Before the
U.S. COPYRIGHT OFFICE

Artificial Intelligence and Copyright	Docket No. 2023–6 Submitted December 6, 2023

REPLY COMMENTS OF THE NEWS/MEDIA ALLIANCE

Contents

1. Introduction

News/Media Alliance (N/MA) welcomes the opportunity to provide these reply comments for the U.S. Copyright Office's study on artificial intelligence (AI). The overwhelming initial response to the Office's Notice illustrates the wide-ranging ways and interest in how generative AI technology may reshape our society, the production and dissemination of creative works, and businesses that power generative AI models and applications.

Unfortunately, some comments also reveal serious misunderstandings and disregard of basic copyright principles by well-heeled developers and financiers, who are incentivized to discount these principles for private gain. An underlying fallacy in many such comments was the assertion of a conflict between the public good purportedly represented by the AI developers on one side and the interests of copyright owners on the other. Developers wrote at length about how positive AI is for society, conflating *generative* uses that take from and compete with copyrighted expression with deployments of AI more generally. N/MA members are certainly not against AI; on the contrary, we strongly support responsible AI technologies. To ensure the benefits of generative AI are more broadly realized, shared, and developed sustainably, however, N/MA asks the Office to provide a tailored, yet direct, clarification that the ingestion of copyright protected content for commercial generative AI training purposes is indeed "reproduced" and is not typically going to be a fair use.

At present, most benefits of generative AI are yet to be realized – and Silicon Valley itself is consumed by internal discussions over AI's prioritization of profits over people. As veteran reporter Kara Swisher put it, "[o]n the hype side, let's try to tone down bountiful future nonsense — we've heard it before and only some people got obscenely wealthy while the rest of us got the bill for the problems. AI will be great and it could be awful — it's complex, so try to be an adult about it."[1] Meanwhile, the dissemination of professional journalism is a cherished public good, with the essential democratic function of the Press enshrined in the Constitution. Public policy conversations should give heavy weight to the risk that this established public interest will be undermined by generative AI development that is parasitic, lacking accountability, and dodging compensation for the media content that fuels these models.

While copyright, creativity, and technology have a long, successful history of evolving together, it does not serve either creators or innovators when the law is interpreted so narrowly as to undermine the goals of copyright, and Congress occasionally must step in if that happens.[2] But typically, existing copyright law ably serves as a check on "too clever by half" efforts to design-around the general, permission-based framework, and the Office can help guide a similarly productive dialogue here.[3]

N/MA's initial comments addressed in detail many questions posed by the Office. This reply will not repeat that forensic, business, and legal analysis. Rather, we wish to pinpoint areas of growing consensus to help guide the Office moving forward; we also respond to select, core misconceptions raised by some commenters.

Overall, N/MA finds considerable support for the recommendations expressed in our initial submission. These recommendations, reiterated here, would also benefit generative AI

[1] @karaswisher, X (Nov. 21, 2023, 9:20 PM), https://twitter.com/karaswisher/status/1727074585336799395.

[2] *Compare White-Smith Music Publishing Co. v. Apollo Co*., 209 US 1 (1908) (finding that player pianos did not make "copies" of musical compositions because they were imperceptible to humans, a decision that was reversed by the 1909 Copyright Act).

[3] *See, e.g., American Broadcasting Cos., Inc. v. Aereo, Inc.*, 545 U.S. 913, at 934 (2014); *Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd.*, 545 U.S. 913, at 934 (2005); *Capitol Records, LLC v. ReDigi Inc*, 910 F.3d 649 (2d. Cir. 2018).

developers and technologies, enabling development in a sustainable manner and increasing legal certainty over their rights and obligations, while safeguarding creativity and an innovative, trustworthy, and vibrant online economy:

- **Infringement and Fair Use:** With regards to the infringing unauthorized use of publishers' expressive content for commercial generative AI training and development, there is little question – based on comments by rightsholders and developers alike – that a copy of protected, creative content is made without permission for AI training purposes. Such copying constitutes prima facie infringement. The bulk of disagreement centers around the application of the fair use defense. As explained in our White Paper, developers of Large Language Model (LLM) systems that use copyright protected media content without permission to power those systems are exceeding the bounds of fair use. Considering the systemic nature and clear harm caused by such uses, we urge the Office to deliver a clarion call to government and industry that the copying of expressive media for commercial generative AI training purposes is not typically going to be a fair use. Such action would help inform the debate on generative AI, establish clear guidelines, and support the Constitutional goals of copyright law.

- **Transparency:** To enable meaningful enforcement of copyrights, the Office should endorse substantial transparency measures around the ingestion of copyrighted materials for use in generative AI training processes. Such measures enjoy strong support among rightsholders, given their necessity in enforcing legal rights and preventing the misappropriation of protected content from pirate websites.

- **Licensing:** The Office should use its expertise in copyright licensing issues to encourage the further development of relevant licensing models, including by acknowledging the feasibility of voluntary collective licensing to facilitate effective solutions for generative AI developers to license content at scale from both small and large publishers alike. The Office can follow decades of its own policy precedent to reject government-mandated solutions in the absence of demonstrated market failure.

- **Competition:** To facilitate a sustainable AI framework, the Office should acknowledge the intersection of copyright law and competition policy and address the potential for anticompetitive actions of generative AI developers to distort traditional copyright discussions and markets for copyrighted works, as we are already seeing in the context of generative search. Indeed, the gravity of these concerns was highlighted to the Office by the Federal Trade Commission.

- **Enforcement:** The Office and Administration should support the development of effective technical measures that prohibit scraping for generative AI training purposes and enforcement efforts against scraping from third-party websites that engage in systemic and clear infringement of publisher content. N/MA encourages the Office to promote these goals through interagency dialogue with the U.S. Trade Representative (USTR), the Department of Justice (particularly the Computer Crime and Intellectual Property Section), and Intellectual Property Enforcement Coordinator (IPEC) initiatives.

Relatedly, we strongly urge the Office to immediately adopt a solution to enable publishers to efficiently register online news website content with identifying materials. This long-awaited registration option is vital for the ability of publishers to meaningfully enforce their copyrights. The copyright registration system – whatever its IT limitations may be – must not act as an effective obstacle to the ability of publishers to enforce their rights and collect damages authorized by the Copyright Act.

2. The Office Can and Should Prioritize the Important Issues Generative AI Presents to Newspaper, Magazine, and Digital Media Publishing.

The Office's study is critical for journalists and publishers of professional media, who form a key pillar of a healthy and informed democracy. Media publishers perform a vital societal function that is gravely threatened by unauthorized use of their expressive content for generative AI training purposes. Among other benefits, trustworthy media consistently serves as an antidote to risks posed by generative AI – issues that are being tackled by various agencies across the government, including deepfakes, election disinformation, false indications of origin, and social manipulation. This crucial reporting and watchdog role cannot be replaced by AI generations that are susceptible to hallucinations and lack editorial oversight.

To survive and flourish, publishers of all sizes rely on the protections afforded by copyright law, the engine of free expression. As News Corp. noted in its initial comments, "while the Copyright Office wades into the complexities presented by generative artificial intelligence, we encourage it to not lose sight of this simple truth: protecting content creators is one of copyright law's core missions, and doing so is necessary to allow publishers to produce the kinds of news and information that News Corp employees generate every day. The implications for publishers, their readers, and democratic values could not be more profound."[4]

As the Copyright Office's AI study tees up a complex set of issues to consider, the core concerns of media publishers and similarly situated creative industries are clear, urgent, and worthy of the Office's focused attention. Throughout development, generative AI developers have scraped and used massive amounts of publisher materials to fuel their systems and models, and in turn, compete with media content. As evidenced in N/MA's White Paper included in our initial submission, ***popular curated datasets underlying LLMs significantly overweight***

[4] NEWS CORPORATION, Comments of the News Corporation to the Copyright Office at 1 (2023) [hereinafter News Corp. Comments].

publisher content by a factor ranging from over 5 to almost 100 as compared to the generic collection of content that the well-known entity Common Crawl has scraped from the web. A separate investigation came to a similar conclusion, finding that ***half of the top 10 sites scraped for LLM training were news sites***.[5] Additional N/MA member content led other categories, such as "home & garden" and "hobbies & leisure."[6]

This mass-scale and systemic infringement poses a real threat to quality media, as evidenced by the unified, shared concerns expressed by media publishers. Reliable media content is also needed as training material to prevent "model collapse" of the LLM systems themselves, which is caused by models being trained on low-quality, AI-generated content, and to facilitate the shared public goal of safe and widely beneficial generative AI innovations.[7]

The comments received by the Office demonstrate the importance of finding solutions that foster innovation while also protecting publishers who invest considerable time, resources, and creativity in producing new, original content. This balance has always been at the heart of the Constitution's objective of "promot[ing] the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." N/MA reiterates our support for responsible, transparent, and accountable generative AI development, in a manner that compensates publishers fairly for the valuable uses of their content used to power generative AI models and applications.

