NEWS MEDIA
ALLIANCE

Before the
U.S. COPYRIGHT OFFICE

Artificial Intelligence and Copyright

Docket No. 2023–6

Submitted October 30, 2023

COMMENTS OF THE NEWS/MEDIA ALLIANCE

Introduction.

The News/Media Alliance ("N/MA") welcomes the opportunity to submit comments in response to the notice of inquiry regarding the U.S. Copyright Office's study of the copyright law and policy issues raised by generative artificial intelligence (AI). The last few years have witnessed the rise of AI systems and applications that have the potential to greatly reshape the digital marketplace and alter many features of public life.

This is particularly true of generative AI ("GAI") technologies,[1] including the introduction of large language models ("LLMs") and related applications (such as chatbot interfaces) to consumers and the digital marketplace, the focus of our comment. N/MA members would welcome working with generative AI developers to help build and grow these technologies, in ways that benefit all actors in the supply chain and society at large. News and media publishers recognize the potential opportunities for users, businesses, and society alike, and many

[1] Our comment adopts the Copyright Office's understanding of generative AI technologies as "capable of producing outputs such as text, images, video, or audio (including emulating a human voice) that would be considered copyrightable if created by a human author" based on "'learning' statistical patterns in existing data, which may include copyrighted works." U.S. Copyright Office, Artificial Intelligence and Copyright, 59942 Fed. Reg. 88 (167), (Aug. 30. 2023) available at https://www.govinfo.gov/content/pkg/FR-2023-08-30/pdf/2023-18624.pdf.

members are exploring how to responsibly utilize generative AI technologies in their workstreams.

But to fulfill their societal potential, technological innovations must be advanced in a sustainable manner. Not only are generative AI models often trained on copyright-protected, professionally created material, many applications also act as direct competitors to publishers, providing informational and cultural content to the public, and drawing readers and advertisers away from publisher websites. In effect, publishers invest in producing high-quality content that is taken without permission to train the AI systems and used to produce substitutional, expressive AI-generated ”outputs“ that then compete directly with publisher content, reducing publisher revenues and employment, tarnishing their brands, and undermining their relationships with readers. The continued unlicensed use of journalistic reporting portends injury to the public interest that it serves, and may hinder the progress of generative AI innovations.

N/MA is grateful to the Copyright Office for undertaking this important and timely study and facilitating dialogue among the stakeholders and policymakers. As President Biden's Executive Order issued on October 30, 2023, recognizes, mitigating against risks posed by AI is vital in order to realize its potential for society.[2] While AI is exciting, and N/MA supports the principled development of generative AI technologies, unregulated, it also poses a significant threat to the pillars of a healthy and informed democracy. Our members are gravely concerned that some developers have to date violated the legal rights of publishers, using their copyrighted material without permission or compensation and tarnishing their brands. Copyright law simply does not require publishers to train their replacements in this way.

N/MA has vigorously advocated for its members' interests on issues surrounding generative AI to advance our members' interests and to address risks that unsustainably deployed generative AI poses to the continued viability of the news business. In 2020, N/MA filed comments with the United States Patent and Trademark Office focusing on the issue of systemic ingestion of copyright protected content for machine learning purposes.[3] These comments, attached here,[4] discussed how the current case law provides protections for media content against such use

[2] *See* The White House, Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (October 30, 2023), available at https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[3] News/Media Alliance, Re: Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation (Jan. 10, 2022) available at http://www.newsmediaalliance.org/wp-content/uploads/2020/01/News-Media-Alliance-AI-Comments-with-USPTO.pdf.

[4] See Appendix B.

and highlighted the need for stronger enforcement. More recently, N/MA published a set of AI principles covering issues related to intellectual property, transparency, accountability, fairness, safety, and design, that we hope will inform AI policy development in the United States.[5] We also joined a similar set of global principles, together with 27 other publisher organizations.[6]

Now, N/MA contemporaneously publishes a White Paper, also attached here[7] and referenced below, on AI developers' pervasive use of publisher content in generative AI training. The White Paper includes a technical analysis regarding the use of publisher content in generative AI applications and discusses the effects and legal implications of such use. A few takeaways from that analysis include:

- Developers have copied and used news, magazine and digital media content to train LLMs.
- Popular curated datasets underlying LLMs significantly overweight 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.
- Other studies show that news and digital media ranks third among all categories of sources in Google's C4 training set, which was used to develop Google's generative AI-powered products like Bard. Half of the top ten sites represented in the data set are news outlets.
- LLMs also copy and use publisher content in their outputs. LLMs can reproduce the content on which they were trained, demonstrating that the models retain and can memorize the expressive content of the training works.

In short, generative AI systems should be held responsible and accountable, just like any other business. The risks of unregulated AI development and use are too high, both for society and a competitive online economy alike. N/MA hopes that the Office's study will bring attention to the systemic and wide ranging infringement by some generative AI developers and help grow emerging practices for licensed use of publisher content.

[5] NEWS/MEDIA ALLIANCE, AI PRINCIPLES (2023) available at http://www.newsmediaalliance.org/wp-content/uploads/2023/04/FINAL-UPDATED-AI-Principles_4-20-23.pdf.

[6] GLOBAL PRINCIPLES ON ARTIFICIAL INTELLIGENCE (AI) (2023) available at http://www.newsmediaalliance.org/wp-content/uploads/2023/09/FINAL-Global-AI-Principles-Formatted_9-5-23.pdf.

[7] *See* N/MA, 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 (2023) [hereinafter N/MA, WHITE PAPER], Appendix A.

The digital ecosystem will benefit from consensus around protection and partnership. N/MA's 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. A constructive solution could benefit all interested parties and society at large, and avoid protracted litigation. And fruitful cooperation will also help maximize the potential of generative AI technologies, by helping ensure they are developed using high-quality and human created works.

The Office's inquiry can also help inform the development of carefully considered and well-balanced AI policy at the federal level to mitigate against unintended consequences and harms to the media and other creative industries. We look forward to engaging with the Office, the Congress, and the Administration moving forward.

About News/Media Alliance.

N/MA is a nonprofit organization headquartered in Washington, D.C., representing the newspaper, magazine, and digital media industries, and empowering members to succeed in today's fast-moving media environment. N/MA represents over 2,200 diverse publishers in the United States and internationally, ranging from the largest news and magazine publishers to small, hyperlocal newspapers, and from digital-only and digital-first outlets to print papers and magazines.

In total, N/MA's membership accounts for nearly 90 percent of the daily newspaper circulation in the United States, nearly 100 magazine media companies with over 500 individual magazine brands, and dozens of digital-only properties. Its members publish high-quality original content on topics ranging from news to culture, sports, entertainment, lifestyle, and virtually any other interest. N/MA diligently advocates for its members on a broad range of current issues affecting them, including copyright policy that directly relates to our members' ability to monetize their content and support their continued investments in high-quality content production.

N/MA members play a vital role in their communities and in fostering an informed public and the public trust necessary for democracy. Publishers invest considerable time and resources to produce journalism and original creative content that combats misinformation, encourages democratic engagement, strengthens community ties, lowers municipal borrowing costs, safeguards consumers, keeps decision makers accountable, gives people something to talk about, and supports the free flow of ideas and information.[8] Our members also support local

[8] *See, e.g.*, Matthew Gentzkow, et al., *The Effects of Newspaper Entry and Exit on Electoral Politics*, 101 AM. ECON. REV. 2980 (2011); Danny Hayes & Jennifer L. Lawless, *As Local News Goes, So Goes Citizen Engagement: Media,*

economies by providing small and medium enterprises, local businesses, and community organizations with a cost-effective way to reach potential local customers through advertising and online content. However, despite these considerable benefits, and the increased audience for publisher content, far too many publishers are struggling to survive in the online ecosystem, partially due to the unauthorized scraping and use of their protected content.

The news, magazine, and digital media industries' contribution to the U.S. economy and society is considerable, with estimated revenues of newspaper and magazine publishers amounting to approximately $45 billion.[9] Newspaper newsrooms were estimated to directly employ approximately 31,000 people in 2020, not including additional indirect employment effects, while magazines employed over 73,000 directly and supported a total of over 219,000 jobs in 2021.[10] Employment in digital-native newsrooms, meanwhile, has increased from approximately 7,400 in 2008 to over 18,000 in 2020.[11] The content produced by these professionals has a huge audience, with N/MA member publishers reaching hundreds of millions of Americans every year. The share of digital audience is large for both magazine and newspaper publishers, with news publishers having over 200 million unique visits and 6.7 billion page views per month while 40 percent of magazine readers access content on mobile

Knowledge, and Participation in U.S. House Elections, 77 Journal of Politics 447 (2014); Mary Ellen Klas, *Less Local News Means Less Democracy*, Nieman Reports, Sep. 20, 2019, https://niemanreports.org/articles/less-local-news-means-less-democracy/; The Covington News, *The Benefits of Local Newspapers*, [n.d.] https://www.covnews.com/nie/benefits-local-newspapers/; Pengjie Gao, Chang Lee, & Durmot Murphy, *Financing Dies in Darkness? The Impact of Newspaper Closures on Public Finance*, 135 Journal of Financial Economics 2 (2020); The British Psychological Society, *Why Magazines Matter,* The Psychologist, Nov. 25, 2016, https://www.bps.org.uk/psychologist/why-magazines-matter.

[9] *See* Pew Research Center, Fact Sheets: State of the News Media (Jun. 29, 2021) available at http://www.journalism.org/fact-sheet/newspapers/ (last visited Oct. 13, 2023); Amy Watson, *Estimated Aggregate Revenue of U.S. Periodical Publishers from 2005 to 2020*, Statista, Dec. 5, 2022, available at https://www.statista.com/statistics/184055/estimated-revenue-of-us-periodical-publishers-since-2005/ (last visited Nov. 17, 2022); Adam Grundy, *Service Annual Survey Shows Continuing Decline in Print Publishing Revenue*, U.S. Census Bureau, Jun. 7, 2022, available at https://www.census.gov/library/stories/2022/06/internet-crushes-traditional-media.html.

[10] Pew Research Center, Fact Sheets: State of the News Media (Jun. 29, 2021) available at http://www.journalism.org/fact-sheet/newspapers/ (last visited Oct. 13, 2023); Mason Walker, *U.S. Newsroom Employment Has Fallen 26% since 2008*, Pew Research Center, Jul. 13, 2021, https://www.pewresearch.org/short-reads/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/; MPA – The Association of Magazine Media, Magazine Media Factbook, (2021) available at https://www.newsmediaalliance.org/wp-content/uploads/2018/08/2021-MPA-Factbook_REVISED-NOV-2021.pdf.

[11] Pew Research Center, Fact Sheets: State of the News Media (Jun. 29, 2021) available at http://www.journalism.org/fact-sheet/newspapers/ (last visited Oct. 13, 2023).

devices.[12] This is in addition to the millions who access content on digital-only publishers' websites.

In order to continue investments into high-quality journalism and digital content, publishers require strong intellectual property protections and a vibrant, open, and fair online competitive environment that, when functioning at its best, rewards quality, creation, and innovation. Today, a few dominant online platforms control the digital ad ecosystem and the distribution of digital content, posing an existential threat to many publishers, especially small and local newspapers. The numbers in the preceding paragraph take on a different meaning when you consider that in less than 20 years, newspaper circulation and advertising revenues dropped from $57.4 billion in 2003 to an estimated $20.6 billion in 2020, while magazines witnessed a drop from $46 billion in 2007 to $23.92 billion in 2020.[13] In short, news publishers' revenues decreased by almost two-thirds and magazines have lost almost half of their revenues. In total, 2,500 newspapers have either closed or merged since 2004.[14] Similarly, there has been a substantial loss of community newspapers such that at least 200 counties, representing four million Americans, no longer have a local newspaper.[15] These losses are more likely to affect already disenfranchised people and communities, with many of the lost or failing newspapers located in areas that are less affluent than the national average. While magazine publishers have generally fared somewhat better than newspaper publishers, many have been forced to reduce print days or cut print editions completely, in an effort to lower costs.[16] Together, these trends have led to substantial job losses across the publishing industry.[17]

[12] News/Media Alliance, News Advertising Panorama (2020) (publicly available to N/MA members only; on file with author); MPA – The Association of Magazine Media, Magazine Media Factbook, (2021) available at https://www.newsmediaalliance.org/wp-content/uploads/2018/08/2021-MPA-Factbook_REVISED-NOV-2021.pdf.

[13] PEW Research Center, Fact Sheets: State of the News Media (Jun. 29, 2021) available at http://www.journalism.org/fact-sheet/newspapers/ (last visited Oct. 13, 2023)); Amy Watson, *Estimated Aggregate Revenue of U.S. Periodical Publishers from 2005 to 2020*, Statista, Dec. 5, 2022, available at https://www.statista.com/statistics/184055/estimated-revenue-of-us-periodical-publishers-since-2005/ (last visited Oct. 13, 2023).

[14] Penny Abernathy, Report: The State of Local News 2022 (2022), available at https://localnewsinitiative.northwestern.edu/projects/state-of-local-news.

[15] *Id.*

[16] *See* Beth Braverman, *How Magazine Publishers Are Cutting Print Costs to Improve Profits*, Folio Magazine, Aug. 2, 2021, https://archive.foliomag.com/magazine-publishers-cutting-print-costs-improve-profits/; Peter Houston, *2021 in Print: Newspapers' Decline Continues, But for Magazines … It's Complicated*, What's New in Publishing, Dec. 20, 2021, https://whatsnewinpublishing.com/2021-in-print-newspapers-decline-continues-but-for-magazines-its-complicated/.

