Maxwell V. Pritt
Written Answers to Questions for the Record

Senate Judiciary Subcommittee on Crime and Counterterrorism
Hearing Titled "Too Big to Prosecute?: Examining the AI Industry's Mass Ingestion of Copyrighted Works for AI Training"

August 6, 2025

<u>**Questions Submitted by Senator Amy Klobuchar**</u>:

Court filings from *Kadrey v. Meta* showed that Meta spoke with multiple companies about licensing training materials, such as books and research papers, but later decided against it because it would be "unreasonably expensive" and "incredibly slow."

1. **Meta employs more than 70,000 people and earned more than $60 billion in profits just last year. Do you believe it is possible for well-resourced companies like Meta to license and pay for high-quality content to train their models?**

Yes. It is not just possible, but in fact is already happening on a large scale. Many AI developers currently license copyrighted material for AI training, including Meta itself in certain circumstances.

Some AI companies have argued that licensing copyrighted works for use with generative AI systems is impossible due to the large amount of material needed to train a model. That self-serving argument ignores the plethora of licensing solutions that already exist and continue developing to meet market demand. The fact that companies like Meta *prefer* to pirate content for free says nothing about the feasibility of paying a fair price for that content. As one federal district court recently put it: "[T]he suggestion that adverse copyright rulings would stop this technology in its tracks is ridiculous. These products are expected to generate billions, even trillions, of dollars for the companies that are developing them. If using copyrighted works to train the models is as necessary as the companies say, they will figure out a way to compensate copyright holders for it."[1]

Meta specifically is better positioned than most companies to pay prevailing market rates for licensing copyrighted content for internal and external uses with its commercial AI models. And that's true even if Meta could substantiate its claim that doing so would be expensive. Internal documents show Meta was prepared to spend hundreds of millions of dollars on licensing copyrighted content for its AI models before it resorted to just pirating that content instead. Meta at one point discussed a $200 million licensing budget, with half of that sum earmarked for licensing books.[2] Apart from data acquisition, Meta has spent astronomical sums on its AI program, including on data infrastructure and talent. Recent reporting shows that Meta pledged hundreds of *billions* of dollars to build AI data centers, invested tens of billions of dollars in deals with AI startups, and

---

[1] *Kadrey, et al. v. Meta Platforms, Inc.*, No. 23-CV-03217-VC, 2025 WL 1752484, at *2 (N.D. Cal., June 25, 2025).

[2] *Kadrey, et al. v. Meta Platforms, Inc.,* No. 23-CV-03417-VC, Dkt. 574 (Pls' Mot. for Partial Summary Judgment) at 8.

offered a $250 million compensation package to a single AI researcher.[3] Meta also projects enormous profit margins for its AI products. Meta's revenue projections for its AI program through 2035 range from a "Base Case" of $460 billion, to a "GenAI Wins Case" of $1.4 *trillion*.[4] Ability-to-pay is not an issue, and the notion that licensing content is prohibitively expensive for a company like Meta is preposterous. If Meta is willing to spend hundreds of millions of dollars to recruit a few AI researchers and hundreds billions of dollars to build AI data centers, then paying hundreds of millions of dollars, or even several billion dollars, to the reporters, authors, publishers, and others in the creative community whose works Meta used to build its AI models is hardly "unreasonably" expensive.

Other large technology companies have already entered into licensing deals to use copyrighted content with their AI systems. For example, in November 2024, Microsoft contracted to license copyrighted works from HarperCollins.[5] More recently, Amazon entered into a similar deal with The New York Times.[6] These contracts demonstrate the feasibility of large-scale licensing. Meta's internal documents also show it knows licensing copyrighted content for use with its commercial AI models is a viable option. The company at one point planned to acquire as much as 20% of its Llama 4 text data corpus through licensed content.[7] However, Meta's licensing strategy remains limited because it still employs what it calls the "gap approach"—pirate as much copyrighted content as possible, and only then use licensing to fill in the gaps of content that cannot be pirated.[8]

With respect to Meta's and other companies' argument that licensing content for use with their AI models is too slow, even assuming the companies devoted adequate resources to licensing (which Meta did not), it is not surprising that respecting intellectual property rights and complying with the law could take longer than breaking it. Naturally, that does not justify the latter.[9] Certainly Meta would not argue that OpenAI could steal its trade secrets because they helped it develop AI