3. Many Comments Point to Common Understandings or Principles Useful to Evaluating Questions Raised by the Office.

N/MA focuses on five main areas related to use of copyrighted material in generative AI training, transparency, licensing, competition, and enforcement.

a. Generative AI Training and Copyright Infringement Concerns

The Copyright Office should clarify that ingestion of copyrighted media materials to develop commercial generative AI products and services is infringing and typically not a fair use.

i. There Is a Growing Consensus That Generative AI Training Implicates the Reproduction Right.

[5] Kevin Schaul, Szu Yu Chen and Nitasha Tiku, *Inside the Secret List of Websites That Make AI Like ChatGPT Sound Smart*, The Washington Post (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

[6] *Id.* (noting, *e.g.*, that *Food & Wine* was a top "hobbies & leisure" publication and *Jalopnik* was a top "home & garden" publication).

[7] *See, e.g.,* Ilia Shumailov, et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget,* ArXiv (May 27, 2023), available at *https://arxiv.org/abs/2305.17493*; Sina Alemohammad, et al., *Self-Consuming Generative Models Go MAD*, ArXiv (Jul. 4, 2023), available at https://arxiv.org/abs/2307.01850.

Stakeholder comments demonstrate that, with respect to media materials copied by leading LLM developers, there is a prima facie infringement of copyright. There is wide agreement that "training" of LLMs requires making copies of copyrighted works, sometimes multiple times, and that such copying is frequently unauthorized. As noted in N/MA's White Paper, "[p]ublisher content is a major category of expressive information contained in the datasets used to build the LLMs. … news and media content is overrepresented in samples of popular curated sets such as C4, OpenWebText, or OpenWebText2 used for LLM training, as compared to the broader category of material captured in the Common Crawl."[8]

Many publishers recounted evidence of copying of protected material into datasets supporting LLMs. News Corp noted, "[i]n one month in 2018 alone, Common Crawl copied more than 180,000 works belonging to the Chicago Tribune, 180,000 works belonging to the Washington Post, and 230,000 works belonging to The Wall Street Journal."[9] The New York Times noted that "a recreated version of WebText, the dataset used to train OpenAI's ChatGPT-2, shows that a stunning 1.2% of the dataset is The Times's content."[10] Book authors and book publishers raised similar concerns that LLMs incorporated copies of pirated or unlicensed books into their development.[11] Other creative industry organizations provided similar evidence.[12]

Many AI developers confirm this copying. In 2019, OpenAI noted to the U.S. Patent and Trademark Office that training "necessarily involves first making copies of the data to be analyzed."[13] Responses to the Office's Notice were similar, *e.g.*:

[8] News/Media Alliance, White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems is Copyright Infringement and Not a Fair Use at 19-20 (2023), http://www.newsmediaalliance.org/wp-content/uploads/2023/10/AI-White-Paper-with-Technical-Analysis.pdf. [hereinafter N/MA White Paper]

[9] News Corp. Comments at 4.

[10] The New York Times Company, Comments of The New York Times Company at 3 (2023). [hereinafter NYT Comments]

[11] *See* The Authors Guild, Comments of the Authors Guild: Artificial Intelligence and Copyright at 8 (2023) ("[A]n independent AI researcher, Shawn Presser, decided to create something similar to Open AI's Books2 for use by open-source developers; he did it by downloading around 200,0000 books from a pirate torrent tracker…') [hereinafter AG Comments]; The Association of American Publishers, Comments from the Association of American Publishers at 8-9 (2023) ("Several Gen AI systems have included pirated books in their training datasets. For example, a Washington Post analysis of Google's C4 dataset … found that 'b-ok.org, a notorious market for pirated e-books that has since been seized by the U.S. Justice Department,' was among the largest data sources in the dataset.") [hereinafter AAP Comments].

[12] *See* National Association of Broadcasters, Comments of the National Association of Broadcasters at 4 (2023); National Music Publishers' Association, National Music Publishers' Association Comments in Response to the Notice of Inquiry at 15 (2023) [hereinafter NMPA Comments]; American Association of Independent Music and Recording Industry Association of America, Comments of the American Association of Independent Music and Recording Industry Association of America, Inc. at 13-14 (2023) [hereinafter A2IM/RIAA Comments]; Getty Images, Response to USCO Inquiry on Artificial Intelligence and Copyright, Appendix A at 14 [hereinafter Getty Comments].

[13] OpenAI, Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation at 2 (2019), https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf.

- **Google:** “If training could be accomplished without the creation of copies, there would be no copyright questions here.”[14]

- **BSA:** “In addition to the reproductions that may be needed to create an AI training database, reproductions may also be made when the training data undergoes the computational analysis that occurs during the machine learning process.”[15]

- **A16z:** “Each of these technologies involves the wholesale copying of one or many copyrighted works.”[16]

- **CCIA:** “Assembling . . . data may entail converting it into a more usable format, e.g., translating image files into mathematical image representations. In addition, backup copies of the materials may be necessary to protect against loss of data in the event of system failure. Temporary reproductions of portions of the material in a computer’s random access memory are a normal part of any computer program, including the process of training an AI algorithm.”[17]

- **Anthropic:** "For Claude, as discussed above, the training process makes copies of information for the purposes of performing a statistical analysis of the data.”[18]

ii. LLM Copying is Neither Transitory Nor Limited to Unprotectable Facts, and Evidence Suggests That the Models Themselves Can Sufficiently Retain and Embody the Expressive Works They Were Trained on.

At least with respect to typical uses of textual material for LLM training, the Office should begin its analysis from the determination that such unauthorized copying constitutes prima facie infringement.[19] While some contend that copies are merely ephemeral,[20] or limited to unprotected elements,[21] these allegations do not jibe with either copyright law or the facts at hand. (Others allege that copying is excused as “intermediate” or for a “non-expressive use,” addressed in the discussion of fair use below).

[14] GOOGLE, Comments of Google LLC at 9 (2023) [hereinafter Google Comments].

[15] BSA - THE SOFTWARE ALLIANCE, Comments from BSA at 8 (2023) [hereinafter BSA Comments].

[16] ANDREESSEN HOROWITZ, Comments of a16z at 7 (2023) [hereinafter a16z Comments].

[17] COMPUTER & COMMUNICATIONS INDUSTRY ASSOCIATION, Comments of the Computer & Communications Industry Association (CCIA) at 8 (2023) [hereinafter CCIA Comments].

[18] ANTHROPIC, Public Comments of Anthropic PBC at 7 (2023) [hereinafter Anthropic Comments].

[19] As noted in N/MA’s initial comment, there are also many examples of training using permissively accessed content, demonstrating the feasibility – and desirability – of driving AI innovations to use licensed content.

[20] *See, e.g.,* CCIA Comments at 8; CREATIVE COMMONS, Response from Creative Commons at 2 (2023).

[21] *See, e.g.*, Anthropic Comments at 7; OPENAI, Comments of OpenAI at 12 (2023) [hereinafter OpenAI Comments]; STABILITY AI, Response to United States Copyright Office Inquiry into Artificial Intelligence and Copyright at 13 (2023).

First, the Office should quickly dismiss suggestions that the copying of digital works is so transitory in nature as to not meet the standard for fixation. While most developer comments did not seriously challenge the fixation requirement, others advanced this argument in passing, without factual support.[22] As N/MA's initial response to question 7 explained, LLM training involves the systematic and often repeated copying and storage of expressive works into datasets, including compiling, cleaning, development, and fine-tuning.[23]

Such activities constitute actionable copying under law, and they raise no metaphysical questions around whether certain data packet transfers create a "copy." Generative AI datasets are downloaded, stored, and cleaned, sometimes involving manual review by low wage workers to scrub out illegal, harmful, or disturbing content.[24] For example, Common Crawl explains that its "crawl data is stored on Amazon's S3 service, allowing it to be bulk downloaded as well as directly accessed" and instructs users on how they can "download the files entirely free using HTTP(S) or S3."[25] Datasets are often retained for fine-tuning.[26] And copying can also occur at the output stage, demonstrating that the models retain, in some fashion, a version of the expressive material they have ingested.