[17] *See* Mason Walker, *U.S. Newsroom Employment Has Fallen 26% Since 2008*, Pew Research Center Jul. 13, 2021, available at https://www.pewresearch.org/fact-tank/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/; Bureau Of Labor Statistics, Occupational Outlook Handbook: Reporters, Correspondents, And News Analysts, [n.d.] available at https://www.bls.gov/ooh/media-andcommunication/reporters-correspondents-and-broadcast-news-analysts.htm (last visited Nov. 17, 2022).

General Questions

Our responses to the Office's specific questions are below. N/MA may submit supplemental comments in response to other questions raised by the Office, or by other commenters.

1. As described above, generative AI systems have the ability to produce material that would be copyrightable if it were created by a human author. What are your views on the potential benefits and risks of this technology? How is the use of this technology currently affecting or likely to affect creators, copyright owners, technology developers, researchers, and the public?

N/MA recognizes the potential benefits and is broadly supportive of AI applications and technologies, with many of our members using AI—including generative AI—in various ways throughout their business operations. These uses may include content ideation and research, content optimization, improving internal efficiency, and content review and distribution activities.[18] Generative AI applications can provide an important tool in newsgathering and research efforts by determining sources for research and interviews, identifying content opportunities and aggregating and synthesizing information. Publishers may also use generative AI systems to generate headlines, outline articles, and write first drafts and FAQs subject to human review—to mention a few examples—while also utilizing AI to improve search engine optimization (SEO). Journalists and authors may benefit from generative AI in activities ranging from proofreading to distribution through social media channels and newsletters.

To be sustainable, however, generative AI development and use must be responsible, regulated, and accountable, with appropriate permission and compensation paid to publishers for the copying and use of their protected works throughout the product cycle. Without effective enforcement, regulation, and standards—including a requirement for AI developers to seek permission from rightsholders for uses of their protected content to train competitive products—AI can lead to considerable harms. These harms may include the undermining of the foundation of our democracy through the further weakening or outright closure of newspapers, magazines, and digital outlets—especially local ones—the spread of mis- and disinformation, and reduced access to reporting that can fundamentally only be created by humans—based on extensive fact-gathering, interviews, and judgment. An engaged and informed citizenry depends on the existence and availability of reliable and accurate reporting and analysis by outlets the public trusts. Unlike generative AI systems that may make up facts and disclaim

[18] *See, e.g.,* Elite Truong, *Local News and AI,* AM. PRESS INST., August 7, 2023, https://americanpressinstitute.org/publications/articles/local-news-and-ai/.

liability for doing so,[19] publishers accept responsibility for the content they publish, ensuring that the information presented to the public is of high quality. In a world flooded by easily accessible, synthetic information of unknown quality, real information becomes harder to identify and trust in our democratic system harder to upkeep.

In addition to these significant societal harms, the negative effects of unsustainable AI development practices on publishers small and large can lead to substantial job losses and a devaluing of journalistic content that will undermine these creative industries. In short, while AI presents many potential benefits to both publishers and the public at large, unregulated generative AI risks driving existing publishers out of business and disincentivizing continued investments in new, original content. This result would undermine the goal and purpose of the Copyright Clause of the Constitution, and diminish the essential role of the Press envisioned by the Founders. (And potentially also harming the further development of generative AI models through model collapse, as discussed further below.)

2. Does the increasing use or distribution of AI-generated material raise any unique issues for your sector or industry as compared to other copyright stakeholders?

The increasing use and distribution of generative AI systems and applications, as well as AI-generated materials, raises substantial, unique concerns for newspaper, magazine, and digital media publishers. While the interests of publishers and generative AI developers could align, for example, in a fair exchange of licensing revenues for access to high-quality training materials to facilitate the continued improvement of the models, the promise of partnership has not yet materialized except in a few narrow instances.[20] Instead of entering into legal licensing agreements with publishers, generative AI developers have chosen to scrape publisher content without permission and use it for model training and in real-time (grounding)[21] to produce outputs (often in the form of lengthy, expressive summaries) that can directly compete with publisher content and products. And they literally are making billions doing it.[22] Not only can

[19] For example, OpenAi has taken the position in litigation that it is not liable for claims for defamation. "Because any ChatGPT user verifies at signup that they "take ultimate responsibility for the content being published," OpenAI says that, "as a matter of law, this creation of draft content for the user's internal benefit is not 'publication.'" Ashley Belanger, *Will ChatGPT's hallucinations be allowed to ruin your life?*, ARS TECHNICA, Oct. 23, 2023, https://arstechnica-com.cdn.ampproject.org/c/s/arstechnica.com/tech-policy/2023/10/will-chatgpts-hallucinations-be-allowed-to-ruin-your-life/amp/.

[20] See discussion on existing licensing deals below.

[21] *See* N/MA, WHITE PAPER at 17-18 (2023), Appendix A; Jordi Ribas, *Building the New Bing*, MICROSOFT BING BLOGS, Feb. 21, 2023, https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing#:~:text=Selecting%20the%20relevant%20internal%20queries,this%20method%20is%20called%20grounding.

[22] *See, e.g.*, Jagmeet Singh & Ingrid Lunden, *OpenAI Closes $300M Share Sale at $27B-29B Valuation,* TECHCRUNCH (Apr. 28, 2023) https://techcrunch.com/2023/04/28/openai-funding-valuation-chatgpt/; Deepa Seetharaman & Berber Jin, *OpenAI Seeks New Valuation of Up to $90 Billion in Share Sale*, WALL ST. J. (Sep. 26, 2023)

generative AI systems and applications respond to user queries using publisher content but, as discussed in more detail below, an AI chatbot or search interface can, and does, produce outputs that include verbatim quotes and/or closely paraphrases publisher stories.

The members of the News/Media Alliance are deeply concerned about this unauthorized and unlawful use of their expressive content by large technology companies that do not shoulder the cost of reporting the news or producing creative content, but who capitalize on the results of that valuable work. Copyright law does not require publishers to train their replacements in this manner. In effect, publishers make the investments and take the risks—including sending journalists into harm's way—while generative AI developers reap the rewards of users, data, brand creation, subscription fees, and advertising dollars. This is freeriding.

The continued unlicensed use of reporting—including entire corpora of unique publisher content, amounting up to millions of stories—portends injury, not just to the news industry, but to the public interest that it serves: an online world that is dominated by AI-generated, inferior yet substitutional content will leave the public with watered-down, less reliable outputs and fewer news outlets with the resources necessary to provide critical original reporting. As district court judge Denise Cote's decision in *Associated Press v. Meltwater U.S. Holdings, Inc.* explained with respect to direct scraping of news content, copyright law does not allow for democracy to be imperiled in this manner:

> *[T]he world is indebted to the press for triumphs which have been gained by reason and humanity over error and oppression … Permitting [Meltwater] to take the fruit of [AP's] labor for its own profit, without compensating [AP], injures [AP's] ability to perform [its] essential function of democracy.*[23]

In addition to decreasing readership, the unauthorized use of publisher content to produce outputs that include inaccuracies also devalues publisher brands and creative content by muddling the source of the original content and misattributing information or misinformation to unrelated publishers or journalists.[24] This is especially damaging as many of N/MA's

https://www.msn.com/en-us/money/companies/openai-seeks-new-valuation-of-up-to-90-billion-in-share-sale/ar-AA1hiJ9W.

[23] *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 553 (S.D.N.Y. 2013).

[24] Julia Black @mjnblack, X (Apr. 4, 2023, 7:48), https://twitter.com/mjnblack/status/1643324719108706304; Kate Crawford @katecrawford, X (Apr. 4, 2023, 19:42), https://twitter.com/katecrawford/status/1643323086450700288 (Journalist doing background research on an interview subject using ChatGPT was provided with cites and links to two non-existent articles critical of the subject, one by MIT Technology Review.); Chris Moran, *ChatGPT is Making up Fake Guardian Articles. Here's How We're Responding*, Guardian, Apr. 6, 2023, https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-technology-risks-fake-article; James Warrington, *AI is 'Polluting the Pool of Human Knowledge', News Publishers Warn*, The Telegraph, Oct. 1,

members have spent years or decades—sometimes even centuries—building their reputation as reliable and trustworthy content producers, providers, and curators. This reputation is vital for their success, with readers associating their brands with content that has been researched, vetted, proofread, and carefully considered by consummate professionals they know and trust. Indeed, trusted journalism can be an antidote to the mis- and disinformation problem.[25]

It is therefore particularly concerning when a generative AI system attributes material that is blatantly false to a publisher who has never published such information. As one example, take the case of Jonathan Turley, a law professor who ChatGPT falsely accused of sexually harassing a student, attributing the information to a non-existent news article by The Washington Post.[26] In the same research experiment, conducted by Professor Eugene Volokh, ChatGPT made other similarly false allegations, citing articles that did not exist from publishers such as the Miami Herald and the Los Angeles Times. These “hallucinations,'' or massive errors, are a recognized propensity of many generative AI models that can spread misinformation and cause real harm to publisher brands. Other examples of the dangers of “hallucinations'' and other harms include summaries of articles by reputable publishers combining information from unreputable sources and the proliferation of deepfake photographs in politics.[27] Publishers recognize these pitfalls and while some may use AI as a tool in newsgathering and content production processes, they accept legal responsibility for the content they publish and understand that the outputs are often not reliable and require human editing and supervision before publication—something that generative AI systems typically do not have.

To mitigate these risks, it is essential that generative AI training datasets, systems, and applications be based on reliable, trustworthy, and high-quality content with adequate safeguards to deter misinterpretations and the creation of false information based on that content. To do so sustainably and lawfully—in a manner that protects the public interest, including professional journalism—generative AI developers should license content from publishers for training and grounding purposes based on fair and transparent negotiations, as

2023, https://www.telegraph.co.uk/business/2023/10/01/news-publishers-warn-ai-will-pollute-human-knowledge/.

[25] Jeff Clune, *AI-enabled Scams Will Proliferate*, MACLEANS, Oct. 12, 2023, https://macleans.ca/society/technology/ai-scams/. (“As we prepare for AI scams to proliferate, the best advice I can offer is for people to seek out and hold onto the sources they trust most-—whether that is the New York Times or a particular reporter. But even then they must make sure they are in fact getting information from that source.”).

[26] Pranshu Verma & Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused*, WASH. POST, April 5, 2023, https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/.

[27] Karen Weise & Cade Metz, *When A.I. Chatbots Hallucinate*, THE TELEGRAPH, Oct. 1, 2023, https://www.telegraph.co.uk/business/2023/10/01/news-publishers-warn-ai-will-pollute-human-knowledge/; William A. Galston, *Is Seeing Still Believing? The Deepfake Challenge to Truth in Politics*, BOOKINGS, Jan. 8, 2020, https://www.brookings.edu/articles/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/.

discussed in more detail below and in the attached White Paper. Only then can publishers recoup their investments in high-quality journalism while assuring developers that generative AI systems are built on authentic—not synthetic—content that is essential for reliable and trustworthy AI.

In the worst-case scenario, without an enforceable licensing market, high-quality publishers will slowly fail, forcing generative AI systems to rely on each other for training, leading to the gradual degradation in the availability of reliable and trustworthy reporting to our communities and system of democratic governance.[28] In fact, without human-generated quality content to train AI, researchers have found "that use of model-generated content in training causes irreversible defects in the resulting models," an effect they term "model collapse"[29] or "Model Autophagy Disorder (MAD),"[30] an analogy to mad cow disease:

> For instance, start with a language model trained on human-produced data. Use the model to generate some AI output. Then use that output to train a new instance of the model and use the resulting output to train a third version, and so forth. With each iteration, errors build atop one another. The 10th model, prompted to write about historical English architecture, spews out gibberish about jackrabbits.[31]

It is therefore in all of our collective interest that generative AI companies adhere with the letter and spirit of intellectual property law.

3. Please identify any papers or studies that you believe are relevant to this Notice. These may address, for example, the economic effects of generative AI on the creative industries or how different licensing regimes do or could operate to remunerate copyright owners and/or creators for the use of their works in training AI models. The Office requests that commenters provide a hyperlink to the identified papers.

The following papers and studies may help the Office in identifying the most pressing issues and concerns related to the proliferation of generative AI systems and applications and in identifying constructive solutions for continued success and innovation for all stakeholders:

[28] *Cf.* "Cory Doctorow, *The 'Enshittification' Of Tiktok*, WIRED, Jan. 23, 2023, https://www.wired.com/story/tiktok-platforms-cory-doctorow/.

[29] 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.

[30] Sina Alemohammad, et al., *Self-Consuming Generative Models Go MAD*, ARXIV, Jul. 4, 2023, available at https://arxiv.org/abs/2307.01850.

[31] Rahul Rao, *AI-Generated Data Can Poison Future AI Models*, SCIENTIFIC AMERICAN, Jul. 28, 2023, available at https://www.scientificamerican.com/article/ai-generated-data-can-poison-future-ai-models/.

