Market-Based Licensing for Publishers' Works Is Feasible. Big Tech Agrees.

Matthew Stratton*

ABSTRACT

Generative AI ("GAI") model developers prioritized speed to market over compliance with copyright law with respect to use of copyrighted works for training their models. Now facing over forty lawsuits, they have asserted fair use to evade responsibility, and they claim that licensing all necessary works is impossible.

This Article focuses on professionally created works only, with an emphasis on publishers' works, and demonstrates that market-based licensing of professionally created works for training GAI models is feasible as measured by the number of licenses and the ability of GAI developers to afford them—both of which are points on which Big Tech agrees. The Article also provides insights on the licensing marketplace for publishers' works as relevant to training GAI models. Finally, the Article underscores that the public interest is squarely on the side of market-based licensing because all stakeholders benefit, and it will help ensure that publishers and authors may continue their vital contributions to America's political, intellectual, and cultural systems.

INTRODUCTION

"The unlicensed use of creative works for training generative AI is a major, unjust threat to the livelihoods of the people behind those works, and must not be permitted."[1]

---

* Matthew Stratton was formerly Deputy General Counsel for the Association of American Publishers ("AAP"), a not-for-profit organization that represents the leading book, journal, and education publishers in the United States on matters of law and policy. This Article is written in his personal capacity, and any views expressed are his own and not necessarily those of AAP.

1. *Statement on AI Training*, AI TRAINING STATEMENT, https://www.aitrainingstatement.org/ [https://perma.cc/PNE6-68BJ] [https://web.archive.org/web/20250125211211/https://www.aitrainingstatement.org/] (last visited Feb. 8, 2025).

On October 22, 2024, creators worldwide signed the above public statement condemning the theft of creative and intellectual works of authorship by Big Tech companies for training their consumer-facing generative AI ("GAI") models.[2] By October 31, 2024, there were over 30,000 signatories including James Patterson, Julianne Moore, Thom Yorke, Sir Kazuo Ishiguro, Roger Daltry, Rosario Dawson, and many other celebrated artists.[3] The campaign was organized by Ed Newton-Rex, CEO of Fairly Trained,[4] who publicly resigned from his position as Vice President of Audio at Stability AI in November 2023 in protest to the company's assertion that training its model on unlicensed, copyright-protected works is fair use.[5]

Professional creators make a particularly compelling argument that market-based licensing for training of GAI models is required, which courts should confirm.[6] In the case of large-language models ("LLMs"), the professional creators of textual works are represented principally in just two sectors: (1) publishing; and (2) news/media organizations. While it is legitimate to question how best to license the voluminous amount of *non-professional* copyrighted textual works available on the *entire internet*, it is illegitimate to confound that question with the case of licensing works from professional creators. This Article reinforces that market-based licensing of professional works for training LLMs is feasible by demonstrating that the number of licenses is manageable and that GAI developers can afford them—both of which are points on which Big Tech agrees—and the Article further underscores that the public interest is squarely on the side of market-based licensing.

While this Article focuses on the publishing sector (since it was my focus professionally at the Association of American Publishers), it is supplemented with data on licensing for news/media organizations to emphasize the feasibility for licensing professionally created copyrighted content *overall* for LLMs.

---

2. *Id.*; *See* Jasmine Scott, *Authors and Other Creatives Sign Statement Rejecting Unlicensed Training of Generative AI Models on Creative Work*, SOC'Y AUTHORS (Oct. 22, 2024), https://societyofauthors.org/2024/10/22/authors-and-other-creatives-sign-statement-rejecting-unlicensed-training-of-generative-ai-models-on-creative-work/ [https://perma.cc/6W69-2M46].

3. *Id.*; Ed Newton-Rex (@ednewtonrex), X (Oct. 29, 2024, 1:49 AM), https://x.com/ednewtonrex/status/1850942855147372749 [https://perma.cc/4L78-4HVL].

4. Fairly Trained is a nonprofit that certifies generative AI models that do not use any copyrighted work for training without a license. *About*, FAIRLY TRAINED, https://www.fairlytrained.org/about [https://perma.cc/DC2N-WFCP]
[https://web.archive.org/web/20250125211813/https://www.fairlytrained.org/about] (last visited Feb. 7, 2025). AAP President & CEO Maria Pallante is an Adviser to Fairly Trained. *Id.*

5. Ed Newton-Rex, *Why I Just Resigned from My Job in Generative AI*, MUSIC BUS. WORLDWIDE (Nov. 15, 2023), https://www.musicbusinessworldwide.com/why-just-resigned-from-my-job-generative-ai/ [https://perma.cc/FMW8-2D5D]
[https://web.archive.org/web/20250125212019/https://www.musicbusinessworldwide.com/why-just-resigned-from-my-job-generative-ai/].

6. At the time of writing in December 2024, no court decisions have addressed whether fair use applies to the training of LLMs on copyrighted works, and the U.S. Copyright Office has not yet released the part of its report on copyright and artificial intelligence addressing this issue. *See generally* Artificial Intelligence and Copyright, 88 Fed. Reg. 59,942 (Aug. 30, 2023).

I.   THE PUBLISHING INDUSTRY TODAY

Today's U.S. publishing industry is diverse, ranging from major commercial book and journal publishers to small, non-profit, university, and scholarly presses, as well as leading publishers of educational materials and digital learning platforms. Publishers' works make vital contributions to the nation's political, intellectual, and cultural systems. They:

> Present novel ideas and new facts unearthed by authors,
>
> Hold governments, businesses, and citizens accountable,
>
> Contribute to a vibrant culture,
>
> Educate and inspire citizens of all ages,
>
> Progress medicine and science, and
>
> Serve as a source of truth at a time when there is rampant misinformation spread through the internet, social networks, and other media—a harm which GAI only risks accelerating.[7]

The vibrancy of the U.S. publishing industry, however, should not be taken for granted. History has amply demonstrated the deleterious effects that an unrestrained digital environment may have on the breadth and quality of new works in other creative sectors. If courts were to find erroneously that fair use applied to the training of GAI models with publishers' works, there risks an irreversible "U-turn" in the breadth and quality of published works to the detriment of the American public and democracy.