Others similarly dismiss suggestions that AI copying is ephemeral.[27] Professors Samuelson, Sprigman, and Sag explain, as "we understand that is considered broadly impractical to proceed without creating a semi-permanent local copy of the training data."[28] A2IM/RIAA put it well:

> That is a purely academic argument. In practice, persistent – and therefore actionable – copies of copyrighted material are made throughout the training process: first, in compiling and cleaning the dataset, and then in the model development and fine-tuning. The development and fine-tuning process is an iterative one, so it is often necessary to keep copies of the dataset on hand throughout each iteration.[29]

[22] *See* ENGINE, Re: Comments of Engine to the U.S. Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright, Docket No. 2023-6 at 5 (2023); CONSUMER TECHNOLOGY ASSOCIATION, Comments of Gary Shapiro at 4 (2023); BSA Comments at 8. As BSA notes, this exception, if applicable, could likely relate only to certain RAM copies employed in AI training.

[23] NEWS/MEDIA ALLIANCE, Comments of the News/Media Alliance at 26-27 (2023) [hereinafter N/MA Comments].

[24] *AI Annotation & Data Labeling Services Ind.*, ISHIR (n.d.), https://www.ishir.com/ai-annotation-services-india.htm (last visited Oct. 25, 2023). Niamh Rowe, Underage Workers Are Training AI, Wired (Nov. 15, 2023), https://www.wired.com/story/artificial-intelligence-data-labeling-children/.

[25] *Frequently Asked Questions*, COMMON CRAWL (n.d.), https://commoncrawl.org/big-picture/frequently-asked-questions/ (last visited Oct. 25,2023); *Get Started*, Common Crawl (n.d.), https://commoncrawl.org/the-data/get-started/(last visited Oct. 25,2023).

[26] Van Lindberg, *Building and Using Generative Models Under US Copyright Law*, 18 RUTGERS BUS. LAW 1, 6 (2023) ("In many cases, the same inputs are re-used in different rounds of training.").

[27] COPYRIGHT ALLIANCE, Comments of the Copyright Alliance at 37, 4 (2023) [hereinafter CA Comments].

[28] Pamela Samuelson et al., Comments in Response to the Copyright Office's Notice of Inquiry on Artificial Intelligence and Copyright at 19 (2023).

[29] A2IM and RIAA Comments at 14-15.

And as Universal Music Group notes, digital copyright litigation history shows that some infringers use these arguments to drag out litigation and drive up enforcement costs.[30] The Office should not countenance serious discussion of this defense as a result.

Similarly, the Office should give little weight to suggestions that the reproduction right is not implicated because only facts or other unprotected elements are copied. As explained in N/MA's initial comments, ingestion and processing for generative AI uses targets media content for its expressive properties, not for its uncopyrightable elements. Others, such as Professor Daniel Gervais, came to similar conclusions: "Yet in the case of GenAI, the use by the machine is not mere character recognition; ***it is semantic in nature***. The machines process the expression of ideas in the works to create new expression."[31] A forthcoming paper from Professor Matthew Sag notes, "[a]lthough there is no machine learning exception to the principle of non-expressive use, the largeness of likelihood models suggest that ***they are capable of memorizing and reconstituting works in the training data,*** something that is ***incompatible with non-expressive use.***"[32]

Copyright Clearance Center (CCC) provides a technical explanation of how this occurs:

> Once the AI system has mapped the input text into tokens, it encodes the tokens into numbers and converts the sequences (even up to multiple paragraphs) that it processed as vectors of numbers that we call "word embeddings." These are vector-space representations of the tokens that ***preserve their original natural language representation that was given as text.*** It is important to understand the role of word embeddings when it comes to copyright because the embeddings are the ***representations (or encodings) of entire sentences, paragraphs, and even documents,*** in a high-dimensional vector space. It is through the embeddings that the AI system captures and stores the meaning and the relationships of the words from the natural language.[33]

CCC's description is not so different from how Microsoft puts it:

[30] UNIVERSAL MUSIC GROUP, Comments of Universal Music Group at 20-21 (2023) [hereinafter UMG Comments] ("Generative AI is not a magical medium and would not be able to replicate a work in its output if it had not copied and retained a reproduction of that work in some digital form in the first instance.").

[31] Daniel Gervais, Comment Submitted by Professor Daniel Gervais, Vanderbilt University at 3-4 (2023) [hereinafter Gervais Comments]. *See also* Authors Guild Comment at 18 ("[T]he works' expressive elements are what is needed for the companies to create a more commercially desirable product—one that can generate outputs that compete with the very works used to build the system."). Unless otherwise indicated, all emphasis in quotations has been added.

[32] Matthew Sag, *Copyright Safety for Generative AI*, 61 HOUSTON L. REV. 2 (2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4438593.

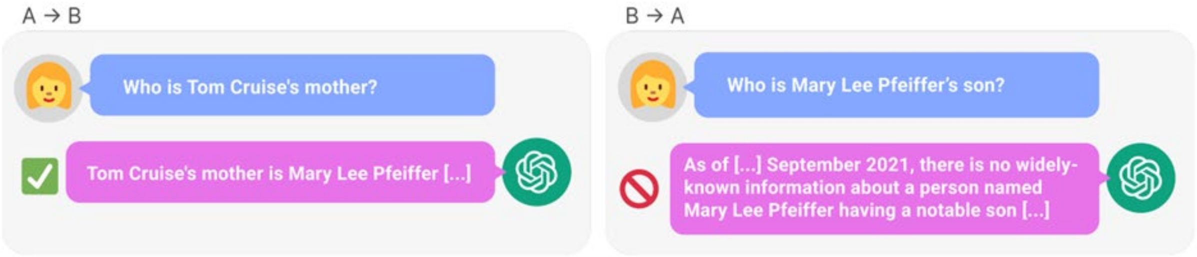
[33] COPYRIGHT CLEARANCE CENTER, Comments of the Copyright Clearance Center at 6 (2023) [hereinafter CCC Comments].

> [W]ords are transformed into 'tokens' that are represented as numerical vectors. These vectors are generated to represent not just words but information about the semantic and contextual meaning of the words and their relationships to other words in the vocabulary. This enables the model to correlate relationships between words.[34]

Using tokenization to map the precise semantic and expressive structure of creative works may be a novel form of memorization, but make no mistake – inherently this is a system that copies and retains the linguistic choices made by creative professionals, not one that "extracts facts" about natural or public domain phenomena or language systems reserved for the public commons.[35] As our White Paper explains, this is why many academics have deemed LLMs to be "stochastic parrots", mimicking syntactical choices made by human writers, but without comprehending or extracting the meaning or facts in a news article or other works.

The "reversal curse" is a vivid example of the shallowness of this parroting:

> "If a model is trained on a sentence of the form '*A* is *B*,'" [a recent research paper found], "it will not automatically generalize to the reverse direction '*B* is *A*.'"[1] In fact, a model that the researchers trained only on facts recited in one direction completely failed to generate equivalent descriptions in reverse. They also found this defect to be evident in the large commercial models that are in use today. For example, GPT-4 is perfectly able to say who Tom Cruise's mother is (Mary Lee Pfeiffer) but it can't answer the reverse question of who is Mary Lee Pfeiffer's son.[36]



Who is Tom Cruise's mother?

✓ Tom Cruise's mother is Mary Lee Pfeiffer [...]

Who is Mary Lee Pfeiffer's son?

As of [...] September 2021, there is no widely-known information about a person named Mary Lee Pfeiffer having a notable son [...]

This memorization appears to be so persistent that generative AI models can sufficiently embody the expressive works they were trained on. This docket includes numerous examples of verbatim, near-identical, or substantially similar outputs. The Technical Annex to N/MA's White Paper includes multiple examples of models reproducing content of news articles,

[34] Microsoft, Comments of Microsoft and Github at 6 (2023).

[35] *Compare Capitol Records, LLC. V. BlueBeat, Inc.,* 765 F. Supp. 2d 1198 (C.D. Cal. 2010) ("[Defendants'] obscure and undefined pseudo-scientific language appears to be a long-winded way of describing 'sampling,' i.e. copying, and fails to provide any concrete evidence of independent creation.").