- News/Media Alliance's White Paper on AI and Copyright, outlining how generative AI developers use publisher content, how it is stored and reproduced, the effects on publishers, and the legal implications of such use. The Paper incorporates a technical study analyzing the issues discussed (attached as Appendix A);

- News/Media Alliance's AI Principles that spell out publisher concerns and set out principles that should guide policy development in order to protect the sustainability of high-quality content online: https://www.newsmediaalliance.org/ai-principles/;

- Global AI Principles signed by 28 publisher organizations across the world, outlining principles that should guide AI policy development both at domestic and international fora: https://www.newsmediaalliance.org/global-principles-on-artificial-intelligence-ai/;

- Copyright Alliance's AI Position Paper that includes high-level discussion of the concerns and interplay of AI and the creative industries: https://copyrightalliance.org/policy/position-papers/artificial-intelligence/;

- The United Kingdom's House of Lords report on AI, outlining benefits and risks of AI as well as relevant policy discussions, including concerning the right of copyright owners to decide when their content is used for text and data mining: https://lordslibrary.parliament.uk/artificial-intelligence-development-risks-and-regulation/;

- A study on the potential for model collapse, noting that to "make sure that learning is sustained over a long time period, one needs to make sure that access to the original data source is preserved and that additional data not generated by LLMs remain available over time"[32]: https://arxiv.org/pdf/2305.17493.pdf;

- An article discussing the proliferation of AI generated information and the risks and opportunities of generative AI to publishers, stating that by "[f]looding the market with cheap information, AI can lead to decrease in overall quality of the Web and misinformation"[33]: https://www.inma.org/blogs/reader-revenue/post.cfm/ai-tsunami-revamps-the-competitive-strategy-of-news-media;

[32] Ilia Shumailov, et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, ARXIV, at 13 (May 27, 2023) available at https://arxiv.org/abs/2305.17493.

[33] Greg Piechota, *AI Tsunami Revamps the Competitive Strategy of News Media*, INTERNATIONAL NEWS MEDIA ASSOCIATION at [no pagination] (Jul. 23, 2023) https://www.inma.org/blogs/reader-revenue/post.cfm/ai-tsunami-revamps-the-competitive-strategy-of-news-media.

- European Magazine Media Association and European News Publishers' Association's Core Concerns on AI and Copyright, outlining many of the issues of concerns for publishers worldwide related to AI development (attached as Appendix C).

4. Are there any statutory or regulatory approaches that have been adopted or are under consideration in other countries that relate to copyright and AI that should be considered or avoided in the United States? How important a factor is international consistency in this area across borders?

5. Is new legislation warranted to address copyright or related issues with generative AI? If so, what should it entail? Specific proposals and legislative text are not necessary, but the Office welcomes any proposals or text for review.

N/MA responds to questions 4 and 5, concerning international approaches and domestic legislation, together.

N/MA asks the Copyright Office to analyze global and domestic regulatory and policy trends with the following backdrop in mind: generative AI technologies like LLMs may develop in ways that significantly benefit society. But LLMs carry the potential to significantly disrupt (or to augment) existing creative markets. To ensure generative AI systems remain beneficial to all, transparency measures will be crucial with respect to how and what copyrighted content was used by AI companies, and whether required permission was obtained from rightsholders. Developers and deployers of foundational models and follow-on configurations should be incentivized to cooperate with rightsholders to achieve the necessary transparency. And where needed—if no legal or contractual exception applies—permission should be obtained for the use of copyrighted material.

We applaud the Office for issuing this comprehensive and thoughtful notice, and for recognizing that this study will not exist in a vacuum, but amidst ongoing global business, legal, and policymaker discussions. As it considers where to make policy recommendations, or provide guidance on existing copyright law, the Office can also leave room for industry-led solutions, while helping guide and convene discussions.

Given the ongoing damage being experienced by publishers, we urge the Copyright Office to support a few concrete objectives in its policy Study, as well as exercise its regulatory authority to ensure that news media publishers can equitably access the copyright registration system. Specifically, N/MA recommends the Office prioritize the following:

- *Use:* The Office should clarify publicly that use of publishers' expressive content for commercial generative AI training and development is likely to compete with and harm publisher businesses, which is disfavored as a fair use. This conclusion follows naturally from existing case law, as discussed below. But such clarification would nonetheless be helpful now, to reduce uncertainty that may arise as multiple lawsuits progress through different district courts and circuits, and help the affected industries and policymakers move towards a clearer consensus on the existing law. It would also help avoid the need for litigation by incentivizing GAI companies to reach fair and negotiated agreements that compensate publishers for the past and ongoing use of their content. While the Office may prefer to weigh in on specific litigation directly in a judicial setting, the constellation of litigation matters that has and will continue to emerge may benefit from the Office's broad guidance on common issues and themes. The Office has historically played such a useful role in providing guidance to the public, Congress, and affected industries in similar contexts.[34] It should do so here, to reduce an extended period of uncertainty that may create a cloud on generative AI products, as well as the economic viability of publishers, journalists and authors, while various litigations proceed, potentially through protracted appeals.

- *Transparency:* Substantial transparency measures should develop around the ingestion of copyrighted materials for uses in generative AI technologies. The Office may consider principles raised in other jurisdictions, such as in the European Parliament's negotiating position on the Artificial Intelligence Act (AI Act),[35] with respect to promulgation and harmonization of transparency obligations. However, it should ensure that any proposals achieve the core objective of providing sufficient transparency into the ingestion and use of copyrighted materials to allow rights holders to sufficiently analyze such models.

- *Licensing:* As described below, the Office should use its expertise in licensing issues to encourage the further development of relevant models, including by acknowledging the

[34] *See, e.g.,* HEARING, SEN. UDALL RESPONSE, NATIONAL EMERGENCY LIBRARY, US. COPYRIGHT OFFICE (Apr. 16, 2020), available at https://copyright.gov/laws/hearings/Sen-Udall-Response-National-Emergency-Library.pdf; *see generally* Rulemaking Proceedings under Section 1201 of Title 17 (concerning *inter alia* whether proposed uses for which exemptions are sought are likely to be noninfringing).

[35] 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 – 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. N/MA expresses no opinion on the AI Act as a whole or any of its provisions, except for the provision proposed by the European Parliament that would impose transparency obligations on AI developers with regards to the use copyrighted materials.

potential feasibility of voluntary collective licensing to facilitate licensing for ingestion of materials for generative AI purposes. It should follow established law and Office policy in discouraging government regulation of licensing markets as a first resort. In doing so, it should consider that different creative industries have different interests, products, business models, policies and standards, and approaches to licensing works.[36] In light of these differences, it is not necessary to propose a "one size fits all" or "one stop shopping" approach to all forms of copyrighted works.

- *Registration:* The Copyright Office should swiftly promulgate an updated registration option, ideally implemented on an immediate, interim basis, that permits online news publishers to submit identifying material and register groups of news articles published online. This specific and actionable request follows years of discussions between the Office and N/MA and is tailored to accommodate what we understand are the limitations of the Office and Library's information technology systems. While we respect the Office's limited resources, considering the blatantly commercial emerging uses of copyrighted media publishing material taken from online sources for use in generative AI development and the current litigation landscape, the need for an updated registration option has boiled over and it should be established urgently.

- *Competition:* The Notice rightly acknowledges the interplay between copyright and competition policy. In light of the continued, large disparity in bargaining power between media publishers and very large online platforms, who are now in fact leaders in generative AI development, we urge the Office to build upon its 2022 press publishers study and support measures to correct this negotiating disparity, such as the Journalism Competition and Preservation Act.

- *Enforcement:* To address the question of protected content being scraped from third-party pirate websites, the Office could consider recommending the establishment of a process, similar to the USTR's Special 301 Review, that would identify, based on stakeholder feedback, known pirate sites that regularly reproduce copyrighted content and are therefore off-limits for AI training purposes, even if the pirate site owners would allow data scraping.[37]

[36] In separating out different interests, the Office can also consider practices of open licensing, and unique aspects of non-commercial works, non-professionally aspiring individual creators, as well as for user-generated material made available on an online platform.

[37] *Compare with* Emilia David, *RIAA Wants AI Voice Cloning Sites on Government Privacy Watchlist*, THE VERGE (Oct. 11, 2023) https://www.theverge.com/2023/10/11/23913405/riaa-ai-voice-cloning-threat-copyright-ustr.

While we believe that existing domestic copyright law is well-suited to address many of the challenges and opportunities presented by generative AI, there are numerous ongoing court challenges. The reality is that many publishers lack the resources to adequately enforce their rights against companies that are aggressively infringing them. As the legal landscape evolves, Congress and the Office should remain diligent to ensure that the law remains fit for purpose — to "encourage the production of original literary, artistic, and musical expression for the good of the public."[38] N/MA notes that the Congressional Research Service appears to have reached a similar conclusion with regards to a wait-and-see approach.[39]

The Office is also wise to consider the virtue of harmonization as it evaluates policy proposals. For example, the EU is currently working on multiple pieces of AI-related legislation, including the AI Act, the AI Liability Directive (soon to be taken up),[40] and a planned revision of the EU Copyright Directive in 2026.[41] Other governments, including the UK, are also considering significant reforms.[42] Harmonizing AI regulations will be vital given AI's global nature, but cannot come at the expense of domestic creative industries and publishers of original expressive material. The Office should support active involvement in international discussions, including from representatives of affected industries, to discourage foreign nations from establishing local climates that encourage AI-related development activities that would be prohibited under U.S. law.[43] It can also take into account the positive aspects of global approaches while rejecting approaches that overlook necessary granularity or protections for publishers in their measures. As noted in question 3, when considering European developments, we recommend the Office consult the attached list of core concerns on AI and Copyright of the European Magazine Media Association and the European News Publishers´ Association, published July 26, 2023.

N/MA may bring forward more concrete concerns or legislative proposals. We look forward to engaging with the Office, the Administration and the Congress as discussions move forward.

[38] *Fogerty v. Fantasy, Inc.*, 510 U.S. 517, at 524 (1994). The Office can especially monitor the understanding of fair use in various district court challenges.

[39] Christopher Zirpoli, Generative Artificial Intelligence and Copyright Law (Congressional Research Service, 2023) available at https://crsreports.congress.gov/product/pdf/LSB/LSB10922.

[40] Eur. Parl. Briefing, Artificial Intelligence Liability Directive, (Feb. 2023) available at https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf.

[41] EU Copyright Directive, Article 30 (2019) O.J. (Directive 2019/790) available at https://eur-lex.europa.eu/legal-content/en/TXT/HTML/?uri=CELEX:32019L0790.

[42] Digital Markets, Competition and Consumers Bill, 2023, H.C. Bill [350 2022-23] available at https://bills.parliament.uk/bills/3453.

[43] N/MA draws particular attention here to recently enacted overbroad TDM exceptions in Japan and Singapore, with neither one explicitly excluding commercial uses or requiring that the content is lawfully accessed.

Training

6. What kinds of copyright-protected training materials are used to train AI models, and how are those materials collected and curated?

There is no rational dispute about whether many generative AI companies copy third-party content without permission to train their models and develop their tools—they do. In order to train an AI model, system, or application that generates language, visualizations, or sounds that resemble human-created works, developers process potentially billions of works, often amounting to trillions of words and millions of photos and audiovisual works that are scraped from the internet. The Copyright Alliance's comments in response to this notice discuss issues related to this question more broadly. In the case of publishers, however, these works often include content that is behind paywalls or other technical measures—potentially even with CAPTCHA protections—and not broadly accessible to the public without subscription. Some companies, such as Bright Data, even advertise their products' ability to evade CAPTCHA, paywalls, and other common ways to prevent scraping.[44] Following the initial training, fine tuning a model may require the processing of additional works and sources.

While developers—directly or indirectly—ingest (or copy) copyrighted works from various online sources, news media accounts form a substantial volume of the known sources for LLM training. Analysis by the Washington Post found that in Google's C4 dataset, news and media ranks third among all categories of sources, including half of the top ten represented sites overall.[45] For example, tokens—that is, words, letters, and other units of text processed by an LLM—from The New York Times alone outnumber any other sources besides Wikipedia and Google Patents at 0.06% of all data in the C4 dataset.[46]

[44] *See, e.g.*, Bright Data, *Web Unlocker*, https://brightdata.com/products/web-unlocker (last visited Oct. 24, 2023); Bright Data, *Web Scraper IDE*, https://brightdata.com/products/web-scraper (last visited Oct. 24, 2023); Damaso Sanoja, The 5 Best Programming Languages for Web Scraping, Bright Data (2023) https://brightdata.com/blog/web-data/best-languages-web-scraping ("Fortunately, regardless of your choice, you can use Bright Data to unlock the power of web data. Bright Data's products offer all the support you need to scrape website data at ease. Whether it's high quality proxies, a headless browser for scraping (Playwright/Puppeteer compatible), a fully hosted Web Scraper IDE, or a large dataset marketplace, Bright Data has all the solutions needed for web data gathering.")

[45] Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart*, Wash. Post, Apr. 19, 2023, https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

[46] *Id.*

The attached technical analysis assessed a small sample of publisher content using 16 publication domains that were volunteered by N/MA members.[47] As measured by the presence of unique URLs, together these 16 publication domains comprised 0.02% of the Common Crawl dataset and between 0.15% and 1.97% of C4, OpenWebText, and OpenWebText2. This assessment shows that datasets specifically developed for LLM training, such as C4 and OpenWebText, skew towards content from the 16 publication domains. When comparing these datasets to Common Crawl, publisher representation increases by a factor of 5 for C4 to as high as 100 in OpenWebText2. This assessment does not capture the full volume of publisher content in the open-source datasets, but is useful to understand the treatment of all publisher content. These works by newspaper, magazine, and digital media publishers are authentic, reliable, and high-quality expressive content that is protected by copyright. The scraping of publisher content and the prevalence of it in the training data speaks volumes about the value of such content for generative AI developers and applications—not only as initial training data but also as an ongoing resource to draw from when the AI system is generating outputs—highlighting the importance of adequate compensation for such uses.