---

[3] *See*, e.g., Jaspreet Singh and Aditya Soni, *Meta's Zuckerberg pledges hundreds of billions for AI data centers in superintelligence push*, REUTERS (July 14, 2025); Billy Perrigo, *How Meta's $14 Billion Scale AI Investment Upended the AI Data Industry*, TIME (June 16, 2025); Mike Isaac, Eli Tan and Cade Metz, *A.I. Researchers Are Negotiating $250 Million Pay Packages. Just Like N.B.A. Stars.*, N.Y. TIMES (July 31, 2025).

[4] *Kadrey v. Meta,* Pls' Mot. for Partial Summary Judgment, Dkt. 574 at 4.

[5] Hannah Miller and Dina Bass, *Microsoft Signs AI-Learning Deal With News Corp.'s HarperCollins*, BLOOMBERG (Nov. 19, 2024).

[6] Alexandra Bruell, *Amazon to Pay New York Times at Least $20 Million a Year in AI Deal,* WALL STREET JOURNAL (July 30, 2025).

[7] *Kadrey v. Meta,* Pls' Mot. for Partial Summary Judgment, Dkt. 537 at 29.

[8] *Kadrey v. Meta,* Pls' Mot. for Partial Summary Judgment, Dkt. 588-1 at 65.

[9] *See*, e.g., *Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd.*, 454 F. Supp. 2d 966, 989 (C.D. Cal. 2006) (StreamCast also blames Plaintiffs for their difficult licensing terms, which StreamCast believes prevented it from launching a successful, legal business with licensed content. . . . Whatever its subjective intentions were about eventually securing licenses and developing revenue streams that did not depend on infringement, the business that actually materialized was one that thrived only because of the massive infringement enabled by Morpheus and OpenNap/MusicCity.")

models faster. The desire to try to keep pace with competitors cannot justify the AI industry's collective decision to "YOLO the legal risk"[10] and commit domestic online piracy at a staggering scale.

2. Are there licensing models that could fairly compensate creators without unnecessarily delaying or hampering AI innovation?

Yes. In addition to a growing number of one-to-one deals between established copyright-holding companies and generative AI developers, collective licensing is available to address issues of scalability. The U.S. Copyright Office conducted a detailed study of this question in its Report on Copyright and Artificial Intelligence titled "Generative AI Training" (the "Report"),[11] finding "available information shows that [licensing] markets exist or are 'reasonable' or 'likely to be developed[.]'"[12]

There is already high demand for corpora of copyrighted works for ingestion by AI systems, and, as discussed above, copyright holders are offering and entering into various licensing agreements. Publishers and copyright holders of scientific and research works such as Elsevier, JSTOR, the Copyright Clearance Center, and many others have either offered or entered into licensing agreements that allow for text and data mining (TDM) or other generative AI uses. Getty Images has struck several licensing deals with generative AI companies for use of portions of its catalog of stock images for training. Multiple news organizations, including NewsCorp, the Associated Press, The Atlantic, The New York Times, and the Financial Times, have reached deals with various AI developers. The list goes on and on, with new licensing deals being announced almost daily.[13]

Importantly, collective licensing is nothing new—it has proven feasible in many contexts and has readily adapted to new uses. With respect to literary works, as just one example, the Copyright Clearance Center was founded in 1978 with the aim of facilitating photocopying permissions in academic settings, and it has been undeniably successful at distributing royalties at scale.[14] Similarly, Performing Rights Organizations (PROs) collect and distribute monies for

---

[10] "YOLO" being a common slang term for "you only live once", so "why worry about the consequences?" *See Tremblay v. OpenAI*, No. 3:23-cv-0322 (N.D. Cal. March 13, 2025), Dkt. 392-8 (Pl's Proposed Second Amended Consolidated Complaint) at 15.

[11] U.S. Copyright Office Copyright and Artificial Intelligence, Part 3: Generative AI Training (May 9, 2025), at 70.

[12] *Id.* at 70.