[36] N/MA White Paper at 12. ("What large language models do *not* do is 'learn' facts or derive 'rules' of language from the large amounts of expression used to train them"); *quoting* Lukas Berglund et al., *The Reversal Curse: LLMs Trained on "A Is B" Fail to Learn "B Is A"* (Sept. 22, 2023), https://doi.org/10.48550/arXiv.2309.12288.

ranging from Pulitzer-winning articles to evergreen reference material carefully drafted, fact-checked and updated to reflect changing developments.[37] The Copyright Alliance noted, "[t]here are several examples of AI models being prompted to reproduce almost verbatim text from ingested books, song lyrics or reproducing ingested pictures, further supporting the notion that these works are embedded in the model itself, to varying degrees."[38] Others encountered the same phenomena,[39] and recent litigation filings document similar replications.[40]

While it is obvious that generative AI models can and do create infringing outputs, these prevalent examples also document a separate issue in these models' ability to retain copyrighted material they have ingested. Whether or not LLMs mitigate after the fact by employing rules that preclude LLMs from providing these outputs, their capability to do so belies claims that expressive material is not memorized by a computer.[41] This retention can give rise to continued harm and actionable conduct beyond initial ingestion activities.[42]

While some generative AI developers call memorization and repetition of expressive works "a bug to be corrected, rather than a feature to be pursued,"[43] it is clear LLMs are infested and overrun by these so-called "bugs." Just as saying "no copyright infringement intended" does not remedy infringements on UGC platforms, wishful and self-serving statements about "overfit" do not excuse infringements by LLM models, and do not reflect the technology actually deployed into consumer markets. A recent academic release noted "by querying the model, we can actually extract some of the exact data it was trained on."[44] In this study, researchers from Google DeepMind, the University of Washington, Cornell, Carnegie Mellon University, the University of California Berkeley, and ETH Zurich were able to bypass "alignment" rules to extract verbatim training data from open source, semi-open, and closed models. As *404* reported:

> A team of researchers primarily from Google's DeepMind systematically convinced ChatGPT to reveal snippets of the data it was trained on using a new type of attack prompt which asked a production model of the chatbot to repeat specific words forever.

[37] N/MA White Paper, Technical Appendix at 23-30 (2023).

[38] CA Comments at 38.

[39] NMPA Comments at 11-12; European Writers' Council, Comments of the European Writers' Council at 11 (2023) [hereinafter EWC Comments]; AAP Comments at 20.

[40] See, e.g., Complaint at 12, *Authors Guild v. OpenAI*, 1:23-cv-08292 (S.D.N.Y.); Complaint, *Getty v. Stability AI*, 1:23-cv-00135-UNA (D. Del.); Complaint at 20-39, *Concord Music Group v. Anthropic*, 3:23-cv-01092 (M.D. Tenn.).

[41] While N/MA's comment does not analyze effects on the exclusive derivative right of copyright owners, others suggest it may also be implicated. *See, e.g.,* AAP Comments at 11.

[42] *Compare Cartoon Network LP, LLLP v. CSC Holdings, Inc.,* 536 F.3d 121, 127 (2d Cir. 2008), cert. denied, 557 U.S. 946 (2009) (finding "embodiment" requirement met when a work is placed in a medium from where it *can be* reproduced).

[43] OpenAI Comments at 7.

[44] Milad Nasr et al., Extracting Training Data from ChatGPT, arXiv:2311.17035 (2023), https://arxiv.org/abs/2311.17035.

> . . . Using this tactic, the researchers showed that there are large amounts of privately identifiable information (PII) in OpenAI's large language models. They also showed that, on a public version of ChatGPT, ***the chatbot spit out large passages of text scraped verbatim from other places on the internet***. . . . It also, crucially, shows that ***ChatGPT's "alignment techniques do not eliminate memorization," meaning that it sometimes spits out training data verbatim***. . . . Some of the specific content published by these researchers is scraped directly from CNN, Goodreads, WordPress blogs, on fandom wikis, and which contain verbatim passages from Terms of Service agreements, Stack Overflow source code, copyrighted legal disclaimers, Wikipedia pages, a casino wholesaling website, news blogs, and random internet comments.[45]

While the paper's researchers focus on the significant privacy risks created if personally identifiable information can be disclosed once guardrails are bypassed, from a copyright perspective it is not enough to simply expect LLMs to cover their tracks better. If a model makes unauthorized use of copyrighted content in a manner that usurps the markets for licensing that content, including by competing with it directly, that ongoing use is concerning, even if the model is instructed not to provide verbatim copies as outputs.

Considering this extensive memorialization and retention of expressive works, as well as the ability for generative AI models to create derivative works, statements such as "there is no copy of the training data — whether text, images, or other formats — present in the model itself"[46] and "[d]espite a common and unfortunate misperception of the technology, the models do not store copies of the information that they learn from"[47] are misleading or, at best, wishful thinking. Regardless of the exact technical processes employed,[48] the models function in a manner that has the same effect as memorization and retention.

Finally, we express concerns over, and ask the Office's infringement analysis to consider the additional process of the unauthorized use of media content in retrieval augmented generation (RAG), also known as "grounding", where LLMs seek out and copy new, current, content (over and above what they were "trained" on) in response to direct queries, including news content, to ensure that the outputs of LLMs remain up to date.[49] RAG processes, which may help correct

[45] Jason Koebler, Google Researchers' Attack Prompts ChatGPT to Reveal Its Training Data, 404 (Nov. 29, 2023), https://www.404media.co/google-researchers-attack-convinces-chatgpt-to-reveal-its-training-data/.

[46] Google Comments at 3-4.

[47] OpenAI Comments at 6.

[48] By analogy, modern digital storage methods are far more distributed and varied than files stored on a hard drive, yet copyright law still applies, such as the clarification that the section 115 license applies to music streaming or video streaming's employment of segmented caching. *See* https://ieeexplore.ieee.org/document/7570510.

[49]*See, e.g., Alan Zeichick, What Is Retrieval-Augmented Generation (RAG)?, Oracle Blog (Sept. 19, 2023),* https://www.oracle.com/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/; Eleanor Berger, Grounding LLMs, Microsoft FastTrack (Jun. 10, 2023), https://techcommunity.microsoft.com/t5/fasttrack-for-azure/grounding-llms/ba-p/3843857.

hallucination or misinformation problems, do so through fresh accessing of reliable copyrighted content, including news and other media content. As a recent github LLaMA project for a "framework based on news contents" describes it, "***the retrieved [news] documents serve as a foundation for generating comprehensive and contextual accurate answers:"***[50] Such techniques can directly compete with news reporting, compounding infringement concerns.

iii. The Fair Use Doctrine Does Not Excuse the Taking of Expressive News Content by Commercial LLM Chatbots

Where protected content has been copied without authorization, the only question that remains for the Copyright Office to consider is whether such copying is excused by a relevant exception or limitation—here, fair use. At least with respect to LLMs copying media content, N/MA's initial comment and accompanying White Paper provides a fair use analysis that explains the answer is generally "no." On reply, we briefly connect that analysis to some of the more misguided positions raised by commenters.[51]

First, some suggest that the so-called "intermediate copying" doctrine excuses the systematic, unauthorized reproduction of copyrighted works.[52] This is a legal misconception. As explained further in N/MA's initial comment,**[53]** intermediate copying is for the purpose of reverse engineering to understand how a computer program works; this doctrine has never been applied to copying of expressive works for the purpose of exploiting the expressive content of the works. Reverse engineering is a fundamentally different use of a copyrighted work: it is for the purpose of understanding ***how*** it works, not to ***absorb*** the work. Broadening the holdings of these cases to encompass the wider range of unlicensed activity involved in "training" LLMs would be destructive to news reporting, media archive access, reprography, media monitoring, genealogy, research uses, educational publishing, and many other established markets, as well as the creative activity they underpin.

Like contentions that LLMs are "reading," some argue that the use of publisher content is fair because the content is only used for "learning" or "knowledge."[54] N/MA's initial comment addressed deficiencies in this argument from factual and legal perspectives.[55] And generative AI development uses material much differently than a person in a library reading a book, or accessing a news article from a reading medium intended to be supported by ads reaching

[50] Ransaka Ravihara, daily-llama GitHub repository (2023), https://github.com/Ransaka/daily-llama.

[51] As it was not feasible to humanly review all 10,000 comments by the reply deadline, we look forward to engaging further as appropriate, including to address overlooked positions.

[52] *See, e.g.,* META, Comments of Meta Platforms, Inc. at 13 (2023) [hereinafter Meta Comments]; Anthropic Comments at 2; BSA Comments at 2-3 ("[T]he reproductions are 'intermediate' in the sense that they are not visible or otherwise made available to the public.").

[53] N/MA Comments at 34.

[54] *See, e.g.,* Google Comments at 9; OpenAI Comments at 6.