For reasons already well-explained by many copyright experts, courts should not find fair use with respect to training GAI models on publishers' works.[8] There is nothing exceptional about the development and commercial exploitation of GAI models by Big Tech to change the basic premise that publishers must be remunerated for the use of their copyrighted works through licensing based on exclusive rights under the Copyright Act. More specifically, there is no question that books and articles are the foundation and "most prized data" inside LLMs;[9] that GAI developers sought

---

7.   Ass'n of Am. Publishers, Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 27 (Oct. 30, 2023) [hereinafter, AAP Comments], https://www.regulations.gov/comment/COLC-2023-0006-9070   [https://perma.cc/ZZ9T-BL89] [https://web.archive.org/web/20250125212011/https://www.regulations.gov/comment/COLC-2023-0006-9070].

8.   In view of the ample and persuasive existing scholarship concluding that the fair use defense should not succeed, this Article will not address fair use in depth. For further reading, see Jacqueline C. Charlesworth, *Generative AI's Illusory Case for Fair Use*, 27 VAND. J. ENT. & TECH. L. (forthcoming 2025).

9.   Cade Metz et al., *How Tech Giants Cut Corners To Harvest Data for A.I.*, N.Y. TIMES (Apr. 8, 2024), https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html [https://web.archive.org/web/20250208030143/https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html] ("The most prized data, A.I. researchers said, is high-quality

the expressive content of these works which they have stored in the models through the encoding of tokens into vectors; and that training violates publishers' exclusive rights. Furthermore, as explained in this Article, there is no "impossibility" of licensing or other countervailing public policy reason to waive the basic operation of copyright law, which requires downstream users to license what they do not own but seek for their derivative uses. "Business as usual" should prevail.

II.   LICENSING PROFESSIONALLY-CREATED COPYRIGHTED WORKS IS FEASIBLE AS EVIDENCED BY THE NUMBER OF LICENSES REQUIRED FOR THE WORKS OF THE PUBLISHING AND NEWS/MEDIA SECTORS AND THE ABILITY OF THE GAI DEVELOPERS TO AFFORD THEM

A.   INITIALLY, LICENSING WAS NOT A PRIORITY FOR BIG TECH IN THE GAI ARMS RACE

Although the licensing market for works to train GAI models is rapidly growing, licensing is not yet ubiquitous. The reason is that licensing has not been a priority for Big Tech. Let us briefly review how we got here.

The Big Tech companies made a calculated decision to not license content.[10] In a "rapidly developing and incredibly competitive marketplace,"[11] they wanted to get to market first and accepted that after the product was out, there would be the prospect of lawsuits and licenses.[12] For an internet-based service with a worldwide market, being first or having an early competitive advantage may separate the billion (or trillion) dollar winners from the also-rans. Winning is all that matters. A trail of lawsuits and the potential for reform is mere ancillary clean-up work. "Move fast and break things" is nothing new. The claim by a Silicon Valley venture capital fund that is one of the most aggressive investors in AI that developers acted on the "settled expectations" of fair use is unconvincing, self-serving *ex post* rhetoric.[13]

---

information, such as published books and articles, which have been carefully written and edited by professionals.").

10.   *E.g.*, AAP Comments, *supra* note 7, at 9 ("AAP members were not approached for authorization or license to use their works in the training of Gen AI systems that publicly launched in 2022.").

11.   *See* Declaration of Jared Kaplan in Support of Defendants' Opposition to Plaintiffs' Renewed Motion for Preliminary Injunction, Concord Music Grp., Inc. v. Anthropic PBC, No. 24 Civ. 3811 (N.D. Cal. Dec. 23, 2024) (claiming that a delay caused by a preliminary injunction would "inflict substantial harm on Anthropic's business by . . . hurting its ability to innovate and iterate in a rapidly developing and incredibly competitive marketplace").

12.   *See, e.g.*, Metz et al., *supra* note 9 ("At Meta, which owns Facebook and Instagram, managers, lawyers and engineers last year discussed buying the publishing house Simon & Schuster to procure long works, according to recordings of internal meetings obtained by The Times. They also conferred on gathering copyrighted data from across the internet, even if that meant facing lawsuits. Negotiating licenses with publishers, artists, musicians and the news industry would take too long, they said.").

13.   Andreessen Horowitz (a16z), Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 6 (Oct. 30, 2023) [hereinafter, Andreessen Horowitz Comments], https://www.regulations.gov/comment/COLC-2023-0006-9057   [https://perma.cc/SB7S-EW99]

As of writing, only two years have passed since OpenAI's ChatGPT was released in November 2022.[14] Quickly it became a public phenomenon, reaching 100 million users in approximately two months.[15] Since then, creators and copyright owners have become aware of the infringement, resulting in over forty lawsuits and also approximately fifty publicly reported LLM licensing deals[16]—plus the nearly 4,000 licenses entered into by Dow Jones alone[17]—and an unknown number of deals that have not been publicly reported. There will be many more licenses if courts properly reject Big Tech's fair use defense. Until such time as the courts so rule, Big Tech may continue to license selectively and at times non-publicly to enhance their chances to prevail on the fair use defense.

More specifically, the fourth statutory factor of the fair use analysis is "the effect of the use upon the potential market for or value of the copyrighted work."[18] GAI developers are aware that evidence of an existing and growing licensing market for the use of copyrighted works to train GAI models will undermine their fair use defense.[19]

[https://web.archive.org/web/20250125212358/https://www.regulations.gov/comment/COLC-2023-0006-9057].

14.   Bernard Marr, *A Short History of ChatGPT: How We Got To Where We Are Today*, FORBES (May 19, 2023), https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/   [https://perma.cc/KSC4-3R4M] [https://web.archive.org/web/20250125212243/https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/].

15.   Dan Milmo, *ChatGPT Reaches 100 Million Users Two Months After Launch*, GUARDIAN (Feb. 2, 2023), https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app   [https://perma.cc/M9GF-4QDG] [https://web.archive.org/web/20250125212455/https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app].