[13] Copyright Alliance, *AI Licensing for Creative Works*, https://copyrightalliance.org/artificial-intelligence-copyright/licensing/.

[14] Mark Seeley, *Evolution of Copyright Law from Guild and Printing Monopolies to Human and Natural Rights*, https://www.copyright.com/wp-content/uploads/2021/01/CCC_CreatingSolutionsTogether_Ebook_2020.pdf, at 25.

musicians where it would otherwise be difficult or inefficient to directly license public performance permissions.[15]

While licensing for internal and external uses in connection with generative AI systems is still in its early stages, the information already available shows there is a clear path towards voluntary licensing that would allow copyright owners to control their works and earn incremental revenue for commercial exploitation of their works by the AI industry. While there isn't a one-size-fits-all solution to licensing for AI systems, there is no reason to doubt that major industry players can develop mutually beneficial solutions so that creators and rightsholders can share in the massive profits expected by the generative AI industry.[16] The feasibility of collective licensing is also demonstrated by models like Audible and Spotify Audiobooks, which already license books at scale.

As happened in the music industry in the 2000s, once online piracy is legally prohibited, market forces react naturally by developing legitimate alternatives. Shortly after Napster was enjoined, record companies made deals with Internet platforms and streaming services to distribute their music.[17] Apple's iTunes proliferated immediately in Napster's aftermath. Streaming models like Pandora and Spotify followed shortly thereafter. The lesson from the music industry is clear: once major participants in pirated markets are forced to use legitimate alternatives to obtain copyrighted content, those markets develop rapidly, including functional systems of collective licensing.[18] In light of the already growing market for licensing copyrighted books and other content

---

[15] Issues Related to Performing Rights Organizations, Comments of the Copyright Alliance, https://copyrightalliance.org/wp-content/uploads/2025/04/AS-SUBMITTED-Copyright-Alliance-Comments_NOI-PRO.pdf, at 2.

[16] *See Kadrey v. Meta*, 2025 WL 1752484, at *22 ("Meta argues that the 'public interest' would be 'badly disserved' by preventing Meta (and other AI developers) from using copyrighted text as training data without paying to do so. Meta seems to imply that such a ruling would stop the development of LLMs and other generative AI technologies in its tracks. This is nonsense. As mentioned earlier, a ruling that certain copying isn't fair use doesn't necessarily mean the copier has to stop their copying—it means that they have to get permission for it. So where copying for LLM training isn't fair use, LLM developers (including Meta) won't need to stop using copyrighted works to train their models. They will need only to pay rightsholders for licenses for that training. Presumably, where copying for AI training isn't fair use, AI developers will simply figure out a way to license the works they wish to use as training data. Meta's contention that markets for this licensing can't or won't develop is hard to believe. If books are as good for LLM training as Meta says they are, then it seems nearly certain that LLM developers would be willing to pay for licenses. (Indeed, Meta itself was willing to pay to license books—it just found licensing too logistically difficult.) Even if the value of any particular book as training data is too low to justify negotiating licensing deals book by book, LLM developers would still presumably be interested in licensing large numbers of books at once .. . . So if it isn't fair use for Meta and other LLM developers to use copyrighted books as training data without permission, they won't have to stop working on their LLMs altogether. They'll just have to pay for licenses or use books that aren't copyrighted. Either way, it may be that LLM companies move somewhat more slowly or make somewhat less money. But the suggestion that the growth of LLM technology would come to a halt (or anything close) doesn't pass the straight face test.").

[17] *See* 1 Lindey on Entertainment, Publ. & the Arts § 2:28 n. 36 (3d ed. 2024).

[18] *See* Jonathan M. Barnett, *The Big Steal: Ideology, Interest, and the Undoing of Intellectual Property* 337 (2024) ("As illustrated by the rise of licensed music and video streaming services, the performance of real-world digital content environments shows that well-functioning markets that support a robust flow of content

for use with AI systems, there is little reason to doubt that a thriving licensing market will continue to develop.

production are generally compelled to assemble a property-rights infrastructure—understood broadly to encompass legal, technological, and contractual devices that enable content owners to regulate and price access to some significant extent. The same argument can be made for licensed platforms in electronic books, digital images, and other creative media.").