[55] N/MA Comment at 40-41; N/MA White Paper at 8-13.

eyeballs. The use is more akin to copying the contents of an entire Barnes & Noble store and taking that material out the door to "use" it repeatedly and commercially over years, to great economic enrichment. In any event, large companies like Google, Meta, or Microsoft who want employees to "learn" by reading the *San Francisco Chronicle*, *The New York Times,* or industry newsletters typically buy subscriptions or otherwise lawfully obtain access.

As a result, we are concerned by overbroad, erroneous statements, such as a16z's proclamation that "[w]here copies of copyrighted works are created for use in the development of a productive technology with non-infringing outputs, our copyright law has long endorsed and enabled those productive uses through the fair use doctrine."[56] Case law has generally not permitted copying for purposes that do not comment on or at least *point to* the original works, outside of defined, limited exceptions, such as to access functional computer code for interoperability purposes. Extending this reasoning to authorize wholesale copying of massive libraries to create expressive works—and even substitutional expressive works—is an overly simplistic perversion of existing doctrine that goes far beyond the traditional contours of fair use.[57]

N/MA's initial response to question 8 explained our position in greater detail, analyzing cases and noting that cases holding "fair" the use of copyrighted materials to develop a new technology or further a technological purpose are grounded on findings that the ultimate use *did not* compete with the copyrighted works.[58]

While other commenters pointed to *HathiTrust* to support a fair use argument, this case, too, does not go so far. Under factor one, the justification is greater when a specific work is necessary for a secondary purpose, or the use is "targeted."[59] The purpose of the *HathiTrust* database—to point a user to information about a work while not competing with the work—is entirely different from the purpose of models that generate images, text, or music that may compete with the ingested works. The purpose of non-profit libraries providing full text access to patrons with certified print disabilities is also quite different, that is, to provide a patron with

[56] A16z Comments at 7.

[57] Indeed, many developer comments give short shrift to the Supreme Court's recent decision in *Goldsmith v. Andy Warhol Foundation,* and even misstate the test for fair use. *See, e.g.,* Google Comments at 2.

[58] N/MA Comments at 42-43 (addressing *Authors Guild v. Google*, 804 F.3d 202, 218 (2d Cir. 2015) and *Kelly v. Arriba Soft Corp*., 336 F.3d 811, 821 (9th Cir. 2002); citing *Fox News Network, LLC v. TV Eyes, Inc*., 883 F.3d 169, 177, 181 (2d Cir. 2018) (media monitoring service, while transformative, was not fair, because it usurped plaintiff's market); *Hachette Book Grp., Inc. v. Internet Archive*, No. 20-CV-4160 (JGK), 2023 WL 2623787, *18-25 (S.D.N.Y. Mar. 24, 2023) (Internet Archive's electronic copying and unauthorized lending of 3.6 million books protected by valid copyrights is not a fair use because it competed with plaintiff's licensing market); *Associated Press v. Meltwater U.S. Holdings, Inc*., 931 F. Supp. 2d 537, 561 (S.D.N.Y. 2013) (crawling and scraping "snippets" of news stories for use in notifying and informing Meltwater's customers directly competed with the Associated Press such that Meltwater's copying would deprive the Associated Press of a stream of income to which it was entitled).

[59] *See Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258, 1275-76 (2023) (quoting *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 580-81 (1994).)

a specific book to consume which would be otherwise unavailable to them, representing a well-defined universe of non-commercial and publicly beneficial uses. With generative AI, copying takes desirable materials for purposes untethered to the specific works, and uses them for massively commercial and substitutional uses in broad consumer markets. The justification for copying a given work is accordingly weak. And the long-underdeveloped market for assistive formats[60] is different from the hot commercial markets for generative AI ingestion and substitutional outputs.

Unlike *HathiTrust,* there is also no legislative history or settled policymaker recognition that generative AI ingestion serves the public good, as there was for copying for the benefit of visually impaired or print disabled persons.[61] Instead, Congress is presently devoting considerable time to assessing risks associated with AI, and prominent AI researchers have called for a 6-month pause to reduce the risk of human extinction.[62]

We are not alone in concerns that some developers appear to have staked bets on flimsy or under-researched understandings of copyright law.[63] The Office received thousands of comments raising reasoned concerns about the effect that these parasitic uses will have on creative cultural production. Just last month Stanford's Human Centered Artificial Intelligence Institute concluded: "[o]ur review of U.S. fair use doctrine concludes that fair use is not guaranteed for foundation models as they can generate content that is not 'transformative' enough compared to the copyrighted material."[64]

In light of all of the above, the Copyright Office should deliver direct and straightforward guidance to government and industry that ingestion of copyright protected media content to develop commercial generative AI products and services is infringing and typically not a fair use.

[60] *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014) ("[I]t is common practice in the publishing industry for authors to forego royalties that are generated through the sale of books manufactured in specialized formats for the blind…").

[61] 755 F.3d 87 (2d Cir. 2014) (noting "making a copy of a copyright works for the convenience of a blind person is expressly identified by the House Committee Report as an example of fair use...").

[62] The Associated Press, *Tech Leaders Urge a Pause in the 'Out-of-Control' Artificial Intelligence Race*, NPR (Mar. 29, 2023), https://www.npr.org/2023/03/29/1166896809/tech-leaders-urge-a-pause-in-the-out-of-control-artificial-intelligence-race.

[63] Michael Hiltzik, *AI Investors Say They'll Go Broke if They Have to Pay for Copyrighted Works. Don't Believe It*, The Los Angeles Times (Nov. 16, 2023), https://www.latimes.com/business/story/2023-11-16/ai-investors-say-theyll-go-broke-if-they-have-to-pay-for-copyrighted-materials-dont-buy-it ("It boils down to the claim that even if the entire AI industry happens to be wrong about the application of copyright law, its investors have staked so much on an erroneous legal interpretation that we should just give them a pass.").

[64] Peter Henderson et al., *Foundation Models and Copyright Questions*, HAI Policy & Society Policy Brief (Nov. 2023), https://hai.stanford.edu/sites/default/files/2023-11/Foundation-Models-Copyright.pdf. *See also* Gervais Comments ("Google Books is on point, but only to a certain extent…. In the case of LLMs, the output is different: new literary and artistic content… Whether that type of use, namely to create content that may compete with the material it was trained on, can be considered fair under Warhol, (of course the case was limited to an analysis under the first factor) is an open question.").

With multiple litigations percolating at the district level, and, to our knowledge, no court having yet ruled on the applicability of the fair use defense to the ingestion of materials into AI models, the Office's study can provide valuable guidance from the expert agency on important questions of copyright, a consideration tacitly recognized by President Biden's executive order on AI.[65]

b. Primary and Secondary Liability Analyses Should Incentivize Technology Companies Towards Responsible Design.

Over past decades, some internet technology companies have benefited from special exceptions to normal rules of corporate responsibility, essentially to encourage their development of a networked infrastructure that would serve others. These companies are now mature and have understandably made use of an advantageous statutory and regulatory playing field to grow their businesses (and supporting investment infrastructures) to incredible heights.

Copyright issues around generative AI development are not the same as the questions surrounding access to third party content at play in the 1990s. As the Copyright Office's Section 512 Study recognized, the conditions that gave rise to the DMCA are not today's conditions.[66] While some commenters have never known it any other way, from a historical perspective, it is highly abnormal for large and influential chunks of industry to be exempted from responsibility or regulatory oversight. While his comments focused on Section 230 of the Communications Decency Act, as opposed to copyright, the Office may be interested in an editorial by a former executive director of Harvard's Berkman Center, titled "*Underregulating tech is a relic of the 90s. AI is an urgent call for change*."[67]

News media publishers also operate platforms that distribute content to be widely accessed by the public. From the Connecticut Courant reprinting Thomas Paine's *Common Sense* in 1775 to cutting-edge digital coverage of generative AI by N/MA members today, we know it is possible to build innovative, quality information distribution businesses in a responsible manner. It will not "break" generative AI to expect its developers and operators to do the same.

[65] Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, Oct. 30, 2023, at 5.2(c)(iii) (noting the study shall address "the treatment of copyrighted works in AI training.").

[66] U.S. COPYRIGHT OFFICE, Section 512 of Title 17: A Report of the Register of Copyrights at 27-34 (2020), https://www.copyright.gov/policy/section512/section-512-full-report.pdf.