16.   *See* PUBLISHERS ASS'N, RESPONSE TO CONSULTATION ON COPYRIGHT AND AI 24 (2025), https://www.publishers.org.uk/wp-content/uploads/2025/02/Publishers-Association-Copyright-and-AI-Consultation-Response-25-February-2025.pdf [https://perma.cc/AAM2-EJVQ] [(listing publicly-reported licensing deals); Sara Gaglione, *2024 in Review: A Timeline of the Major Deals Between Publishers and AI Companies*, DIGIDAY (Dec. 27, 2024), https://digiday.com/media/2024-in-review-a-timeline-of-the-major-deals-between-publishers-and-ai-companies/   [https://perma.cc/9R7R-N6HB] [https://web.archive.org/web/20250323121234/https://digiday.com/media/2024-in-review-a-timeline-of-the-major-deals-between-publishers-and-ai-companies/]; Anna Tong et al., *Exclusive: Reddit in AI Content Licensing Deal with Google*, REUTERS (Feb. 22, 2024), https://www.reuters.com/technology/reddit-ai-content-licensing-deal-with-google-sources-say-2024-02-22/ (last visited Apr. 3, 2025); *Associated Press, OpenAI Partner To Explore Generative AI Use in News*, REUTERS (July 14, 2023), https://www.reuters.com/business/media-telecom/associated-press-openai-partner-explore-generative-ai-use-news-2023-07-13/ (last visited Apr. 3, 2025).

17.   Andrew Deck, *Dow Jones Negotiates AI Usage Agreements with Nearly 4,000 News Publishers*, NIEMANLAB (Dec. 5, 2024), https://www.niemanlab.org/2024/12/dow-jones-negotiates-ai-usage-agreements-with-nearly-4000-news-publishers/   [https://perma.cc/Q828-EP9H] [https://web.archive.org/web/20250125212457/https://www.niemanlab.org/2024/12/dow-jones-negotiates-ai-usage-agreements-with-nearly-4000-news-publishers/].

18.   17 U.S.C. § 107(4).

19.   *See, e.g.*, Princeton Univ. Press v. Mich. Document Servs., Inc., 99 F.3d 1381, 1387 (6th Cir. 1996) ("Where . . . the copyright holder clearly does have an interest in exploiting a licensing market—and especially where the copyright holder has actually succeeded in doing so—'it is appropriate that potential licensing revenues . . . be considered in a fair use analysis.'" (quoting Am. Geophysical Union v. Texaco Inc., 60 F.3d 913, 930 (2d Cir. 1994)).

Accordingly, they have been incentivized to avoid entering into licensing agreements, arguing that doing so would be impossible, and they have gone so far as to label license agreements as access agreements, perhaps to avoid a prejudicial finding on the fourth factor.[20]

One should not be distracted by Big Tech's claims about challenges with trying to license every copyright-protected work on the internet.[21] This is a red herring. Whether there are certain categories of nonprofessionally created copyrighted works that may be amenable to licensing solutions other than or in addition to market-based licensing merits further analysis (and is beyond the scope of this Article). However, works of professional content creators like publishers, the professional news media, and others clearly are among those that must be subject to market-based licensing for LLMs.

B.   ENTERING INTO THE NUMBER OF LICENSES REQUIRED FOR TRAINING LLMs ON THE WORKS OF PUBLISHERS AND NEWS/MEDIA ORGANIZATIONS IS FEASIBLE

Focusing on the works of professional publishers and news/media organizations—since they constitute the vast majority of copyright-protected professionally created works used for training LLMs by far—it is undeniable that they are susceptible to market-based licensing.[22]

Licensing publishers' works is feasible in terms of volume of titles and licenses. For example, regarding the volume of book titles, it has been reported that the Books3 pirate training dataset with around 190,000 books in the corpus was used by large GAI

---

20.   *See, e.g.*, Defendant Anthropic PBC's Opposition to Plaintiffs' Renewed Motion for Preliminary Injunction at 23, Concord Music Grp., Inc. v. Anthropic PBC, No. 24 Civ. 3811 (N.D. Cal. Aug. 22, 2024) ("OpenAI's fair use position casts doubt on [Plaintiffs' expert's] assumption that these deals must be about licensing data for LLM training."); *but see* Reply Declaration of Michael D. Smith in Support of Plaintiffs' Reply in Support of Motion for Preliminary Injunction ¶¶ 16–21, Concord Music Grp., Inc. v. Anthropic PBC, No. 24 Civ. 3811 (N.D. Cal. Sept. 12, 2024) (setting forth ample evidence from press coverage and press statements that the licenses include terms permitting AI training).

21.   *See, e.g.*, OpenAI, Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 13 (Oct. 30, 2023) [hereinafter, OpenAI Comments], https://www.regulations.gov/comment/COLC-2023-0006-8906 [https://perma.cc/ZF8M-TN55] [https://web.archive.org/web/20241108223225/https://www.regulations.gov/comment/COLC-2023-0006-8906] ("The diversity and scale of the information available on the internet is thus both necessary to training a 'well-educated' model . . . and also makes licensing every copyrightable work contained therein effectively impossible."); Declaration of Steven R. Peterson, PhD in Support of Defendant's Opposition to Plaintiffs' Renewed Motion for Preliminary Injunction ¶ 10, Concord Music Grp., Inc. v. Anthropic PBC, No. 24 Civ. 3811 (N.D. Cal. Aug. 22, 2024) ("In short, the claim that a market involving many thousands or millions of individual transactions to license huge volumes of copyrighted text covered by millions of copyrights would fail is unremarkable as a matter of economics.").