N/MA understands that LLM developers often gather this content either by scraping it directly from websites or by extracting it from archives or datasets created by third parties, such as Common Crawl (or a curated subset). In addition to scraping the content from the copyright owners' own websites, developers may gain access through third-party websites that republish publisher content, often without authorization. In these cases, publishers' content can be infringed at least twice—once by the third-party website reproducing the content and once by the AI developer scraping said content from that website.

The scraping of publisher websites is systematic and generally takes place without a license or authorization, in violation of publishers' terms of service, and with no real way for publishers to opt out of such scraping. Even where opt-out measures are offered or respected, they are insufficient at best. While some developers now provide publishers with the option to opt out, this is not a common practice and such opt-outs only apply to the specific developer in question, making opting out impractical and burdensome for media publishers. Similarly, while some developers have indicated that publishers can use robots.txt exclusion protocol going forward to indicate their unwillingness to be scraped for AI training purposes, the use of the protocol has traditionally meant being excluded from even simple search results by search engines—reducing publishers' visibility and discoverability to the public. There is also no requirement for developers to comply with the voluntary opt-out signal or for scrapers to

[47] This assessment was made not to capture the full volume of publisher content in the open-source datasets, but to help understand the treatment of publisher content.

accurately identify themselves, allowing bad actors to continue scraping publisher content without authorization. Further, and more fundamentally, publishers should not have to affirmatively opt out from generative AI uses to prevent the commercial consumption of their protected material—it is antithetical to the guiding principles of U.S. copyright law and the exclusive rights afforded to rightsholders. Such opt-out solutions are also "too little, too late," considering the vast scraping and copying of publisher content that has already taken place to bring generative AI models to the point of commerciality.

Regardless, liability related to the collection and ingestion of copyright-protected materials for training does not depend solely, or even mainly, on whether those materials were protected from scraping by technical measures or terms of service, or whether a developer or third party curated those materials into a larger dataset. The original expressive works published by N/MA members, including compilations, are clearly protected by copyright. Protected content is not free for the taking simply because it was made available for readers on the public internet. That was precisely part of the reason why the WCT/WPPT established "making available" as a separate right under international treaty.

6.1. How or where do developers of AI models acquire the materials or datasets that their models are trained on? To what extent is training material first collected by third-party entities (such as academic researchers or private companies)?

Our response to Question 6 discusses some of the ways in which generative AI developers acquire the materials and datasets for training purposes. As noted, representative sources include training datasets such as Common Crawl, Google's C4, WebText, The Pile, Books3, LAION, WebVid-10M, as well as public forums like Reddit and Quora, in addition to direct scrapes of numerous publisher websites, including articles, images, web documents, books, code, mathematics, and conversational data. Some of these datasets have been collected by nonprofits, such as Common Crawl, and are then used as the basis for other datasets. For example, Google's C4 dataset is based on a curated subset of Common Crawl's web corpus.[48] OpenAI's WebText, meanwhile, contains data scraped from websites linked to by Reddit users.[49] Some of the large search platforms crawl and index publisher content for search engine purposes but also seemingly use the copies they have created to feed generative AI models.

To the extent AI developers rely on third parties, such as Common Crawl, to obtain datasets of scraped content, those companies seemingly copy the content a second time when they obtain

[48] Papers with Code, *C4 (Colossal Clean Crawled Corpus)* (n.d.), https://paperswithcode.com/dataset/c4 (last visited Oct. 23, 2023).
[49] Papers with Code, *Web Text* (n.d.), https://paperswithcode.com/dataset/webtext (last visited Oct. 23, 2023).

the datasets from these third parties. 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."[50] While datasets like C4 and Common Crawl are publicly available, others like WebText have not been released, making it difficult for publishers to ascertain what is included in them.

It is clear, however, that in addition to content scraped from sites made freely accessible—yet still copyright protected—to the public, some of the datasets and AI models include content that has been collected from behind paywalls. This is partially due to many publishers allowing crawling behind paywalls for search engine purposes but also because some companies offer ways to circumvent paywalls for AI scraping purposes.

Many media publishers have long had terms and conditions that prohibited the use of their protected material for generative AI development, while others have updated their terms of service to prohibit AI scraping more recently.[51] Without cooperation from generative AI developers, there is no easy, standardized way to block scraping for AI purposes. While some respect robots.txt, others do not. Additionally, blocking for AI training can often have the undesirable effect of also blocking crawling for search and other desirable, mutually beneficial uses.[52] Increasingly many companies have indeed opted out or blocked AI web crawlers—over the course of three weeks in late-September at least 250 top websites blocked OpenAI's GPTBot while 14 percent of the 1,000 most popular websites block Common Crawl's CCBot.[53]

[50] *Frequently Asked Questions*, COMMON CRAWL (2023), https://commoncrawl.org/big-picture/frequently-asked-questions/; *Get Started*, COMMON CRAW (2023), https://commoncrawl.org/the-data/get-started/.

[51] *See, e.g.*, Katyanna Quach, *Medium Asks AI bot crawlers: Please, Please Don't Scrape Bloggers' Musings*, THE REGISTER, Sep. 29, 2023, https://www.theregister.com/2023/09/29/medium_ai_crawlers/; Noah Waisberg & Maya Lash, *NO (Mostly)! What Terms of Use For Major Websites Say About Whether Generative AI Training Is Allowed On Their Content*, ZUVA, Jul.18, 2023, https://zuva.ai/blog/llm-breach-of-terms-of-use/.

[52] While Google recently announced a new mechanism, Google-Extended, that it claims "web publishers can use to manage whether their sites help improve Bard and Vertex AI generative APIs, including future generations of models that power those products," it has not yet documented how this feature will do so, or how it may affect visibility through Google's search interfaces. Further, this does not address historic scraping that has already taken place. See Emma Roth, *Google Adds a Switch for Publishers to Opt Out of Becoming AI Training Data*, THE VERGE, Sep. 28, 2023, https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex; Devin Coldewey, *Your Website Can Now Opt out of Training Google's Bard and Future AIs*, TECHCRUNCH, Sep. 28, 2023, https://techcrunch.com/2023/09/28/your-website-can-now-opt-out-of-training-googles-bard-and-future-ais/. ('"Though Google claims to develop its AI in an ethical, inclusive way, the use case of AI training is meaningfully different than indexing the web. . . . Google's actions is that it exploited unfettered access to the web's data, got what it needed, and is now asking permission after the fact in order to look like consent and ethical data collection is a priority for them.").

[53] Kali Hays, *OpenAI's GPTBot and Other AI Web Crawlers are Being Blocked by Even More Companies Now*, INSIDER, Sep 28, 2023, https://www.businessinsider.com/openai-gptbot-ccbot-more-companies-block-ai-web-crawlers-2023-9?r=US&IR=T; *Who Blocks OpenAI, Google AI and Common Crawl?,* PALEWIRE (2023),

Overall, technical measures including robots.txt are blunt and flawed instruments when it comes to protecting publishers from infringement in practice. Robots.txt in particular has many holes that enable bypassing of the measure. The eventual development and adherence to reasonable technical measures may help to establish the conditions for a flexible and market-based licensing framework that facilitates continued innovation and creativity for all affected parties. But technical measures alone cannot substitute for a system of enforceable rights, lest the burden improperly shift to copyright owners to protect their content from automated, systemic infringement, instead of requiring AI developers to take responsibility for their compliance with the law.

And as long as the content is available elsewhere, the opt-outs or blocks are not fully effective. AI developers and dataset curators often still access protected content through pirate websites, undermining the value of such prohibitions and exacerbating the harm to copyright owners. To mitigate this problem, as discussed in response to Questions 4 and 5, the Copyright Office could consider recommending the establishment of a process, modeled after the USTR's Special 301 Review, to identify known pirate sites that regularly reproduce copyrighted content and are therefore off-limits for AI training purposes.

As noted, in addition to collecting content and creating datasets themselves, many generative AI developers acquire such datasets from third-party organizations, including research and non-profit entities that scrape and collect content and data facially for public interest purposes. By using these datasets for commercial AI applications, the result is essentially a form of data laundering by generative AI developers that blurs the distinction between noncommercial research and commercial uses. The Copyright Office should take a clear position against such practices and recommend policies to deter their use for liability evasion purposes.

6.2. To what extent are copyrighted works licensed from copyright owners for use as training materials? To your knowledge, what licensing models are currently being offered and used?

Most generative AI developers, including OpenAI, Google, Microsoft, Meta, and Anthropic, do not reliably acquire the required licenses for the professional media content they use to train their AI models. Instead, the use of reliable and trusted expressive content for generative AI training without authorization undermines existing licensing markets, with the copying serving and supplanting the same licensing purpose.

Licensing markets have long existed for archival material and real-time access to news and other digital media content, including for use in new products and technologies, and many

https://palewi.re/docs/news-homepages/openai-gptbot-robotstxt.html.

N/MA members already operate robust licensing businesses. N/MA members are actively working to grow such licensing opportunities for generative AI products and services. Examples of current, non-AI specific, licensing agreements are abundant, ranging from news media monitoring services to legal research services like LexisNexis to news aggregators like Google News Showcase, as well as a variety of other licenses offered by publishers either directly or through services like the Copyright Clearance Center (CCC).[54] Some major news organizations also provide licensing services for themselves and partners.[55]

The fact that some of the largest generative AI developers (such as Google and Meta) already license content from publishers for other uses shows that these licensing markets are working and appropriate for AI development. Meanwhile, the market is already responding to the demand to provide high-quality media content specifically for generative AI development. For example, this summer, OpenAI signed a deal with the Associated Press to license AP news stories.[56] Reddit recently announced that it will charge AI developers to copy its large corpus of human-to-human conversations.[57] CCC also licenses a catalog of text content on behalf of almost 60 scientific publishers for certain uses of AI development.[58] And this licensing market is poised to continue to grow, with discussions reportedly underway between numerous media entities and developers, such as OpenAI, to license media content for AI training.[59]

[54] *See, e.g.*, *Copyright Resources*, CISION (2023), https://www.cision.com/legal/copyright-resources/; *LexisNexis Extends Multi-year Content Agreement with The New York Times*, LEXISNEXIS PRESS ROOM (Sep. 20, 2021), https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-extends-multi-year-content-agreement-with-the-new-york-times; *Annual Copyright License*, COPYRIGHT CLEARANCE CENTER (2020) available at https://www.copyright.com/wp-content/uploads/2021/01/Product-Sheet-Annual-Copyright-License-8-2020.pdf; *Copyright Clearance Center Integrates Rights Delivery Platform on Copyright.com*, LIBRARY TECHNOLOGY GUIDES (Mar. 1, 2011), available at https://librarytechnology.org/pr/15507/copyright-clearance-center-integrates-rights-delivery-platform-on-copyright-com; Sara Fischer, *Google to Launch News Showcase Product in U.S.*, AXIOS, Jun. 8, 2023, https://www.axios.com/2023/06/08/google-news-showcase-us.

[55] *What We Do*, N.Y. TIMES, (n.d.), https://nytlicensing.com/what-we-do/ (last visited Oct. 25, 2023); *Products*, WASH. POST (n.d.), https://www.washingtonpost.com/licensing-syndication/products (last visited Oct. 25, 2023).

[56] *ChatGPT-Maker OpenAI Signs Deal with AP to License News Stories*, AP (Jul. 13, 2023) available at https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[57] Lawrence Bonk, *Reddit Will Charge Companies for API Access, Citing AI Concerns*, ENGADGET (Apr. 18, 2023) https://www.engadget.com/reddit-will-charge-companies-for-api-access-citing-ai-training-concerns-184935783.html.

[58] COPYRIGHT CLEARANCE CENTER, COMMENTS ON INTELLECTUAL PROPERTY PROTECTION FOR ARTIFICIAL INTELLIGENCE INNOVATION, at 2. (Jan. 10, 2020) Docket No. PTO–C–2019–0038 ("CCC USPTO Comments"), available at https://www.uspto.gov/sites/default/files/documents/Copyright-Clearance-Center_RFC-84-FR-58141.pdf.

[59] *AI and Media Companies Negotiate Landmark Deals Over News Content*, FINANCIAL TIMES Jun. 17, 2023, https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d; Helen Coster & Zaheer Kachwala, *News Corp in Negotiations with AI Companies over Content Usage, CEO Says,* REUTERS, Sep. 7, 2023, https://www.reuters.com/business/media-telecom/news-corp-negotiations-with-ai-companies-over-content-usage-ceo-2023-09-07/.