[67] John Palfrey, *Underregulating Rech Is a Relic of the 90s. AI Is an Urgent Call for Change, The Hill* (Nov. 12, 2023), https://thehill.com/opinion/technology/4304297-underregulating-tech-is-a-relic-of-the-90s-ai-is-an-alarm-for-urgent-change/ ("Policymakers in the U.S. have historically eschewed regulating the tech sector for fear of stifling innovation and American competitiveness in the global marketplace. But that argument is a relic of the late 90s when cyberspace was new and its impact unknown.").

i. "*Sony-Betamax*" Does Not Excuse AI Developers From Contributory Liability Risks.

With respect to contributory liability of a generative AI developer or system for infringements caused by user prompts, multiple commenters contend that only the user should be liable, based on the claim that the system is "capable of substantial non-infringing uses" or because of the "*Sony* safe harbor."[68] Others compare generative AI software to physical tools like word processors.[69] Explicitly or implicitly, these arguments invoke *Sony v. Universal,* but gloss over important distinctions between the physical device at-issue in *Sony* (the Betamax recorder) and generative AI software development. In *Sony,* the Court considered whether the Betamax manufacturer was contributorily liable because consumers could copy copyrighted television programming and held that it was not. In reaching this opinion, the Court stressed that the manufacturer had no ongoing relationship with its customers after the sale of the device.[70]

Generative AI development raises quite different considerations than *Sony*, which concerned a physical device sold to consumers. LLMs, for example, are trained on copyrighted material that in turn shapes the output; it is the AI developer that is assembling or curating the training content, and the LLM (or instantiation) that is typically primarily directing the output. In *Sony,* the end-user completely directed the output, with no input from the Betamax machine. But the influence of AI models is precisely why the Copyright Office has issued guidance addressing registrability of material where the only asserted human authorship is via model prompts. This scenario is more similar to preloaded set-top boxes, which were found to be infringing by facilitating access to unauthorized material.[71] (It is obviously also very different from use of a word processor, where the user is the only creator of the infringing material).

Generative AI software further differs from physical devices because developers often have an ongoing relationship to users. LLM developers often control what outputs can be requested and have shown a willingness to update their services, host apps, provide fine-tuning services, implement safeguards, collect subscription monies, provide customer service, and many other actions that show an ongoing and direct interest in the activity of users. Thus, the right and ability to control and supervise also makes *Sony* largely inapplicable.[72] Subsequent decisions, including *Napster* and *ReDigi,* found the right and ability to control users to be significant,

[68] *See, e.g.,* CCIA Comments at 9 & 11; ELECTRONIC FRONTIER FOUNDATION, Comments of Electronic Frontier Foundation at 5 (2023); Google Comments at 14.

[69] TECHNET, Comments of TechNet at 6 (2023).

[70] *See* 464 US 417 at 437-438. The opinion also considered the record that consumers' end-use was typically for "time-shifting," which it found was a fair use, and thus did not raise contributory liability concerns for Sony.

[71] *See Universal City Studios Productions v. TickBox TV, LLC*, No. CV 17-7496-MWF (ASX), 2018 WL 1568698, at *10 (C.D. Cal. Jan. 30, 2018).

[72] 464 US 417 at 437.

regardless of whether the service was capable of significant non-infringing uses.[73] A number of peer-to-peer file sharing services (not outwardly encouraging of piracy) were subsequently shut down, or chose to shutter for similar reasons.[74]

Also of note, the Supreme Court clarified in *Grokster* that "*Sony* did not displace other theories of secondary liability,"[75] and found against *Grokster* on an intentional inducement theory; similarly, vicarious liability remains an alternate basis for potential developer liability.[76] The Office may also consider evaluating the various secondary liability questions raised by generative AI at a later stage, given the emergent and black-box nature of generative AI technologies.

ii. The Office Should Encourage Responsible Generative AI Development and Resist Calls to Pass Responsibility on to Individual Users or Others.

Too many comments exhibit disregard for the responsibility that comes with the development and commercialization of powerful and impactful technologies. Not only do some developers want to avoid licensing requirements, they demonstrate considerable hesitancy in accepting liability when it comes to infringing outputs generated by their models. For example, CCIA noted that "[g]enerally, any liability should lie on the end-user who requests and publishes a copyright-infringing work",[77] with Google echoing, "[w]hen an AI system is prompted by a user to produce an infringing output, any resulting liability should attach to the user as the party whose volitional conduct proximately caused the infringement."[78] These positions would help create a liability bubble around generative AI applications and developers, shielding them from the consequences of their actions and decisions, similar to the immunities provided by Section 230 of the Communications Decency Act or Section 512 of the DMCA. It would also greatly expand the burdens on rightsholders and the general public to sort through liability issues, minimizing the likelihood of preventing and obtaining redress for infringement, while alleviating compliance costs on the developers whose technology is at the heart of these disputes.

[73] *A M Records, Inc. v. Napster, Inc.*, 239 F.3d 1004, 1022 (9th Cir. 2001) ("The ability to block infringers' access to a particular environment for any reason whatsoever is evidence of the right and ability to supervise."); *Capitol Records, LLC v. ReDigi Inc., 934 F. Supp. 2d 640, 660 (S.D.N.Y. 2013)* ("Clearly, ReDigi Vicariously infringed Capitol's copyrights. As discussed, ReDigi exercised complete control over its website's content, user access, and sales.").

[74] Alex Bracetti, *A History of P2P Sites Being Shut Down*, Complex (Jan. 28, 2012), https://www.complex.com/pop-culture/a/alex-bracetti/a-history-of-p2p-sites-being-shut-down.

[75] *Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd.*, 545 U.S. 913, at 934 (2005).

[76] *Id.* at 930 ("One infringes contributorily by intentionally inducing or encouraging direct infringement, see *Gershwin Pub. Corp.* v. *Columbia Artists Management, Inc.,* 443 F. 2d 1159, 1162 (CA2 1971), and infringes vicariously by profiting from direct infringement while declining to exercise a right to stop or limit it, *Shapiro, Bernstein & Co.* v. *H. L. Green Co.,* 316 F. 2d 304, 307 (CA2 1963)").

[77] *See* CCIA Comments at 9 & 21.

[78] Google Comments at 12-13.

N/MA strongly urges the Office to reject such arguments and not facilitate attempts to avoid liability when it comes to copyright infringement. Providing broad immunity for generative AI developers would be contrary to baseline legal frameworks of corporate responsibility, not only in copyright law but in other areas of law as well—including consumer protection law, product liability, torts, and environmental law, to name a few.

The immunities provided by Sections 512 and 230[79] both relate narrowly to third-party content hosted on online platforms (in the case of Section 512, after having been placed there by a user) and should not be applied to generative AI models and applications.[80] Unlike internet service providers in their infancy, designed to host user-uploaded content and allow users to communicate and interact with each other, generative AI developers have control over the systems they design, the services they offer, and the software they update. While users can input a prompt, outputs are highly dependent upon the system itself and what it contains or has "memorized."

Generative AI models can produce an infringing work even in response to a completely innocent user query or prompt without any intention to generate infringing output. Users are also unable or unlikely to know what materials power the generative AI model, and whether the content was appropriately licensed, creating a massive information imbalance between the user and the AI developer. It would therefore be wrong to overlook the necessary contributions of the software in generating an infringing output—the law should encourage design that is responsible and safe, rather than reckless at the expense of users.

Generative AI developers and deployers are best positioned to bear the regulatory burden of infringing outputs. CCIA represents the largest companies in modern history. Leading generative AI developers account for over 50 percent of the top-10 S&P 500 stocks, with unprecedented amounts of capital fueling other start-up endeavors.[81] It does not make sense to create or extend a special liability exception to the largest and most well-funded companies in the world at the expense of creators, users and other sectors of the economy. Rather,

[79] Section 230 exempts intellectual property law from its scope and applies to "interactive computer services."

[80] There is considerable uncertainty as to whether Section 230 applies to generative AI applications with some AI industry groups advocating for clarifying the law to include AI applications. See Ashley Johnson, *Generative AI Is the Next Challenge for Section 230*, ITIF Innovation Files (Apr. 12, 2023), https://itif.org/publications/2023/04/12/generative-ai-is-the-next-challenge-for-section-230/ and Peter Henderson, *Law, Policy, & AI Update: Does Section 230 Cover Generative AI?*, Stanford University Law, Regulation, and Law Blog (Mar. 23, 2023), https://hai.stanford.edu/news/law-policy-ai-update-does-section-230-cover-generative-ai.