22.   It is also undeniable that licensing song lyrics for training LLMs is feasible. Anthropic concedes that "[t]here is an established market for the display of song lyrics on the internet . . . ." Defendant Anthropic PBC's Surresponse to Plaintiffs' Renewed Motion for Preliminary Injunction at 2, Concord Music Grp., Inc. v. Anthropic PBC, No. 24 Civ. 3811 (N.D. Cal. Oct. 23, 2024). While license terms for training would be different, the established market for display shows that the scope of works and rightsholders is not prohibitive.

developers to train their LLMs.[23] In comparison, a license with a single publisher could potentially cover more than 100,000 titles. As for the volume of licenses, for a sense of scale, the AAP has approximately 110 members across its trade, education, and professional and scholarly publishing sectors.[24] For training purposes within any given sector, negotiating licenses with some percentage of that subset of the industry is quite feasible.

Similarly, the volume of professional news content is equally susceptible to licensing. Licenses with fifteen news portfolio publishers[25] would cover 9.91% of the URLs in OpenWebText, 12.04% of the URLs in OpenWebText2, and 13.47% of the URLs in the top 1,000 domains in WebText.[26] With up to fifteen counterparties, GAI developers could easily license broad news media coverage for training purposes. Moreover, according to the News/Media Alliance, its members "stand ready to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy expressed content, something that will benefit all parties and society at large . . . ."[27]

Certain Big Tech (and related party) submissions in response to the U.S. Copyright Office's Notice of Inquiry and Request for Comments on Artificial Intelligence and Copyright acknowledge, albeit with varying degrees of directness, that it is the licensing of unprofessional works across the entire internet that they contend presents certain licensing challenges, not the licensing of professionally created works such as those of publishers and news/media organizations. For example:

---

23. Kate Knibbs, *The Battle Over Books3 Could Change AI Forever*, WIRED (Sept. 4, 2023), https://www.wired.com/story/battle-over-books3/ [https://perma.cc/UCA6-PYQF] [https://web.archive.org/web/20250323123006/https://www.wired.com/story/battle-over-books3/].

24. Ass'n of Am. Publishers, *Our Members*, https://publishers.org/who-we-are/our-members/ [https://perma.cc/PF7G-RQT5] [https://web.archive.org/web/20250120142631/https://publishers.org/who-we-are/our-members/] (last visited Jan. 20, 2025).

25. The fifteen portfolio publishers (with relevant subsidiary brands in parentheses) include: Advance (Condé Nast, Advance Local), Alden Global Capital (Tribune Publishing, MediaNews Group), Axel Springer, Bustle Digital Group, Buzzfeed, Inc., Future plc, Gannett, Hearst, IAC (Dotdash Meredith and other divisions), News Corp, The New York Times Company, Penske Media Corporation, Vox Media, The Washington Post, and Ziff Davis. George Wukoson & Joey Fortuna, *The Predominant Use of High-Authority Commercial Web Publisher Content To Train Leading LLMs*, at 9 (Nov. 4, 2024), https://www.ziffdavis.com/wp-content/uploads/2024/11/The-Predominant-Use-of-High-Authority-Commercial-Web-Publisher-Content-to-Train-Leading-LLMs.pdf [https://perma.cc/B2HD-ZM2D] [https://web.archive.org/web/20250214201601/https://www.ziffdavis.com/wp-content/uploads/2024/11/The-Predominant-Use-of-High-Authority-Commercial-Web-Publisher-Content-to-Train-Leading-LLMs.pdf].

26. *Id.* at 2–4, 11–12. Note that the WebText percentage is limited to the top 1,000 domain names because OpenAI only released a list of the top 1,000 domain names in that dataset. *Id.* at 3.

27. News/Media Alliance, Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 39–40 (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-8956 [https://perma.cc/UG28-MWHR] [https://web.archive.org/web/20250120142809/https://www.regulations.gov/comment/COLC-2023-0006-8956].

- Although Andreessen Horowitz initially states that "[t]he unique considerations involved in training AI models make direct, voluntary licensing impossible," it reveals that its claimed licensing concern is for nonprofessional content:

  > [T]he reason AI models are able to do what they can do today is that the internet has given AI developers ready access to a broad range of content, *much of which can't reasonably be licensed—everything from blog posts to social media threads to customer reviews on shopping sites.*[28]

- Anthropic observes that "a regime that *always* requires licensing for use of material in training would be inappropriate; it would, at a minimum, effectively lock up access to the vast majority of works, since *most works are not actively managed and licensed in any way.*"[29]
- Meta provides a hypothetical where it *concedes licensing of professionally created works including books* but makes the case for the importance of access to a broader array of "*user-generated or orphaned*" *content* (e.g., internet forum comments or memes) for fuller context around the subjects of the professionally created works.[30]
- OpenAI states that the "diversity and scale of the information available on the internet is thus both necessary to training a 'well-educated' model . . . and also makes licensing *every copyrightable work contained therein* effectively impossible."[31]

Likewise, in internal meetings at Meta, managers, lawyers, and engineers discussed that in the interest of speed they preferred to avoid negotiating licenses with publishers, artists, musicians, and the news industry, not that negotiating licenses would be infeasible.[32] And former employees of GAI developers like Ed Newton-Rex and Suchir Balaji have effectively acted as whistleblowers on their former employers' stances on unlicensed training by quitting in protest.[33]

---

28.   Andreessen Horowitz Comments, *supra* note 13, at 8–9 (emphasis added).

29.   Anthropic PBC, Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 9 (Oct. 30, 2023) (emphasis added) [hereinafter, Anthropic Comments], https://www.regulations.gov/comment/COLC-2023-0006-9021 [https://perma.cc/D9Z5-N822] [https://web.archive.org/web/20250120143216/https://www.regulations.gov/comment/COLC-2023-0006-9021].

30.   Meta Platforms, Inc., Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 4–5 (Dec. 6, 2023) (emphasis added) [hereinafter, Meta Comments], https://www.regulations.gov/comment/COLC-2023-0006-10332 [https://perma.cc/C485-C2ES] [https://web.archive.org/web/20250120143244/https://www.regulations.gov/comment/COLC-2023-0006-10332].

31.   OpenAI Comments, *supra* note 20, at 13 (emphasis added).

32.   *See, e.g.*, Metz et al., *supra* note 9.