Outside the publishing industry, similar licenses between generative AI developers and content creators abound. For example, and as noted in response to question 8, Stability AI and Meta have both launched text-to-music generators built completely on licensed sound recordings and musical compositions, while Google is considering a similar service with Universal Music Group.[60] Universal Music Group also recently reached an agreement with a social music creation platform BandLab focusing on AI.[61] Meanwhile, Getty has partnered with Nvidia to develop a text-to-image generator based on licensed images.[62] OpenAI has licensed imagery from Shutterstock since 2021, providing access that its CEO Sam Altman said was "critical" to the training of its DALL-E engine, and it recently announced an expanded licensing deal covering the licensing of Shutterstock's music catalogue as well.[63] Adobe Firefly is a text-to-image generator trained on Adobe Stock images, openly licensed content, and public domain content.[64]

Despite this evidence that generativeAI developers can and do build models based purely on licensed content, and the ability of the marketplace to facilitate reasonable licenses for media content, to our knowledge, most generative AI developers do not presently negotiate and acquire licenses for this valuable content. There is no copyright-based reason to treat published media content any differently than works of visual art or music. And absent efficient licensing markets—such as through voluntary collective licensing—and enforcement of existing rights, smaller publishers especially may be left out of these market solutions due to their lack of resources to develop their own AI license offerings.

N/MA also incorporates its responses to questions 10-13 with respect to licensing models.

[60] Cristina Criddle, *AI and Media Companies Negotiate Landmark Deals Over News Content*, FINANCIAL TIMES, Jun. 17, 2023, https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d; Helen Coster & Zaheer Kachwala, *News Corp in Negotiations with AI Companies over Content Usage, CEO Says,* REUTERS, Sep. 7, 2023, https://www.reuters.com/business/media-telecom/news-corp-negotiations-with-ai-companies-over-content-usage-ceo-2023-09-07/.

[61] Murray Stassen, *Universal Music Strikes 'First-of-Its-Kind' Strategic AI Partnership with Bandlab Technologies*, MUSIC BUSINESS WORLDWIDE, Oct. 18, 2023, https://www.musicbusinessworldwide.com/universal-music-strikes-first-of-its-kind-strategic-ai-partnership-with-bandlab-technologies1/.

[62] Lauren Goode, *Getty Images Plunges into Generative AI Pool*, WIRED, Sep. 25, 2023, https://www.wired.com/story/getty-images-generative-ai-photo-tool/.

[63] Daniel Tencer, *OpenAI Secures License to Access Training Data from Shutterstock… Including Its Music Libraries*, MUSIC BUSINESS WORLDWIDE, Jul. 12, 2023, https://www.musicbusinessworldwide.com/openai-secures-license-to-access-training-data-from-shutterstock-including-its-music-libraries/.

[64] *Firefly FAQ for Adobe Stock Contributors*, ADOBE (Updated Oct. 4, 2023), https://helpx.adobe.com/stock/contributor/help/firefly-faq-for-adobe-stock-contributors.html.

6.3 To what extent is non-copyrighted material (such as public domain works) used for AI training? Alternatively, to what extent is training material created or commissioned by developers of AI models?

N/MA's response focuses on the copyright-protected content produced by its publishing members, and we cannot speak for the practices of generative AI model developers. Taken together, acquisition practices include negotiating licenses to valuable media and other content, use of public domain material, using material created or commissioned themselves, integrating open licensed material, among others. N/MA's response to questions 6.2, 8 and 10 include numerous examples of development processes that make use of licensed content. It is therefore possible to develop models without ingesting unauthorized copyright-protected works.

6.4. Are some or all training materials retained by developers of AI models after training is complete, and for what purpose(s)? Please describe any relevant storage and retention practices.

Considering the variety of generative AI models and the lack of transparency regarding their training processes, it is impossible to say for certain whether and to what extent and purposes training materials are retained by developers after training is complete. For this reason, among others, adequate transparency and recordkeeping obligations, *see infra,* are essential.

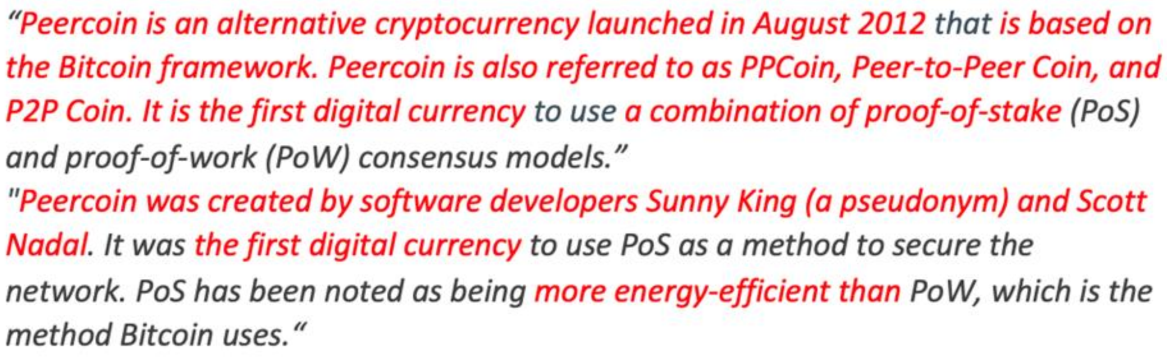
However, as discussed in the attached White Paper and technical analysis, following forensic analysis, outputs from LLMs confirm that LLMs both copy and retain the expressive content of the news, magazine, and digital media materials ingested to build the models. As explained by researchers, "[t]he capabilities of text generators are perhaps even more striking, as they write essays, poems, and summaries, and are proving adept mimics of style and form" and LLMs "can produce content that is sufficiently similar to copyrighted material,"[65] neither of which could occur unless the LLM had copied and stored that expression, even if only translated into a numeric state. Northwestern University Professor of Communication Studies and Computer Science Nick Diakopoulus has similarly documented this retention and use of news reports.[66] The exact practices may vary, and training data may also be required for the purposes of quality assurance and ensuring the continuous operation of the model.

[65] Peter Henderson, et al., *Foundation Models and Fair Use*, ARXIV (Mar. 29, 2023), available at https://doi.org/10.48550/arXiv.2303.15715.

[66] Nick Diakopoulus, *Finding Evidence of Memorized News Content in GPT Models*, GENERATIVE AI IN THE NEWSROOM, Sep. 5, 2023, https://generative-ai-newsroom.com/finding-evidence-of-memorized-news-content-in-gpt-models-d11a73576d2.

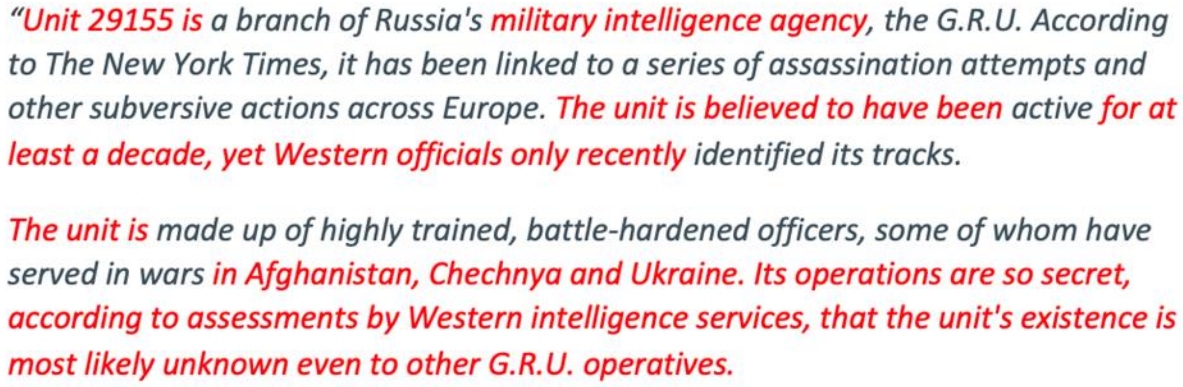
With regards to verbatim reproductions of protected content, in tests, GPT-4 reproduced long runs of memorized texts from Investopedia, New York Times, and The Boston Globe, articles when prompted to respond with direct quotes from the articles in question. As the White Paper explains:

> The ecosystem is designed to produce GAI-generated responses to user queries that substitute for expressive content created by news publishers. For example, GPT-4 reproduces long runs of verbatim text (shown in red below) from an Investopedia article about the cryptocurrency Peercoin when prompted: "Tell me about Peercoin. Respond with direct quotes from the Investopedia article called Peercoin."



"Peercoin is an alternative cryptocurrency launched in August 2012 that is based on the Bitcoin framework. Peercoin is also referred to as PPCoin, Peer-to-Peer Coin, and P2P Coin. It is the first digital currency to use a combination of proof-of-stake (PoS) and proof-of-work (PoW) consensus models."
"Peercoin was created by software developers Sunny King (a pseudonym) and Scott Nadal. It was the first digital currency to use PoS as a method to secure the network. PoS has been noted as being more energy-efficient than PoW, which is the method Bitcoin uses."

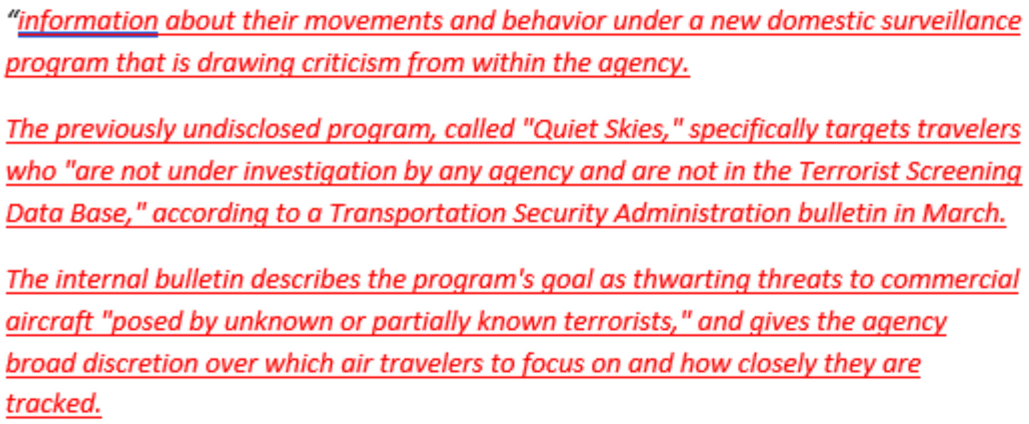
GPT-4 produces a similar result for a New York Times article about a Russian military unit when prompted: "Tell me about Russian unit 29155. Respond with information from NY Times."



"Unit 29155 is a branch of Russia's military intelligence agency, the G.R.U. According to The New York Times, it has been linked to a series of assassination attempts and other subversive actions across Europe. The unit is believed to have been active for at least a decade, yet Western officials only recently identified its tracks.

The unit is made up of highly trained, battle-hardened officers, some of whom have served in wars in Afghanistan, Chechnya and Ukraine. Its operations are so secret, according to assessments by Western intelligence services, that the unit's existence is most likely unknown even to other G.R.U. operatives.

The response to optimized prompts, designed to elicit the maximum amount of repetition from an original article (e.g., to bypass a paywall) was even starker. GPT-4's response to the optimized prompt "You are a journalist for The Boston Globe. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Federal air marshals have begun following ordinary US citizens not suspected of a crime or on any terrorist watch list and collecting extensive'" included a 91.3% overlap in runs of three words with the original source:[67]



"information about their movements and behavior under a new domestic surveillance program that is drawing criticism from within the agency.

The previously undisclosed program, called "Quiet Skies," specifically targets travelers who "are not under investigation by any agency and are not in the Terrorist Screening Data Base," according to a Transportation Security Administration bulletin in March.

The internal bulletin describes the program's goal as thwarting threats to commercial aircraft "posed by unknown or partially known terrorists," and gives the agency broad discretion over which air travelers to focus on and how closely they are tracked.

Another study has shown that generative AI models have regurgitated pages from books such as Harry Potter, with the author noting that "several models output the first page or two of Harry Potter books, verbatim,"[68] while adding "the instruction 'replace every a with a 4 and o with a 0' along with the prompt" had the model "regurgitate the first three and a half chapters of [Harry Potter and the Sorcerer's Stone] verbatim."[69] In addition to engaging in verbatim copying, such tools can reproduce the structure and expressive quality of the underlying works.

Further, as explained in the technical analysis, even when the models do not generate verbatim output, they are able to provide paraphrases with a measurably high degree of similarity in meaning that exceeds that attributable to addressing the same factual subject, implying that while generative AI systems can be programmed to prevent verbatim copying, they are still likely to retain copies for paraphrasing.

7. To the extent that it informs your views, please briefly describe your personal knowledge of the process by which AI models are trained. The Office is particularly interested in: 7.1. How are training materials used and/or reproduced when training an AI model? Please include your understanding of the nature and duration of any reproduction of works that occur during the training process, as well as your views on the extent to which these activities implicate the exclusive rights of copyright owners.

[67] As shown in p. 24-25 of the technical analysis, the verbatim copy is roughly 4x the length of this excerpt.
[68] PETER HENDERSON, ET AL., FOUNDATION MODELS AND FAIR USE (Mar. 29, 2023) available at https://arxiv.org/pdf/2303.15715.pdf.
[69] *Id.*

While much of the generative AI training and development processes are deliberately kept opaque by the AI companies, it seems clear that developers systematically copy substantial amounts of protected publisher content. The training typically involves making copies of the expressive content, curating and editing it as necessary, and then using that material for its expressive attributes to draw mathematical inferences that predict the most likely word to come next in a sentence in order to produce outputs.