[81] *See, e.g.*, Gabe Alpert, *Top 10 S&P 500 Stocks by Index Weight*, Investopedia (Aug. 29, 2023), https://www.investopedia.com/top-10-s-and-p-500-stocks-by-index-weight-4843111 (for a list of the top 10 S&P 500 stocks); Cindy Gordon, *AI Start-Up Investments Bucking Venture Capital Decline Trends*, Forbes (Aug. 31, 2023), https://www.forbes.com/sites/cindygordon/2023/08/31/ai-start-up-investments-bucking-venture-capital-decline-trends.

developers should be encouraged to allocate risks from user activities through terms of service and other built-in safeguards.

c. Transparency Requirements Enjoy Broad Support and Should Be Enacted.

Transparency measures, discussed in N/MA's initial comments, are critical for copyright owners to understand whether and how their works are being used in generative AI systems, to negotiate permissions and licenses when required, and to enforce their rights when necessary. In addition to benefiting rightsholders, transparency is needed to incentivize the responsible development of generative AI models, prevent uncertainty and risk from forming "clouds" over exciting innovations and new systems, and to resolve questions over secondary liability.

A wide range of commenters across industries support the adoption of adequate transparency measures to ensure that copyright owners can efficiently and accurately identify the use of their content in training datasets, including: Copyright Alliance,[82] NMPA,[83] A2IM and RIAA,[84] Authors Guild,[85] The New York Times,[86] AAP,[87] AFL-CIO,[88] Getty,[89] WGA East & West,[90] European Writers' Council,[91] and CEDRO.[92] Such requirements also form a key part of various other domestic and international efforts, including the European Union's proposed AI Act, where the European Parliament suggested adding in a transparency requirement concerning the use of copyrighted materials in AI training. It would therefore advance efforts to harmonize policymaking across the administration and internationally for the Office to support and help facilitate development of federal transparency requirements with respect to copyright and other matters and generative AI use of materials.

As noted in our initial comments, the usefulness and efficacy of such transparency requirements depends on the level of transparency required. For example, the European Parliament's proposed version of the AI Act would require a "sufficiently detailed summary of the use of training data protected under copyright law."[93] We believe that such a formulation is

[82] CA Comments at 17.

[83] NMPA Comments at 6.

[84] A2IM/RIAA Comments at 29.

[85] AG Comments at 29.

[86] NYT Comments at 6.

[87] AAP Comments at 7.

[88] Department for Professional Employees, AFL-CIO, Comments of AFL-CIO at 4 (2023).

[89] Getty Comments at 8.

[90] Writers Guild of America West & Writers Guild of America East, Writers Guild of America West and Writers Guild of America East Comment on USCO Notice of Inquiry on Copyright & Artificial Intelligence at 4 (2023).

[91] EWC Comments at 15-16.

[92] Centro Español de Derechos Reprográficos, Comments of the Centro Español de Derechos Reprográficos at 12-13 (2023).

[93] EUR. PARL., Amendments Adopted by the European Parliament on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council on Laying Down Harmonised rules on artificial intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021)0206 – C9-0146/2021 –

too vague, leaving the determination of "sufficiently detailed" up to the generative AI developers and the courts, and N/MA thereby strongly supports more detailed transparency requirements. These requirements should include a comprehensive, meaningful obligation to identify copyright-protected content sources for training, fine tuning, and other purposes—including identifying the work in question and the time of scraping—to allow rightsholders to construct a full chain of use.

With commitment, we believe implementing comprehensive data recording and transparency requirements is feasible both operationally and financially. Developers should be able to track data origins through metadata, ensuring transparency and responsible usage compliance, and potentially increasing performance analysis and output quality in the process. Indeed, other commenters noted that there is already a market for services and platforms that help with and automate such recordkeeping for AI developers.[94] Others operating in the licensing space noted that it is implausible that recordkeeping would place an unbearable burden on generative AI developers,[95] with ASCAP stating that digital streaming services "are able to maintain sufficient data to enable PROs . . . to identify the use of protected content and to compensate their members accordingly."[96]

While N/MA believes that self-regulatory or regulated requirements to proactively identify works used in training data are the preferable solution, we also support a suggestion, proposed by A2IM & RIAA, to consider establishing an administrative subpoena process, loosely modeled after section 512(h), where a subpoena can be issued upon the assertion of a good faith belief that one or more of the owners' copyrighted works have been used by an AI developer without authorization.[97] Such a provision would incentivize adequate recordkeeping because failure to comply by the developer would provide the copyright owner with an evidentiary presumption that the works identified in the subpoena were, in fact, reproduced.

d. Marketplace Licensing of News Media and Other Content Should be Supported.

The comments reveal strong support for the ability of marketplace licensing to respond to the needs of GAI, and little evidence of the marketplace failure that would be required to consider a compulsory regime. The strongest objections to marketplace licensing by developers appear primarily motivated by investment margins. Especially in light of the tremendous economic

2021/0106(COD))1 at Art. 28b(4)(c) (2023), available at https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf.

[94] *See, e.g.*, CA Comments at 90-91; A2IM and RIAA Comments at 32.

[95] *See* ASCAP, Comments of the American Society of Composers, Authors and Publishers on Artificial Intelligence and Copyright at 47-48 (2023); CCC Comments at 15.

[96] ASCAP Comments at 47-48.

[97] A2IM and RIAA Comments at 31.

benefits these companies and their backers are poised to enjoy, they should be required to factor content acquisition costs into their models, just like any other cost of doing business.

N/MA's initial comment detailed the viability of marketplace licensing for media content, noting its "members are by and large willing to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy and authoritative expressive content."[98] Others expressed similar views from their perspective, including:

- **ASCAP:** As demonstrated by the hundreds of thousands of businesses that currently license public performance rights from ASCAP, voluntary collective licensing is practically feasible and mutually beneficial for both creators and businesses that derive value from the use of copyrighted musical works. … The main obstacle to voluntary collective licensing is the lack of willingness on the part of AI providers to come to the negotiation table with the creators.[99]

- **Association of American Publishers:** Professional and scholarly publishers already employ licensing arrangements to facilitate access to their databases, whether for non-commercial research purposes or for commercial use. Other sectors of the publishing industry are exploring how they may facilitate access to their copyrighted works, consistent with the rights their authors have assigned to them.[100]

- **CCC:** Among many others, the prominent copyright holders Associated Press, Getty Images and vAIsual all offer licenses… CCC already offers market-based, global non-exclusive voluntary licenses to support AI in the commercial research, schools, and education technology sectors. These licenses were built with rightsholders and users based on agreed understandings of needs and market conditions.[101]

- **Shutterstock:** Shutterstock has built robust demand for ethically sourced AI training data. It has partnered with multiple companies that are interested in training their AI models on licensed data from Shutterstock, including LG and Meta.[102]

- **Getty:** [C]ompulsory or extended collective licensing schemes are not desirable when a marketplace for direct licensing already exists, which is the case with the licensing visual works and metadata to use in connection with the training and development of AI Models.[103]

[98] N/MA Comments at 4.
[99] ASCAP Comments at 4.
[100] AAP Comments at 24.
[101] CCC Comments at 12.
[102] SHUTTERSTOCK, Comments of Shutterstock, Inc. at 3 (2023).
[103] Getty Comments at 21.

- **NMPA:** While specific terms will vary between licensees, the process of obtaining voluntary licenses to musical works would not be fundamentally different for AI model developers than it is for the many other digital platforms that license music in the free market. Indeed, many of the major companies in the AI model development space, including Alphabet, Amazon, Apple and Meta, have significant experience negotiating voluntary licenses for music on an industry-wide basis for their other digital services.[104]
- **A2IM & RIAA:** There is no need or basis for government intervention in the licensing market for recorded music. The market is demonstrably working.[105]

Additional comments spoke to the ability of open licenses to work at mass scale.[106] And last month, following the close of comments, OpenAI announced a data partnership initiative to work with organizations to "create open-source and private datasets for AI training."[107]

Developer comments opposed to marketplace licensing were generally unpersuasive. For example, a16z objected that "[t]he fact that large rights owners are willing to strike deals is irrelevant, as such deals would only permit use of a small amount of the content needed to adequately train AI systems."[108] A few misconceptions underpin this argument, which rather brazenly urges that the more the developers copy, the less they should have to pay, yet crucially ***admits the viability of obtaining permission for some of the most valuable content to train systems.*** As noted, N/MA's research finds news media content overweighted up to 100 times in training datasets, and a separate Washington Post investigation revealed that N/MA member content is significantly represented amongst the top training sources for LLMs.[109] There is no technical reason why GAI models must ingest these copyright-protected expressive works apart from the desire to incorporate that very expression, and to use it requires permission. A16z's statement also ignores the emergence of licensed and other ethically-sourced models, and does not quantify the amounts at issue (*e.g.,* how much content is needed for training, how much is unavailable under license). It ignores the potential for voluntary collective licensing options to facilitate access to aggregated content from numerous publishers, and separately, other rightsholders. And it ignores public domain material, open

[104] NMPA Comments at 24.