33.   *See* Newton-Rex, *supra* note 5 and accompanying text; Cade Metz, *Former OpenAI Researcher Says the Company Broke Copyright Law*, N.Y. TIMES (Oct. 23, 2024), https://www.nytimes.com/2024/10/23/technology/openai-copyright-law.html

Finally, it should be noted that failure to reach agreement with certain rightsholders will not be prejudicial to the development of any LLM. As Meta has observed:

> [N]umerous comments agreed that the quality of Generative AI models does not depend on the expressive elements of a particular piece of content, or the inclusion or exclusion of any individual piece of training data, but rather on the *quantity* and *diversity* of the training content as a whole. Training data, in other words, is highly substitutable: as long as the model's overall training corpus is large and diverse, the model will function just as effectively with or without any specific piece of content.[34]

Consistent with this observation on the lack of prejudice from the exclusion of particular works, in practice "OpenAI has worked closely with both music publishers (including through the [National Music Publishers' Association]) and book authors (including through the Authors Guild) to identify third-party sites that may contain infringing or unauthorized copies of their works, and has specifically excluded sites from being part of training data."[35]

C. BIG TECH CAN AFFORD THE LICENSING FEES FOR PROFESSIONALLY CREATED WORKS

Big Tech can clearly afford to pay reasonable license fees to the copyright owners whose works are the very building blocks of GAI and whose livelihoods are threatened by the same system.[36] GAI developers count among their investors some of the largest and most profitable technology companies and venture capitalist firms in the world, and the value of the technology is astronomical. In October 2024, OpenAI reached a

---

[https://web.archive.org/web/20250208043720/https://www.nytimes.com/2024/10/23/technology/openai-copyright-law.html].

34.   Meta Comments, *supra* note 30, at 3.

35.   OpenAI, Reply Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 11 (Dec. 6, 2023), https://www.regulations.gov/comment/COLC-2023-0006-10340   [https://perma.cc/G8NY-247S] [https://web.archive.org/web/20250208044150/https://www.regulations.gov/comment/COLC-2023-0006-10340].

36.   Although ability to afford licenses is not relevant to the fourth statutory fair use factor, this Article posits nonetheless that affordability is not hindering licensing of professionally created works. *See* Copyright Alliance, Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 87 (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-8935 [https://perma.cc/PE8H-5JPY] [https://web.archive.org/web/20250208044618/https://www.regulations.gov/comment/COLC-2023-0006-8935]

> ("Notably, the fourth factor *does not require* courts to consider the economic impact of securing a license, but it does explicitly require them to consider the economic impact of *not* securing a license. The text of section 107(4) of the Copyright Act does not say anything about the potential impact to *the user or the user's market.* Instead, courts shall consider 'the effect of the use upon the potential market for or value *of the copyrighted work.*' This distinction is critical, and it is one that must be taken into account under any fair use analysis related to the unauthorize[d] ingestion of copyrighted works by AI developers.").

$157 billion valuation.[37] As of December 31, 2024, Microsoft's market capitalization exceeded $3.13 trillion,[38] Amazon's market capitalization exceeded $2.32 trillion,[39] Alphabet's market capitalization exceeded $2.32 trillion,[40] and Meta's market capitalization exceeded $1.48 trillion.[41]

Moreover, several Big Tech commentators have conceded that they could afford licenses for all copyrighted content used in training—above and beyond the professionally created works that are the focus of this Article.[42]

In other contexts, emerging technology companies repeatedly have said they could not afford licenses but have been proven wrong, with licensing and successful services prevailing.[43] The increasing number of agreements for training demonstrates that license rates and transaction costs do not make mutually agreeable licenses between

---

37.   Cade Metz, *OpenAI Completes Deal that Values Company at $157 Billion*, N.Y. TIMES (Oct. 2, 2024), https://www.nytimes.com/2024/10/02/technology/openai-valuation-150-billion.html?smid=url-share [https://web.archive.org/web/20250123225753/https://www.nytimes.com/2024/10/02/technology/openai-valuation-150-billion.html?smid=url-share].

38.   *Microsoft Corporation (MSFT) Statistics*, YAHOO FIN., https://finance.yahoo.com/quote/MSFT/key-statistics/ [https://web.archive.org/web/20250323123818/https://finance.yahoo.com/quote/MSFT/key-statistics/] (last visited Mar. 23, 2025).

39.   *Amazon.com, Inc. (AMZN) Statistics*, YAHOO FIN., https://finance.yahoo.com/quote/AMZN/key-statistics/ [https://web.archive.org/web/20250323131010/https://finance.yahoo.com/quote/AMZN/key-statistics/] (last visited Mar. 23, 2025). Amazon has invested $8 billion USD in Anthropic. *See* Daniel Tencer, *Anthropic Lands Another $4BN Investment from Amazon, Amid Ongoing Copyright Battle with Universal, Concord and ABKCO*, MUSIC BUS. WORLDWIDE (Nov. 25, 2024), https://www.musicbusinessworldwide.com/anthropic-just-landed-another-4bn-investment-from-amazon-amid-its-ongoing-copyright-fight-with-universal-concord-and-abkco/ [https://perma.cc/N7QV-U5MN] [https://web.archive.org/web/20241127035502/https://www.musicbusinessworldwide.com/anthropic-just-landed-another-4bn-investment-from-amazon-amid-its-ongoing-copyright-fight-with-universal-concord-and-abkco/].

40.   *Alphabet Inc. (GOOG) Statistics*, YAHOO FIN., https://finance.yahoo.com/quote/GOOG/key-statistics/ [https://web.archive.org/web/20250323131014/https://finance.yahoo.com/quote/GOOG/key-statistics/] (last visited Mar. 25, 2025).

41.   *Meta Platforms, Inc. (META) Statistics*, YAHOO FIN., https://finance.yahoo.com/quote/META/key-statistics/ [https://web.archive.org/web/20250323131543/https://finance.yahoo.com/quote/META/key-statistics/] (last visited Mar. 25, 2025).