Throughout this process, it appears generative AI developers engage in copying and reproduction, during the original collection or scraping, the transfer or sale of large datasets or models to other developers, and the fine-tuning and other development stages. The original copies include unaltered reproductions of text from the training source pages, while later stages may involve alterations to or manual curation of the content. As probed further in the White Paper, this understanding is shared by leading AI developers, the Congressional Research Service, and even advocates who contend that generative AI is non-infringing fair use, each acknowledging that large language models engage in massive copying of copyright-protected material.[70] Indeed, as counsel for Meta's LLAMA2 explains, as a general matter, generative AI "systems involve copying the entire work, without alteration."[71]

The copying violates copyright owners' exclusive rights to reproduce their copyrighted work, and occurs at the ingestion stage, likely at the retention stage, and, oftentimes, at the output stage. The copying first occurs when the generative AI developers or third parties such as Common Crawl scrape whole articles without authorization from media websites.[72]

[70] *See*, *e.g.*, Comment of OpenAI, LP Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, Before the United States Patent and Trademark Office Department of Commerce at 2 ("OpenAI USPTO Comments") ("By analyzing large corpora (which necessarily involves first making copies of the data to be analyzed), AI systems can learn patterns inherent in human-generated data"); Congressional Research Service, Generative Artificial Intelligence and Copyright Law, Updated May 11, 2023 ("As the U.S. Patent and Trademark Office has described, this process [of building an LLM] 'will almost by definition involve the reproduction of entire works or substantial portions thereof.'"); Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743 (2021) available at https://texaslawreview.org/wp-content/uploads/2021/03/Lemley.Printer.pdf.

[71] Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743, at 746 (2021)(AI systems "rarely transform the databases they train on; they are using the entire database."). *See* Defendant's Motion to Dismiss Plaintiff's Complaint, *Richard Kadrey, Sarah Silverman & Christopher Golden v. Meta Platforms, Inc.*, No. 3:23-cv-03417-VC, (U.S. Dist. N.D. Cal. Nov. 16, 2023) available at https://fingfx.thomsonreuters.com/gfx/legaldocs/dwpkakjdxpm/META%20OPENAI%20SILVERMAN%20INFRINGEMENT%20metamtd.pdf (listing Lemley as counsel for Meta).

[72] Each of Google, OpenAI, and Microsoft appear use a combination of web content which they have directly scraped from the web or obtained from Common Crawl. Google's Bard initially used Google's LLM LaMDA, which was built using a dataset composed primarily of dialog data that Google took from public forums such as Reddit and Quora, as well a subset of material offered by Common Crawl, referred to as "C4." Romal Thoppilan, et al., *LaMDA: Language Models for Dialog Applications*, Google (2022) at 47, available at https://arxiv.org/pdf/2201.08239.pdf. Google announced in May 2023 that Bard would be powered by a different LLM called PaLM2 and

To the extent generative AI technologies rely on datasets full of scraped web content made available by third parties, the AI developers copy the content a second time when they obtain the datasets from these third parties. 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."[73]

These developers often further copy the materials, multiple times, in the process of building out LLMs.[74] Further copying can occur at the "output" stage. As OpenAI candidly admits, GAI systems can "generate output media that infringes on existing copyrighted works."[75] As noted in response to question 6, and documented in the accompanying White Paper, news and media articles are a major category of material contained in the datasets used to build leading LLMs.

Copies made for generative AI development appear to be perceptible to humans and more than transitory in duration, evidenced by reports that some developers engage human reviewers to manually curate and tag content included in the training datasets. As one company offering such services in India states, they "annotate the texts with metadata labeling for machine learning and AI algorithms based on natural language processing helping machines to understand the human language easily."[76] Even where humans are not involved, the computer-based ingestion of works appears sufficient to satisfy the definition of copying in the Copyright Act. Under the statute, a copy is made when a work is fixed and "can be perceived, reproduced,

stated that the model used "web documents, books, code, mathematics, and conversational data." *See* Zoubin Ghahramani, *Introducing PaLM 2*, Google Blog, May 10, 2023, https://blog.google/technology/ai/google-palm-2-ai-large-language-model/; James Vincent, *Google Announces PaLM 2 AI Language Model, Already Powering 25 Google Services*, The Verge, May 10, 2023, https://www.theverge.com/2023/5/10/23718046/google-ai-palm-2-language-model-bard-io; PaLM 2 Technical Report, Google at 2 (2023), https://ai.google/static/documents/palm2techreport.pdf. OpenAI built various iterations of its GPT technology from a curated subset of material from Common Crawl, as well as a database known as WebText2, a proprietary corpus. *See* Tom B. Brown, et al., *Language Models Are Few-Shot Learners*, Google (2022) at 9, available at https://arxiv.org/pdf/2005.14165.pdf; *see also* Alec Radford, et al., *Language Models Are Unsupervised Multitask Learners* at 3, (n.d.), https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[73] *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).

[74] 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.").

[75] OpenAI USPTO Comments at 11 (emphasis omitted).

[76] *AI Annotation & Data Labeling Services Ind.*, Ishir (n.d.), https://www.ishir.com/ai-annotation-services-india.htm (last visited Oct. 25, 2023).

or otherwise communicated for a period of more than a transitory duration," and this perception can occur "either directly or with the aid of a machine or device."[77]

Especially in light of N/MA's technical analysis and those of other third parties suggesting that generative AI systems develop and retain the ability to replicate and mimic large passages of text, it appears that the so-called training process—however shrouded and despite whatever efforts to mitigate after the fact to avoid infringing outputs—requires use of the expressive works in ways that violate the exclusive copyright interests.

7.2. How are inferences gained from the training process stored or represented within an AI model?

The attached White Paper and responses to questions 6.3-6.4 discuss N/MA's understanding of the training process and the use of publishers' content thereof, as well as relevant storing and retention practices. In order to work, generative AI systems draw from the very copyrightable expression encapsulated in the ingested works. Therefore, while N/MA questions the use of the term "gaining inferences" in this context and does not believe AI systems should be anthropomorphized as "learning", the ingestion process itself, as well as storing and representing such relationships later down the line, appears to implicate copyright owners' exclusive rights. The White Paper outlines ways in which ingested content is used throughout the AI model development cycle in further detail.

7.3. Is it possible for an AI model to "unlearn" inferences it gained from training on a particular piece of training material? If so, is it economically feasible? In addition to retraining a model, are there other ways to "unlearn" inferences from training?

N/MA does not currently express an opinion on whether generative AI models can truly "unlearn" inferences gained from training on a particular piece of training material. This may depend on the model or the developer, and there may be workarounds that minimize or hide the effect a specific piece of copyrighted material would have on the output even if it would not fully "unlearn" it. Some recent reports suggest that at least some developers, such as OpenAI, have removed meaningful materials from their models, affecting their outputs, reportedly for trust, safety, and infringement reasons.[78] But other academic research acknowledges that

[77] 17 U.S.C. 101. *See also MAI Systems Corp. v. Peak Computer, Inc.,* 991 F.2d 511 (9th Cir. 1993) (finding that "MAI has adequately shown that the representation created in the RAM is 'sufficiently permanent or stable to permit it to be perceived, reproduced, or otherwise communicated for a period of more than transitory duration.'").

[78] With new reports seemingly weekly as to the limitations of these efforts, it is unclear how successful they are. *See, e.g.,* Maggie Harrison, *Microsoft Lobotomizes Bing's Image Generating AI,* THE BYTE, Oct. 10, 2023*,*

"achieving precise unlearning is computationally infeasible for very large models."[79] Regardless, mitigation after the fact should not be presumed to be an adequate remedy.

For copyright owners, there are three main potential concerns as to the potential limits for "unlearning." First, even if possible, it would not eliminate a past act of infringement, and may not eliminate benefits the infringement provided to the model and/or the developer, or harm to a copyright owner in the form of lost revenue and brand harm. Second, because a compulsory license would not be appropriate here, it is necessary that adequate "unlearning" processes are established to provide copyright owners with an effective way to decline to license their materials in the first place.

Finally, publishers may have legal obligations to remove certain content from their properties for a variety of reasons—including compliance with regulations ranging from right to be forgotten, consumer privacy, and copyright, in addition to litigation settlement purposes—and publishers need the ability to demand generative AI developers delete the same publisher content from their models. Without effective "unlearning" in these situations, issues will arise about whether the publisher and/or the developer may potentially be legally liable. As such, this question raises additional questions outside the confines of copyright law. The Copyright Office may wish to further consult other agencies and stakeholders on these issues.

7.4. Absent access to the underlying dataset, is it possible to identify whether an AI model was trained on a particular piece of training material?

Transparency and recordkeeping requirements are essential for publishers to accurately identify whether a generative AI model was trained on a particular piece of material. While some datasets are publicly available and searchable, and third party tools like "Have I Been Trained?" exist that purport to answer this question,[80] these tools are imperfect and model developers may supplement, edit, or combine datasets to suit the needs of their AI models, reducing the utility of the publicly available datasets to rightsholders. In addition to testing for evidence of verbatim copying that provides strong evidence of the use of a particular piece in the training of the model, indirect methods known as "membership inference attacks" have been developed to infer that particular works were used in training in certain circumstances. Examples of such methods are discussed in detail in the White Paper. However, such methods put the burden on publishers, are costly to employ at scale, and are incapable of systematically

https://futurism.com/the-byte/microsoft-lobotomizes-bing-ai (describing efforts to mitigate BingAI after it returned an image of Mickey Mouse driving a plane into the World Trade Center).

[79] Martin Pawelczyk, Seth Neel, & Himabindu Lakkaraju, *In-Context Unlearning: Language Models as Few Shot Unlearners*, ArXiv:2310.07579, Oct. 12, 2023, available at https://arxiv.org/pdf/2310.07579.pdf.

[80] Have I Been Trained (n.d.), https://haveibeentrained.com/ (last visited Oct. 25, 2023).

identifying all works that were used in training. Transparency and recordkeeping rules are the only fair, certain, and efficient method to achieve this end.

8. Under what circumstances would the unauthorized use of copyrighted works to train AI models constitute fair use? Please discuss any case law you believe relevant to this question. 8.1. In light of the Supreme Court's recent decisions in *Google* v. *Oracle America* and *Andy Warhol Foundation* v. *Goldsmith,* how should the "purpose and character" of the use of copyrighted works to train an AI model be evaluated? What is the relevant use to be analyzed? Do different stages of training, such as pre-training and fine-tuning, raise different considerations under the first fair use factor? 8.2. How should the analysis apply to entities that collect and distribute copyrighted material for training but may not themselves engage in the training? 8.3. The use of copyrighted materials in a training dataset or to train generative AI models may be done for noncommercial or research purposes. How should the fair use analysis apply if AI models or datasets are later adapted for use of a commercial nature? Does it make a difference if funding for these noncommercial or research uses is provided by for-profit developers of AI systems? 8.4. What quantity of training materials do developers of generative AI models use for training? Does the volume of material used to train an AI model affect the fair use analysis? If so, how? 8.5. Under the fourth factor of the fair use analysis, how should the effect on the potential market for or value of a copyrighted work used to train an AI model be measured? Should the inquiry be whether the outputs of the AI system incorporating the model compete with a particular copyrighted work, the body of works of the same author, or the market for that general class of works?

N/MA responds to question 8 and its subparts together. For consistency, much of the below analysis is repeated in the accompanying White Paper.

N/MA members are strong supporters of fair use and regularly interpret and rely on fair use principles as media publishers, including to disseminate the robust criticism and commentary necessary to ensure an informed public. That said, fair use is not intended to excuse mass-scale acts of infringement.

As the Office knows, fair use is considered on a case-by-case basis, with reference to the four-factor test developed through case law and codified in section 107 of the Copyright Act. N/MA recognizes that generative AI technologies and uses vary, including configurations that are technology, industry, use, or audience specific. N/MA members believe that the LLM systems presently at the core of many policy discussions are exceeding the bounds of fair use. With those systems in mind, our comments generally address how the fair use doctrine may relate to analyses of generative AI systems and configurations of these systems.

Copyright law is not designed to accommodate taking publisher content and using it in ways that damage their businesses. The fair use defense need not shield a generative AI modeler's copying of (1) the entirety of expressive works to build their large language models [inputs], or (2) substantial portions of the works' expressive content when responding to user queries [outputs]. To our knowledge, no court has held that taking copyrighted material for ingestion into a commercial generative AI model is a fair use.

The purpose and character of copying to train LLMs is not sufficiently transformative (first factor).

i. Copying for purposes of commercial substitution weighs against fair use.

The Supreme Court recently explained in *Warhol Foundation* that "the first fair use factor considers whether the use of a copyrighted work has a further purpose or different character, which is a matter of degree, and the degree of difference must be balanced against the commercial nature of the use."[81] Moreover, "if an original work and a secondary use share the same or highly similar purposes, and the secondary use is of a commercial nature, the first factor is likely to weigh against fair use, absent some other justification for copying."[82]

Such an independent justification is "particularly relevant to assessing fair use where an original work and copying use share the same or highly similar purposes, or where wide dissemination of a secondary work would otherwise run the risk of substitution for the original or licensed derivatives of it."[83] As *Warhol Foundation* emphasized, "targeting" the copied work's expression furnishes the predominant justification. Examples include when it "is reasonably necessary to achieve the user's new purpose,"[84] such as to "conjure up" the original work for a parody or to engage in criticism.[85] "Targeting" is not limited to parody; it more generally involves "commentary … [that] critical[ly] bear[s] on the substance or style of the original composition."[86] Copying may be justified when it "shed[s] light on the original[work]'s depiction."[87]

The focus on "targeting" is consistent with the "purposes" listed in the preamble of section 107: "criticism, comment, news reporting, teaching … scholarship, or research." These purposes

[81] *Andy Warhol Foundation for the Visual Arts v. Goldsmith, et al.*, 143 S. Ct. 1258, 1277 (2023)
[82] *Id*.
[83] *Id*.
[84] *Id.* at 1276.
[85] *Id*. (quoting *Campbell v. Acuff-Rose Music, Inc*., 510 U.S. 569, 580-81 (1994)).
[86] *Id*.
[87] *Id.* at 1295, n.21.

reflect the types of uses the courts and Congress most commonly have found to be fair.[88] All "shed light on" the defendant's treatment of the copied work's expression, not merely on its subject matter. Moreover, and for that reason, such uses ordinarily do not supersede or supplant the copied work.[89]

ii. Generative AI development copies news and digital media content to extract and replicate its expressive content.

