety of large-scale scraping, it directs attention to the economic conditions that determine whether critical content remains available at all. The governance challenge is not to guarantee universal access or to suppress strategic gating; it is to ensure that the loss of fragile but essential sources does not degrade the quality, diversity, and reliability of generative AI systems in ways that undermine long-term social welfare. A contingent, evidence-based licensing fallback that is narrow in scope, modest in ambition, and grounded in market failure, could offer one path toward that goal.