42.   *See* Andreessen Horowitz Comments, *supra* note 13, at 8 ("A multi-billion-dollar company might be able to afford to license copyrighted training data, but smaller, more agile startups will be shut out of the development race entirely."); Anthropic Comments, *supra* note 29, at 10 ("Only the most highly resourced entities would be able to engage in costly and burdensome data licensing processes."). Although the essential point has been conceded, it is also worth noting that the supposed concerns of Big Tech about small startups are unconvincing and misleading. *See* Kevin Madigan, *Generative AI Licensing Isn't Just Possible, It's Essential*, COPYRIGHT ALL. (Nov. 21, 2024), https://copyrightalliance.org/generative-ai-licensing/ [https://perma.cc/RUA7-FJBM] [https://web.archive.org/web/20250123225921/https://copyrightalliance.org/generative-ai-licensing/].

43.   Declaration of Michael D. Smith in Support of Plaintiffs' Motion for Preliminary Injunction, Concord Music Grp., Inc. v. Anthropic PBC, No. 24 Civ. 3811, ¶ 72 (N.D. Cal. Aug. 1, 2024) (citing, among other things, (i) internet radio (e.g., Pandora); (ii) social media platforms (e.g., YouTube and Instagram); and (iii) streaming media (e.g., Apple Music, Spotify, and Deezer)).

rightsholders and GAI developers impractical.[44] Inasmuch as copyrighted works are as integral to GAI technologies as they are to services like Netflix and Spotify, both of which license the works made available through their platforms, GAI developers are capable of doing the same, and publishers are willing to work with them to effect such licensing.

III. PUBLISHERS WELCOME ALL MARKET-BASED LICENSING OPTIONS FOR TRAINING LLMS WITH PUBLISHERS' WORKS

The marketplace for publishers' works is sufficiently flexible to accommodate licenses for training through both existing licensing practices and new licensing models based on the exclusive rights afforded by copyright law. For publishers, licensing is about marketplace innovation and competition. The drafters of the Copyright Act set forth the exclusive rights in broad terms "to ensure that copyright protection would encompass not only the breadth of technological uses known at the time of enactment, but also future technological uses."[45] Time and time again publishers have vindicated the drafters' intentions by adapting their licensing models to the technologies of the day. For example, publishers have innovated in parallel with the development of the internet and handheld devices by making books available in eBook and/or digital audiobook formats for lending, for a subscription term, or for purchase of a usage license. Professional and scholarly publishers have a multitude of licensing options for their journal databases, including licenses for text-and-data mining purposes that have existed for several years.[46]

---

44.   *See, e.g.*, Reply Declaration of Ed Newton-Rex in Support of Plaintiffs' Reply in Support of Motion for Preliminary Injunction, Concord Music Grp., Inc. v. Anthropic PBC, No. 24 Civ. 3811, ¶ 9 (N.D. Cal. Sept. 12, 2024).

45.   Ass'n of Am. Publishers, Reply Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 7–8 (Dec. 6, 2023), https://www.regulations.gov/comment/COLC-2023-0006-10298 [https://perma.cc/DR9K-PU28] [https://web.archive.org/web/20250208055757/https://www.regulations.gov/comment/COLC-2023-0006-10298] (citing H. COMM. ON THE JUDICIARY, 89TH CONG., SUPPLEMENTARY REGISTER'S REPORT ON THE GENERAL REVISION OF THE U.S. COPYRIGHT LAW 18 (Comm. Print 1965)).

46.   *See, e.g.*, AAP Comments, *supra* note 7, at 5 n.9 and accompanying text (citing *Text & Data Mining*, ACS PUBL'NS, https://solutions.acs.org/solutions/text-and-data-mining [https://perma.cc/HCM5-LKCE] [https://web.archive.org/web/20250208060101/https://solutions.acs.org/solutions/text-and-data-mining/] (last visited Feb. 8, 2025); *Elsevier Text and Data Mining (TDM) License*, ELSEVIER, https://beta.elsevier.com/about/policies-and-standards/text-and-data-mining/license?trial=true [https://perma.cc/59S6-ZBJU] [https://web.archive.org/web/20250208060209/https://www.elsevier.com/about/policies-and-standards/text-and-data-mining/license?trial=true] (last visited Feb. 8, 2025); *Sage Journals Text and Data Mining License Agreement*, SAGE JS., https://journals.sagepub.com/page/policies/text-and-data-mining-license [https://web.archive.org/web/20250208060459/https://journals.sagepub.com/page/policies/text-and-data-mining-license] (last visited Feb. 8, 2025); *Text and Data Mining*, TAYLOR & FRANCIS, https://taylorandfrancis.com/our-policies/textanddatamining [https://perma.cc/6M8J-UPKX] [https://web.archive.org/web/20250208060624/https://taylorandfrancis.com/our-policies/textanddatamining/] (last visited Feb. 8, 2025); *Text and Data Mining*, WILEY, https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining

While the latest technological advance, GAI, holds extraordinary potential to transform society, there is nothing extraordinary about the licensing marketplace for training GAI models on publishers' works. Publishers welcome all voluntary marketplace developments. Market-based licensing solutions are the superior tool for facilitating development of GAI systems while respecting and protecting the exclusive rights of authors and publishers. Fundamentally, authors and publishers should remain free to exercise their exclusive rights, to determine how and in what ways their works are to be used, and by whom, and to not exercise them if the circumstances and/or terms do not merit a deal. Publishers and GAI developers have the flexibility to decide whether direct licensing or voluntary collective licensing is best suited for the particular circumstances of any deal they wish to conclude.

Voluntary collective licensing has worked well for publishers in certain sectors for certain specific use cases, such as corporate licensing for certain uses of works of professional and scholarly publishers, and academic licensing for certain uses of works of educational publishers. For example, the Copyright Clearance Center ("CCC") offers non-exclusive voluntary collective licensing solutions that include the Annual Copyright License for Business,[47] Multinational Copyright License for Business,[48] Annual Copyright License for Higher Education,[49] Annual Copyright License for Curriculum & Instruction,[50] and Annual Copyright License for Student Assessments.[51] In July 2024, the CCC announced a new licensing option within its Annual Copyright

---

[https://web.archive.org/web/20250211082002/https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining] (last visited Feb. 11, 2025)).