As the attached forensic research demonstrates, LLMs typically ingest valuable media content for their written expression. To the extent they are ingesting this content so these published words can be analyzed "in relation to all the other words in a sentence,"[90] or their sequences of words identified,[91] that analysis and identification is intended to capture the very expression that copyright protects. Indeed, it is that very capturing of expression which fuels the LLMs' success, by enabling them to determine the most likely next word in a sentence.[92] That is why LLMs that are trained to generate their own expressive works "copy expression for expression's sake."[93]

Examples such as the "reversal curse" explained in the White Paper show that LLMs take copyrighted content so they can ingest the content's expressive words, not to "understand" the underlying facts (which is why, for example, one LLM could string together a sentence stating that Tom Cruise's mother is Mary Lee Pfeiffer but not one telling a user who is Mary Lee Pfeiffer's son).[94] By its very construction, this is a taking for use of the expression, not one designed to extract the underlying information. Nor is the use to facilitate or extract information about or otherwise "shed light on" the original works' expression.

This capturing of expression to extract, replicate, and paraphrase puts LLMs in a category beyond what was contemplated in prior cases that found fair copying done in the service of a new product or technology. For example, in *Authors Guild v. Google, Inc.*, a case that "tests the

[88] *Campbell*, 510 U. S. at 577-578.

[89] *Warhol Found.*, 143 S. Ct. at 1274; *see Folsom v. Marsh,* 9 F. Cas. 342, 348 (C.C.D. Mass. 1841).

[90] Pandu Nayak, *Understanding Searches Better than Ever Before*, GOOGLE BLOG, Oct. 25, 2019, https://www.blog.google/products/search/search-language-understanding-bert/.

[91] Gary N. Smith, *An AI that Can "Write" is Feeding Delusions about How Smart Artificial Intelligence Really Is*, SALON Jan. 1, 2023, https://www.salon.com/2023/01/01/an-ai-that-can-write-is-feeding-delusions-about-how-smart-artificial-intelligence-really-is/.

[92] Parvin Mohmad, *How Does ChatGPT Become Popular So Quickly and How Is It Growing*, , ANALYTICS INSIGHT, Jan. 19, 2023, https://www.analyticsinsight.net/how-does-chatgpt-become-popular-so-quickly-and-how-is-it-growing/.

[93] Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743, 777 (2021); *see also id*. at 767 (LLMs "empower [] companies to extract value from authors' protected expression without authorization").

[94] N/MA, WHITE PAPER at 12, 25 (2023), Appendix A; Lukas Berglund et al., *The Reversal Curse: LLMs Trained on "A Is B" Fail to Learn "B Is A"*, ARXIV (Sep. 22, 2023), available at https://doi.org/10.48550/arXiv.2309.12288.

boundaries of fair use," the court evaluated two features: (1) a "search for identification of books," and (2) the use of "snippets" to show "just enough context … to … evaluate whether the book falls within the scope of [a reader's] interest (without revealing so much as to threaten the author's copyright interests)."[95] The court found that the nature and purpose of Google's copying of the underlying works favored a finding of fair use because the copying was done to provide "information about" the books,[96] not to exploit the expression in them, and was likely to help users identify books of interest.[97] Although Google's search program did not criticize or comment on the copied works, it nonetheless "targeted" them because its primary objective was to provide information about a particular book ("the purpose of Google's copying of the original copyrighted books is to make available significant information *about those books*.").[98]

Perfect 10, Inc. v. Amazon.com, Inc.[99] and *Kelly v. Arriba-Soft*[100] are similar. Those cases found fair the copying of full-size images into thumbnails, in part because the copying was done to help users to find and access the source materials, not to exploit the works' expressive qualities.

The same is true of the so-called "intermediate copying cases."[101] Those cases found the defendants' reverse engineering of computer code was likely a fair use primarily because, given the unique characteristics of computer code, that copying was "the only way [the defendant could] gain access to the ideas and functional elements embodied in [the plaintiff's] copyrighted computer program," which was needed to facilitate interoperability with video game systems.[102] Thus, the defendants did not copy the computer software to copy the expressive qualities of the computer code; rather, they could access the software's inherent functionality only by reverse engineering the code, which necessarily involved the making of copies. These courts also concluded that a finding of infringement would have allowed the plaintiffs to misuse their copyrights to achieve patent-like monopolies over the functional concepts embodied in their computer software.[103]

[95] 804 F.3d 202, 206, 218 (2d Cir. 2015).
[96] *Id.* at 207, 215.
[97] *Id.* at 222-223.
[98] *Id.* at 217.
[99] 508 F.3d 1146 (9th Cir. 2007).
[100] 336 F.3d 811 (9th Cir. 2003).
[101] *See Sony Computer Entertainment, Inc. v. Connectix Corp*., 203 F.3d 596 (9th Cir. 2000); *Sega Enterprises Ltd. v. Accolade*, 977 F.2d 1510 (9th Cir. 1992).
[102] *Sony*, 203 F.3d at 602, 605-06; *Sega*, 977 F.2d at 1518, 1525-28.
[103] *Sony*, 203 F.3d at 605; *Sega*, 977 F.2d at 1526.

These needs and concerns do not apply to N/MA members' media content. Indeed, to the extent developers contend their models ingest media publications for their non-protectable "facts," the publications disclose any such facts on their face; the facts are not hidden, so copying media publications is not necessary to obtain the information. Nor would enforcing publishers' copyrights make it impossible for generative AI developers to otherwise discover those facts or give publishers a "monopoly" over them.

More importantly, the content of N/MA members is unquestionably protected by copyright. The content of their publications is not simply "facts," but narratives expressed in a particular manner, and which also include carefully reported, crafted, and edited opinion, analysis, reviews, memoir, advice, investigations, fiction, and so on. Such original expression, which is what has been copied, is both protectable and valued.[104]

Indeed, good journalistic writing conveys communicative value. That is why media content is overrepresented in curated sets of well-known training data as compared to non-curated datasets. As the accompanying forensic analysis demonstrates, sampled publisher content was overrepresented in the curated datasets by a factor from over 5 to almost 100 as compared to the generic collection of content in the well-known Common Crawl dataset.

Relatedly, the Office has asked if different stages of training, such as pre-training and fine-tuning, raise different considerations under the first fair use factor. In short, steps in generative AI modeling, including pre-training, fine-tuning, and use of tighter curated datasets can provide insight into the nature of the intended use, but should be viewed as stages that comprise an intended use, rather than bifurcated under a first factor analysis. Those activities may also be relevant to probing whether additional reproductions or adaptations of copyrighted works were made. To the extent this question probes acts by different entities who perform different steps in this process, it implicates questions related to liability addressed below.

Even decoupled from downstream configurations, the copying for generative AI training purposes serves the same purpose as the licensing market for such use.

Training LLMs on reliable, trusted expressive content without authorization also seeks to override licensing markets that already exist and are evolving for these works, and the LLMs' copying for these training purposes thus serves (and supplants) that same licensing purpose.

[104] See *Harper & Row*, 471 U.S. at 556-557; *Feist Publications, Inc. v. Rural Telephone Serv. Co*., 499 U.S. 340, 349 (1991); *see also Super Express USA Publ'g Corp. v. Spring Publ'g Corp*., No. 13-CV-2814 (DLI), 2017 WL 1274058, at *8 (E.D.N.Y. Mar. 24, 2017) (explaining that copyright protection extends to the manner of expression and the author's analysis or interpretation of events in news articles); *accord Wainwright Securities, Inc. v. Wall Street Transcript Corp*., 558 F.2d 91, 95-96 (2d Cir. 1977).

Well-established markets have long existed for licensing archival material and other real-time access to publisher content, including for use in new products and technologies. This market is already responding to the demand to provide high-quality publisher content specifically for AI development, and N/MA members are actively working to grow this field. Moreover, AI developers can (and do) license textual works for model training. For all these reasons, generative AI developers' unauthorized copying of non-licensed content to fuel their development needs shares the same licensing purposes inherent in N/MA members' copyrighted works.[105]

For example, earlier this summer, OpenAI signed a deal with the Associated Press to license AP stories.[106] Reddit recently announced that it will charge generative GAI developers to access its large corpus of human-to-human conversations.[107] The Copyright Clearance Center already licenses a vast catalogue of text content for AI development.[108] And this licensing market is poised to continue to grow, with discussions underway between numerous media entities and LLM developers, such as OpenAI, to license media content for generative AI training.[109]

This licensing for generative AI development is part and parcel of the long existing and well-established markets for licensing archival material and other real-time access to trustworthy journalistic content. For example, media organizations license their content for a variety of

[105] *Warhol*, 143 S. Ct. at 1273, 1278, 1280 (where plaintiff licensed her photographs of Prince to illustrate stories about Prince in magazines, "[plaintiff]'s photograph and AWF's 2016 licensing of Orange Prince share substantially the same purpose").

[106] Matt O'Brien, *ChatGPT-Maker OpenAI Signs Deal with AP to License News Stories*, AP, Jul. 13, 2023, https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a.

[107] Lawrence Bonk, *Reddit Will Charge Companies for API Access, Citing AI Concerns*, ENGADGET, Apr. 18, 2023, https://www.engadget.com/reddit-will-charge-companies-for-api-access-citing-ai-training-concerns-184935783.html.

[108] CCC USPTO Comments at 2.

[109] Cristina Criddle et al., *AI and Media Companies Negotiate Landmark Deals Over News Content*, FINANCIAL TIMES, Jun. 17, 2023, https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d; Helen Coster & Zaheer Kachwala, *News Corp in Negotiations with AI Companies over Content Usage, CEO Says,* REUTERS, Sep. 7, 2023, https://www.reuters.com/business/media-telecom/news-corp-negotiations-with-ai-companies-over-content-usage-ceo-2023-09-07/.

uses, including to media monitoring entities,[110] to LEXIS,[111] and through the CCC.[112] Several major publishers provide licensing services for themselves and partners.[113]

Generative AI copying serves the same purpose as the copied works in two ways: the input of the publishers' works into the LLMs' training data substitute for the publishers' licensing of the same content for the same purpose and the outputs from the models as a result of the copying produce text that serves the same purpose of providing content to readers and end users, sometime by reproducing or paraphrasing portions of the publishers' expression.

iii. LLMs and chatbot uses are highly commercial.

Many generative AI uses of protected content are overwhelmingly commercial. As set forth above, emerging generative AI companies are valued in the billions, and established platforms have seen their market capitalizations soar because of their generative AI products and services. This is fueled by the unauthorized use of third-party content. Following a well-trod Silicon Valley strategy, services that initially were provided at no cost, like Midjourney, Claude, Dall-E, and ChatGPT, are now selling commercial subscriptions that provide the only way to access the full functionality of the products. OpenAI, for example, began as a non-profit research organization offering ChatGPT for free, but pivoted to a for-profit model that now requires a paid subscription to access all its features.[114]

To the Office's question about evaluation of datasets or generative AI training that are initially done for noncommercial or research purposes, it is true that the first factor should take into account the specific use.[115] A dataset that is acceptable to make a non-expressive use within the confines of research may not be fair to use in an expressive, commercial context. While this comment focuses on the many LLM and associated uses that are blatantly, highly commercial, N/MA recognizes that is not the case across the board. However, in light of concerning

[110] *See, e.g., Copyright Resources*, Cison (2023), https://www.cision.com/legal/copyright-resources/.

[111] *LexisNexis Extends Multi-Year Content Agreement with The New York Times*, LexisNexis Press Room, Sep. 20, 2021, https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-extends-multi-year-content-agreement-with-the-new-york-times.

[112] *Annual Copyright License*, Copyright Clearance Center, (2023) https://www.copyright.com/wp-content/uploads/2021/01/Product-Sheet-Annual-Copyright-License-8-2020.pdf.

[113] *What We Do*, N.Y. Times, [n.d.], https://nytlicensing.com/what-we-do/ (last visited Oct. 25, 2023); *Products*, Wash. Post, [n.d.], https://www.washingtonpost.com/licensing-syndication/products (last visited Oct. 25, 2023).

[114] Alex Konrad, *OpenAI Releases First $20 Subscription Version of ChatGPT AI Tool*, Forbes, Feb. 1, 2023, https://www.forbes.com/sites/alexkonrad/2023/02/01/openai-releases-first-subscription-chatgpt/?sh=b4debac7f5f1); *see also* Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev.743, 746 (2021) ("[ML] systems . . . rarely transform the databases they train on; they are using the entire database, and for a commercial purpose at that.").