[105] A2IM and RIAA Comments at 26.

[106] *See, e.g.,* Wikimedia Foundation, Responses to the United States Copyright Office at 6-7 (2023).

[107] OpenAI, *OpenAI Data Partnerships*, Blog (Nov. 9, 2023), https://openai.com/blog/data-partnerships. As noted in our initial comments, OpenAI has also entered into a licensing agreement with the Associated Press. *See* AP, *ChatGPT-maker OpenAI signs deal with AP to license news stories,* AP (July 13, 2023), https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[108] A16z Comments at 9.

[109] *Supra* note 5. The Post's report also revealed a substantial amount of government and public domain content, including Google Patents, and publications by the NIH or UK Parliament, which would remain available for use.

licenses, and, if eventually proven necessary, the ability to explore a separate treatment limited to addressing orphan works.

More fundamentally, it deems "irrelevant" the core balance of copyright in the interest of facilitating a frictionless path for developers, and greater returns for venture capital investment. Claiming that the need to license makes a business model too expensive would have fair use swallow the purpose of copyright (*i.e.,* to incentive the creation and distribution of new works) just to make content-consuming, enormously valued business models more lucrative.

Meta objected on the grounds that monetary benefits to publishers and other content creators would be negligible.[110] This frankly specious position misleadingly measures a single snapshot of time rather than the long-term value of using copyrighted content. Aggregating smaller amounts of revenue over time is a standard and typical foundation for internet, media, and other digital business models (*e.g.*, subscription, advertising, or as-a-service models). The power of these business models is demonstrated throughout the economy, including in media publishing, which depends on subscription and advertising revenue over time, cloud computing,[111] and music and video streaming.[112] Indeed, venture capital values generative AI companies based on projections that revenue will accrue over time: Bloomberg Intelligence recently predicted that generative AI will become a $1.4 ***trillion*** market by 2032, mainly due to incremental revenue projections.[113] Mark Zuckerberg himself noted that revenue from Meta's LLaMA2 will not be a large amount in the near term, but will grow over time.[114] Meta's position is also undermined by a16z's warning that paying for content could cost developers "tens or hundreds of billions of dollars a year in royalty payments."[115] And, perhaps presuming that the best path is that which is most efficient for Meta, Meta's statement overlooks that licensing

[110] Meta Comments at 20.

[111] Amazon Web Services, for example, saw $80 billion of revenue in 2022. *See* Simon Sharwood, *Google Cloud Makes Its First Profit, 15 Years After Launching*, The Register (Apr. 26, 2023), https://www.theregister.com/2023/04/26/alphabet_q1_2023/.

[112] The global value of music streaming was $41.5 billion in 2022, the highest ever for the music industry. *See* Stuart Dredge, *Global Value of Music Copyright Grew 14% to $41.5bn in 2022*, Music Ally (Nov. 6, 2023), https://musically.com/2023/11/06/global-value-of-music-copyright-grew-14-to-41-5bn-in-2022/.

[113] Bloomberg, *Generative AI to Become a $1.3 Trillion Market by 2032, Research Finds* (Jun. 1, 2023), https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/.

[114] Alexandra Barinka, *Meta to Charge Cloud Providers for AI Tech That It Said Was Free*, Bloomberg (Jul. 26, 2023), https://www.bloomberg.com/news/articles/2023-07-26/meta-to-charge-cloud-providers-for-ai-tech-that-it-said-was-free.

[115] Kali Hays, *Andreessen Horowitz Would Like Everyone to Stop Talking about AI's Copyright Issues, Please*, Business Insider (Nov. 7, 2023), https://www.businessinsider.com/marc-andreessen-horowitz-ai-copyright-2023-11.

valuations do not need to be the same for all types of content, nor would all permissive uses be expected to be royalty bearing.

e. Copyright Office Should Pay Particular Attention to Competition Concerns That Have the Potential to Distort the Copyright Marketplace.

The Office should pay special attention to the distortive effects competition concerns may play in the copyright marketplace. Public policy should discourage a system that allows one sector of the economy to prosper at the expense of another that makes all the investments and accepts all of the risk and responsibility. These concerns were highlighted by the submission of the Federal Trade Commission, which stated the FTC's intention to focus on the intersection of copyright and antitrust concerns in the generative AI marketplace, noting that "under certain circumstances, the use of pirated or misuse of copyrighted materials could be an unfair practice or unfair method of competition under Section 5 of the FTC Act."[116] It warned that incumbents could use their data and computing resources to "unlawfully entrench their market positions in AI and related markets."[117] N/MA strongly supports and encourages the FTC's focus on the issue. The failure of generative AI developers to seek licenses and adequately compensate rightsholders for the use of their content in AI training poses significant competition concerns, especially given larger developers' dominance in related verticals and their ability to impose conditions, access decisions, or preferences across a range of products and services. In light of the Office's current exploration in the interrelation between copyright and competition interests, N/MA strongly urges the Office to follow the logical conclusion of its Study on Ancillary Copyright Protections for Publishers and support competition-based changes and solutions, such as the *Journalism Competition and Preservation Act* (JCPA).

f. The Office Should Recommend Policies to Address Notorious Pirate Sites and to Step Up Enforcement Efforts.

Initial comments submitted to the Office reveal a widespread, shared concern over the ingestion of materials from sources known to contain pirated content.[118] While some developers may attempt to curb this practice and prevent content being scraped from unscrupulous sources, developer-by-developer approaches are unlikely to be effective in the long run. N/MA encourages the Office and the Administration to publicize the risks of notorious pirate sites and step up enforcement efforts so they are not included in generative AI training datasets. As explained in our initial comments and supported by other commenters, an "opt-

[116] FEDERAL TRADE COMMISSION, Comment of the United States Federal Trade Commission at 5 (2023) [hereinafter FTC Comments].
[117] FTC Comments at 4.
[118] AG Comments at 7-9; A2IM and RIAA Comments at 12; AAP Comments at 8; Getty Comments at 14 fn 24.

out" or rightsholder reporting regime is neither feasible nor consistent with U.S. copyright principles, and the burden of enforcement should not be placed on the rightsholders alone.

4. Conclusion

N/MA and our member publishers remain optimistic about the exciting opportunities AI and generative AI technologies present for society and media publishers themselves. We are also heartened that the Office received so many thoughtful, reasonable comments from a wide range of stakeholders. However, we strongly encourage the Office to reject the views of a minority of vocal commenters who misapply and misinterpret copyright laws against the basic principles of copyright in the United States.

Copyright law does not accommodate a structure where generative AI companies get all the benefits of using creative content without carrying any of the burdens—no licensing or compensation of intellectual property without which these systems could not exist, no transparency, no standards and practices review for defamatory or otherwise harmful content, no liability for infringing outputs. No other industry works like this, and the nascent generative AI industry should not be encouraged or enabled to develop in this manner. When balancing policy goals, the Office should consider the critical role that journalism and media publishing play in our democratic society and processes—and has played since the founding of our country—and work to minimize outcomes that deviate from core copyright principles.

The countervailing risk, articulated by commenters like Andreessen Horowitz, appears to be that if companies turn out to have bet wrong on how courts will interpret fair use, it would "significantly disrupt" an "enormous investment of private capital."[119] We believe these deep pocketed actors[120] can withstand any potential disruption, and that AI innovation will be safer, more reliable, and more sustainable, if developed in accordance with copyright law.

There is also no reason to believe that enabling rightsholders to enforce their copyrights would lead to a competitive disadvantage vis-à-vis companies based in other countries and regions with explicit laws allowing for text and data mining in certain circumstances. Many of these laws are untested in the AI context and regions such as the EU are considering transparency measures to enable copyright owners to identify the use of their content in training datasets.

N/MA thanks the Office for the opportunity to provide these comments and looks forward to the Office's report(s). We stand ready to answer any further questions the Office may have.

[119] *Supra* note 115.

[120] *Supra* note 63 ("The truth is that the investment community sees AI as a potential goldmine. One study placed the infusion of investment cash into the market in the last quarter alone at nearly $18 billion — rising higher even as investments in other startup categories have been shrinking.").

Respectfully submitted,

Danielle Coffey
President & CEO
News/Media Alliance

Regan Smith

Regan Smith
Senior Vice President & General Counsel
News/Media Alliance

December 6, 2023