47. *Annual Copyright License*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/solutions-annual-copyright-license/ [https://perma.cc/KNK8-ZSLT] [https://web.archive.org/web/20250123230115/https://www.copyright.com/web/20250123230115/https://www.copyright.com/solutions-annual-copyright-license/] (last visited Feb. 8, 2025).

48. *Multinational Copyright License*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/wp-content/uploads/2018/09/Product-Sheet-Multinational-License.pdf [https://perma.cc/EL9U-7DGH] [https://web.archive.org/web/20250123230153/https://www.copyright.com/wp-content/uploads/2018/09/Product-Sheet-Multinational-License.pdf] (last visited Feb. 8, 2025).

49. *Annual Copyright License for Higher Education*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/solutions-annual-copyright-license-higher-education/ [https://perma.cc/T9WQ-ZL4Z] [https://web.archive.org/web/20250123230137/https://www.copyright.com/web/20250123230137/https://www.copyright.com/solutions-annual-copyright-license-for-curriculum-instruction/] (last visited Feb. 8, 2025).

50. *Annual Copyright License for Curriculum & Instruction*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/solutions-annual-copyright-license-for-curriculum-instruction/ [https://perma.cc/5GWU-UKYU] [https://web.archive.org/web/20250123230137/https://www.copyright.com/web/20250123230137/https://www.copyright.com/solutions-annual-copyright-license-for-curriculum-instruction/] (last visited Feb. 8, 2025).

51. *Annual Copyright License for Student Assessments*, COPYRIGHT CLEARANCE CTR., https://www.copyright.com/solutions-annual-copyright-license-student-assessments/ [https://perma.cc/6PRH-5GL8] [https://web.archive.org/web/20250130021543/https://www.copyright.com/web/20250130021543/https://www.copyright.com/solutions-annual-copyright-license-student-assessments/] (last visited Jan. 29, 2025).

License that permits the internal use of copyrighted materials in AI systems—the first ever voluntary collective licensing solution for this use.[52]

Notwithstanding the foregoing, for most publishing houses and in most circumstances, their experience and preference is direct licensing. It is likely that in the majority of circumstances publishers will prefer direct licensing with GAI developers. Publicly available sources have reported that publishers, including DeGruyer Brill,[53] HarperCollins,[54] Informa (parent of Taylor & Francis),[55] Oxford University Press, and Wiley,[56] have each concluded one or more licensing deals. Outside of publishing, there have also been reported over forty other content licensing agreements with LLMs, including with Associated Press, The Atlantic, Axel Springer, Condé Nast, Dotdash Meredith, Financial Times, LA Times, Le Monde, News Corp., Prisa Media, Reddit, Reuters, Time, and Vox Media, plus the nearly 4,000 licensing agreements Dow Jones concluded.[57] Like licensing markets before it, the licensing marketplace for GAI is growing and developing.[58]

There are also startups such as Calliope Networks, Created by Humans, and Human Native AI that are seeking to facilitate licenses, particularly for smaller AI developers.[59] The Authors Guild, the nation's oldest and largest professional organization for

---

52. Press Release, *CCC Pioneers Collective Licensing Solution for Content Usage in Internal AI Systems*, COPYRIGHT CLEARANCE CTR. (July 16, 2024), https://www.copyright.com/media-press-releases/ccc-pioneers-collective-licensing-solution-for-content-usage-in-internal-ai-systems/ [https://perma.cc/B4S7-39UT] [https://web.archive.org/web/20250130022544/https://www.copyright.com/web/20250130022544/https://www.copyright.com/media-press-releases/ccc-pioneers-collective-licensing-solution-for-content-usage-in-internal-ai-systems/].

53. *See* PUBLISHERS ASS'N, *supra* note 16, at 24.

54. Ella Creamer, *HarperCollins To Allow Tech Firms To Use Its Books To Train AI Models*, GUARDIAN (Nov. 19, 2024), https://www.theguardian.com/books/2024/nov/19/harpercollins-tech-firms-books-train-ai-models-nonfiction-artificial-intelligence [https://perma.cc/E4PJ-UC4E] [https://web.archive.org/web/20250130023512/https://www.theguardian.com/books/2024/nov/19/harpercollins-tech-firms-books-train-ai-models-nonfiction-artificial-intelligence].

55. Press Release, *Market Update*, INFORMA PLC (May 8, 2024), https://www.informa.com/globalassets/documents/investor-relations/2024/informa-plc—-market-update.pdf [https://perma.cc/D2JP-YKN5] [https://web.archive.org/web/20250223154404/https://www.informa.com/globalassets/documents/investor-relations/2024/informa-plc%E2%80%94-market-update.pdf].

56. Heloise Wood, *Wiley and Oxford University Press Confirm AI Partnerships as Cambridge University Press Offers 'Opt-in'*, BOOKSELLER (Aug. 1, 2024), https://www.thebookseller.com/news/wiley-cambridge-university-press-and-oxford-university-press-confirm-ai-partnerships [https://perma.cc/ST7R-DD3F] [https://web.archive.org/web/20250130025539/https://www.thebookseller.com/news/wiley-cambridge-university-press-and-oxford-university-press-confirm-ai-partnerships].

57. *See* sources cited *supra* note 16; Deck, *supra* note 17.

58. Beyond LLMs, there have also been many publicly reported licensing agreements for images, music, and/or video for GAI models.

59. CALLIOPE NETWORKS, https://calliopenetworks.ai/ [https://perma.cc/S6LP-5TC9] [https://web.archive.org/web/20250124101832/https://calliopenetworks.ai/] (last visited Jan. 29, 2025); CREATED BY HUMANS, https://www.createdbyhumans.ai/ [https://perma.cc/3XF5-GM7B] [https://web.archive.org/web/20250130030651/https://www.createdbyhumans.ai/] (last visited Jan. 29, 2025); HUMAN NATIVE AI, https://www.humannative.ai/ [https://perma.cc/NQZ2-X78M] [https://web.archive.org/web/20250130031112/https://www.humannative.ai/] (last visited Jan. 29, 2025).

published writers, announced an official partnership with Created by Humans in October 2024.[60] Whether through one of these models or a different voluntary licensing model, publishers will innovate and compete to serve smaller AI developers too.

IV. MARKET-BASED LICENSING FOR PUBLISHERS' WORKS BEST SERVES THE PUBLIC INTEREST

Market-based licensing for publishers' works best serves the public interest, benefiting all stakeholders, including the GAI developer, downstream users, authors, publishers, and the general public:

Rights Certainty. Publishers, authors, GAI developers, and the downstream users of GAI technologies can have certainty with respect to rights and obligations, including how copyrighted works can be used to train GAI systems and be incorporated in outputs.

Sustaining America's Political, Intellectual, and Cultural Systems Through New Works. Market-rate licensing fees are an important source of income for U.S. authors and publishers, and best promote their investment in new, high-quality, human-created works that will sustain the nation's political, intellectual, and cultural systems, ultimately benefiting the public.[61]

Better Performing LLMs. Payment of market-rate licensing fees by GAI developers is also in the long-term interest of GAI developers, downstream users, and the general public because the new, high-quality, human-created works that licensing fees support will fuel higher quality performance from future LLMs on a continuing, ongoing basis.[62]

Higher AI Reliability. Publications licensed from authorized sources are more reliable. For example, in the case of professional and scholarly journal articles, it is important for any AI system to use the Version of Record ("VoR") with appropriate licensing. Earlier versions or pirated versions may be subject to post publication modification or retraction, which could create serious and cascading scientific or medical errors in AI generated outputs.[63] All stakeholders in the AI ecosystem benefit.

High Human Risk Demands Highest Quality. Since AI technologies are being and will be used in ways that impact the lives and well-being of individuals, whether financially, physically, mentally, or professionally, it is critically important that the highest quality textual works are used to create current and future training sets.[64]

---

60.   Press Release, *Authors Guild Partners with Created by Humans To Empower Authors in the AI Era*, AUTHORS GUILD (Oct. 9, 2024), https://authorsguild.org/news/ag-partners-with-created-by-humans-to-empower-authors-in-ai-era/   [https://perma.cc/8SCC-SZLS] [https://web.archive.org/web/20250130031429/https://authorsguild.org/news/ag-partners-with-created-by-humans-to-empower-authors-in-ai-era/].

61.   AAP Comments, *supra* note 7, at 27.

62.   *See id.*

63.   *Id.* at 27–28.

64.   *Id.* at 28.

V. PUBLISHERS STRONGLY OPPOSE THE INTRODUCTION OF A COMPULSORY LICENSING REGIME OR EXTENDED COLLECTIVE LICENSING REGIME AS RELATED TO PUBLISHERS' WORKS

Publishers strongly oppose the introduction of a compulsory licensing regime or extended collective licensing to redress unauthorized uses of publishers' works to train GAI systems. As AAP stated in its comments to the U.S. Copyright Office in response to its Copyright and AI Study:

> The Copyright Office has correctly observed, "Compulsory licensing has been, and should be, regarded as an extreme last resort in copyright law," and it has held this view firmly for decades. Congress shares this view, recognizing that compulsory licenses fly in the face of the exclusive rights of authors, and it has enacted compulsory licenses reluctantly, sparingly, and narrowly—and only after clear and demonstrated need."[65]

It further observed that "[b]ecause extended collective licensing also acts in derogation of the exclusive rights of copyright owners, it raises many of the same concerns as compulsory licensing. As with compulsory licensing, extended collective licensing is not currently needed or justified."[66]

Fundamentally, authors and publishers should remain free to exercise their exclusive rights, to determine how and in what ways their works are to be used, and by whom. There has also been no market failure with respect to publishers' works that would necessitate such an "extreme last resort" as compulsory licensing or would otherwise necessitate extended collective licensing.[67] To the contrary, licensing publishers' works is eminently feasible for the reasons explained in this Article.

VI. CONCLUSION

Publishers welcome market-based licensing solutions for facilitating development of GAI systems because they benefit all stakeholders, they are undoubtedly feasible, and

---

65. *Id.* at 25 (internal citations omitted).
66. *Id.* at 26.
67. *Id.* at 25 (quoting U.S. COPYRIGHT OFF., SECOND SUPPLEMENTARY REGISTER'S REPORT ON THE GENERAL REVISION OF THE U.S. COPYRIGHT LAW 60 (1975) https://ipmall.law.unh.edu/sites/default/files/hosted_resources/lipa/copyrights/Second%20Supplementary%20Register's%20Report%20on%20the%20General%20Revision.pdf [https://web.archive.org/web/20240725150851/https://ipmall.law.unh.edu/sites/default/files/hosted_resources/lipa/copyrights/Second%20Supplementary%20Register's%20Report%20on%20the%20General%20Revision.pdf]). Extended collective licensing also "presupposes the existence of a voluntary license between relevant parties representing a substantial majority of both licensors and licensees for a specifically-described, licensable use that, for explicit reasons, is unable to encompass all necessary works of potential-but-absent licensors." *See* Copyright Clearance Center, Comment Letter on the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, at 13 (Oct. 30, 2023), https://www.regulations.gov/comment/COLC-2023-0006-8601 [https://perma.cc/SPD4-2YUS] [https://web.archive.org/web/20250130031930/https://www.regulations.gov/comment/COLC-2023-0006-8601]. These circumstances do not exist in the market for publishers' works at the timing of writing and do not appear likely to exist in the future.

they will help ensure that publishers and authors may continue to make their vital contributions to society. The licensing market is rapidly growing, and but for the utter disregard towards the rights of copyright owners by LLM developers hiding behind the assertion of fair use, it would be even larger. Whether the non-professionally created copyright works that comprise the vast majority of content on the internet present unique challenges for market-based licensing is an issue on which this Article takes no opinion but to note that it merits further study. Instead, this Article concludes that the works of professional content creators like publishers, the professional news media, and others clearly are among those that must be subject to market-based licensing for LLM training and use. The number of licenses is manageable, and Big Tech can afford them. Big Tech would seem to agree.