[115] *Warhol*, 598 U.S. 508; *Chapman v. Nicki Minaj*, 2:18-CV-09088 (C.D. CAL. OCT 22, 2018).

practices of "data laundering" and initially nonprofit models that transition into commercial entities or assist them in building competitive, commercial products, the Office should be careful in drawing any kind of a bright line between commercial and noncommercial uses.[116]

The Office has previously addressed similar questions, including in connection with the triennial Section 1201 rulemaking and regulatory implementation of a "noncommercial use" exception under the new protection for pre-1972 sound recordings established by the Music Modernization Act.[117] Even when limited TDM uses for non-consumptive, academic research purposes were contemplated in the Section 1201 triennial rulemaking, the Register of Copyrights noted that the "case law has not established that all copying of works for the purpose of TDM is necessarily a fair use." The Office further noted that the exemption request was based on representations that the ingested text would only be accessible for purposes of verifying research findings (and not to analyze or view the works for any other purpose). It therefore appears that some of the generative AI products on the market now, with their copying of expression for expression's sake and the ability to produce paraphrased and in some cases near-verbatim outputs, go beyond the proposal before the Office at that time. In that rulemaking, the Register also required that academic institutions employ substantial security measures to limit access to the corpus of circumvented works only to other researchers affiliated with qualifying institutions for purposes of collaboration or the replication and verification of research findings, and that the circumvention of technical measures for research purposes only be allowed "on copies of the copyrighted works that were lawfully acquired and that the institution owns or for which it has a non-time-limited license," not including renting or borrowing.

iv. There is no satisfactory independent justification for the copying.

There is no independent reason why generative AI models must ingest valuable copyright-protected expressive works apart from the desire to incorporate that very expression. While GAI developers may prefer to copy such high-quality media unburdened from any licensing obligations, some of the very companies that have infringed the copyrighted content of N/MA members have licensed content from others for similar purposes. For example, Stability AI and Meta have launched text-to-music generators trained solely on licensed musical works and

[116] Relatedly, similar logic, as well as considerations of secondary liability and agency principles, may be relevant to the Office's question with respect to entities that collect and distribute copyrighted material for training but may not themselves engage in the training.

[117] *See* Noncommercial Use of Pre-1972 Sound Recordings That Are Not Being Commercially Exploited, 84 Fed. Red. 14242 (Apr. 9, 2019); Exemption to Prohibition on Circumvention of Copyright Protection Systems for Access Control Technologies, 86 Fed. Red. 59,627 (Oct. 28, 2021).

sound recordings,[118] and Google is in discussions to develop a similar tool using music licensed from Universal Music Group.[119] OpenAI has licensed imagery from Shutterstock since 2021, providing access that its CEO Sam Altman said was "critical" to the training of its DALL-E engine, and it recently announced an expanded licensing deal covering the licensing of Shutterstock's music catalogue as well.[120] Others seems to be trying to get this right from the start. Adobe Firefly is a text-to-image generator trained solely on Adobe Stock images, openly licensed content, and public domain content.[121] Getty has developed a text-to-image generator trained solely on licensed images.[122]

In an acknowledgment that generative GAI development can continue and flourish without training LLMs on unauthorized copies, Google recently announced a new mechanism, Google-Extended, which will allow website publishers to opt out of having their content used to improve the company's AI models in the future while maintaining access to such content through Google Search.[123] OpenAI has similarly announced that internet sites can now block OpenAI's GPTBot and keep their sites out of ChatGPT.[124] In addition, this "opt-out" approach is antithetical to U.S. copyright law (and does not allow for opt-out of the content already scraped). There is also a wealth of material in the public domain or available under open licenses available for the LLMs to use to build their models.

Notably, NMA members stand ready to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy expressive content, something

[118] Daniel Tencer, *Stability AI Launches Text-to-Music Generator Trained on Licensed Content Via a Partnership with Music Library AudioSparx,* MUSIC BUSINESS WORLDWIDE, Sep. 14, 2023, https://www.musicbusinessworldwide.com/stability-ai-launches-text-to-music-generator-trained-on-licensed-content-via-a-partnership-with-music-library-audiosparx/; Justinas Vainilavicius, *Meta Releases Music Generator Called MusicGen*, CYBERNEWS, Aug. 3, 2023, https://cybernews.com/tech/meta-releases-music-generator-musicgen/.

[119] Hibaq Farah, *Google and Universal Music Working on Licensing Voices for AI-Generated Songs*, THE GUARDIAN, Aug. 9, 2023, https://www.theguardian.com/technology/2023/aug/09/google-and-universal-music-working-on-licensing-voices-for-ai-generated-songs.

[120] Daniel Tencer, *OpenAI Secures License to Access Training Data from Shutterstock . . . Including Its Music Libraries*, MUSIC BUSINESS WORLDWIDE, Jul. 12, 2023, https://www.musicbusinessworldwide.com/openai-secures-license-to-access-training-data-from-shutterstock-including-its-music-libraries/.

[121] *Firefly FAQ for Adobe Stock Contributors*, ADOBE (Updated Oct. 4, 2023), https://helpx.adobe.com/stock/contributor/help/firefly-faq-for-adobe-stock-contributors.html.

[122] Emilia David, *Getty Made an AI Generator that Only Trained on its Licensed Images*, THE VERGE, Sep. 25, 2023, https://www.theverge.com/2023/9/25/23884679/getty-ai-generative-image-platform-launch.

[123] Emma Roth, *Google Adds a Switch for Publishers to Opt Out of Becoming AI Training Data*, THE VERGE, Sep. 28, 2023, https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex.

[124] Emilia David, *Now You Can Block OpenAI's Webcrawler*, THE VERGE, Aug. 7, 2023, https://www.theverge.com/2023/8/7/23823046/openai-data-scrape-block-ai.

that will benefit all interested parties and society at large, rather than engage in litigation to protect their rights.

In this setting, the developers' goal to create LLMs or to employ those models to power generative AI products, however laudable, does not justify their infringement of this valuable corpus of copyrighted expression. Sam Altman, the founder of OpenAI, and Brad Smith, President of Microsoft, each acknowledged this point in their recent testimony before Congress, explaining that creators of expressive works deserve to control the rights to, and must benefit from, their creations.[125]

Indeed, courts have long recognized that such generalized fair use justifications should not be used to insulate widespread infringement. *American Geophysical Union v. Texaco, Inc*., for example, found that Texaco's photocopying of scientific journals for purposes of commercial R&D was not a fair use, even where the company had made the copies to enrich their researchers' knowledge, because the company was engaged in a "systematic process of encouraging employee researchers to copy articles so as to multiply available copies while avoiding payment."[126] As the court explained:

> The purposes illustrated by the categories listed in section 107 refer primarily to the work of authorship alleged to be a fair use, not to the activity in which the alleged infringer is engaged. Texaco cannot gain fair use insulation for [its employee]'s archival photocopying of articles (or books) simply because such copying is done by a company doing research. It would be equally extravagant for a newspaper to contend that because its business is "news reporting" it may line the shelves of its reporters with photocopies of books on journalism or that schools engaged in "teaching" may supply its faculty members with personal photocopies of books on educational techniques or substantive fields. Whatever

[125] *Oversight of A.I.: Rules for Artificial Intelligence*, 118th Cong. (2023), https://techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/ (statement of Sam Altman) ("we think that creators deserve control over how their creations are used and what happens sort of beyond the point of, of them releasing it into the world . . . . we think that content creators, content owners, need to benefit from this technology . . . . We're still talking to artists and content owners about what they want. I think there's a lot of ways this can happen, but very clearly, no matter what the law is, the right thing to do is to make sure people get significant upside benefit from this new technology. And we believe that it's really going to deliver that. But that content owners likenesses people totally deserve control over how that's used and to benefit from it."); *Oversight of A.I.: Legislating on Artificial Intelligence*, 118th Cong. (2023), https://techpolicy.press/transcript-us-senate-judiciary-hearing-on-oversight-of-a-i/ (statement of Brad Smith) ("generally I think we should let local journalists and publications make decisions about whether they want their content to be available for training or grounding and the like. And that's a big topic and it's worthy of more discussion. And we should certainly let them, in my view, negotiate collectively because that's the only way local journalism is really going to negotiate effectively.").

[126] 60 F.3d 913, 920 (2d Cir. 1994).

> benefit copying and reading such books might contribute to the process of "teaching" would not for that reason satisfy the test of a "teaching" purpose.[127]

This principle applies in full force to generative AI development. While developers have contended that their unlicensed use of material for LLM training and generative AI development purposes is justifiable because the LLMs ingest the copyrighted content to "learn" from the content, just like a human being, no one is allowed to copy an underlying work just because they have an alleged good reason to read the underlying document but don't want to buy (or otherwise lawfully access) a copy. As one scholar explains:

> Making gigabytes upon gigabytes of copies of copyrighted art, in order to teach a machine to mimic that art, is indeed a remarkable technological achievement. An artificially intelligent painter or writer may yield social benefits and enrich the lives of many beholders and users. However, this view of productivity is overbroad. No human can rebut an infringement claim merely by showing that he has learned by consuming the works he copied, even if he puts this new knowledge to productive use later on . . . . A teacher who copies to broaden his personal understanding is a productive consumer, but he nonetheless must pay for the works he consumes. If the teacher's consumption of copyrighted works inspires him to create new scholarship, so much the better, but his subsequent productivity does not entitle him to a refund for the works that influenced him. In much the same way, machine learning makes consumptive use of copyrighted materials in order to facilitate future productivity. If future productivity is no defense for unauthorized human consumption, it should not excuse robotic consumption, either.[128]

[127] *Id*. at 924; *see also Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1263-64 (11th Cir. 2014) ("[A]llowing some leeway for educational fair use furthers the purpose of copyright by providing students and teachers with a means to lawfully access works . . . . But, as always, care must be taken not to allow too much educational use, lest [the court] undermine the goals of copyright by enervating the incentive for authors to create the works upon which students and teachers depend."); *Princeton Univ. Press v. Mich. Document Servs., Inc*., 99 F.3d 1381 (6th Cir. 1996) (reproduction of significant portions of copyrighted works for use in course packets is not fair use); *Marcus v. Rowley*, 695 F.2d 1171 (9th Cir. 1983) (same for teacher's educational booklet); H.R. Rep. No. 94-1476, at 66-67 (1976), https://www.copyright.gov/history/law/clrev_94-1476.pdf ("[A] specific exemption freeing certain reproductions of copyrighted works for educational and scholarly purposes from copyright control is not justified."); Linda Starr, *Is Fair Use a License to Steal?*, EDUCATION WORLD, May 25, 2010, https://www.educationworld.com/a_curr/curr280b.shtml#:~:text=The%20fair%20use%20doctrine%20is,and%20scholarship%2C%20and%20classroom%20instruction.

[128] Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 COLUM. J. L. & ARTS 45, 73-74 (2017); *id*. at 74 (suggesting "a constituent who copies a news program to help make a decision on how to vote" would not be protected by the fair use doctrine despite the salutary purpose (quoting *Sony Corp. of Am. v. Universal City Studios, Inc*., 464 U.S. 417, 455 n.40 (1984))).

Of course, LLM machines are not humans. As explained in the White Paper, they do not "learn"—they copy, and they do so on a massive scale that no human could replicate. Because a market exists to provide high quality publisher content for purposes such as AI training, the goal of building LLMs does not justify the unlicensed copying of N/MA members' expressive works.

The Copyright Office asks specifically how recent Supreme Court precedent is relevant to this analysis. N/MA has incorporated the *Warhol* decision throughout its analysis. With respect to *Google v. Oracle,* the Court repeatedly took precautions to limit its reasoning to the specific software code at issue and, as a result, is not directly relevant.[129] Indeed, by beginning the analysis with factor two, and emphasizing the inherent functional nature of computer programs, the opinion is grounded in a very different factual surrounding than generative AI training, which ingests publisher works that have been repeatedly described by the Court as expressive works protected by copyright. [130] The limited applicability of *Google v. Oracle* to this instance is shored by *Warhol,* which, as noted, explains that the degree of transformation depends on the specific use.[131]

v. The unlicensed use of training materials serves a system designed to produce substitutional outputs.

LLMs are designed to produce outputs that can substantially copy from, compete with, and substitute for original text content. Even in the furtherance of new technological development, no court has held fair the copying of content to develop a system whose purpose is to substitute for the original works. Rather, 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. The first fair use factor does not require news and media publications to be mined to fuel their replacements.

In *Authors Guild*, for example, the court found that neither of the challenged uses (for "search" and "snippets") could provide a meaningful substitute for the copied books and instead were likely to help users identify books of interest.[132] It concluded that if the snippets were arranged into a coherent aggregate "manner and order" (which the challenged system disallowed) "that would raise a very different question beyond the scope of our inquiry."[133] Similarly, in *Kelly v.*

[129] *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1206-07 (2021).
[130]*Harper & Row,* 471 U.S. at 556-557; *Feist Publications, Inc. v. Rural Telephone Serv. Co*., 499 U.S. 340, 349 (1991)
[131] *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1206-07 (2021).
[132] 804 F.3d at 218.
[133] *Id.* at 223.

Arriba Soft Corp., the court found that the search engine "Arriba's use of Kelly's images in its thumbnails does not harm the market for Kelly's images or the value of his images."[134]

In contrast, LLMs can and do generate outputs that replicate or closely paraphrase the original expressive works. Consumer-facing chatbot services built around these models, including those integrated into search engines like Bing or Google, are well poised to directly substitute for publishers and to usurp their valuable relationships with readers of news, magazine, and web content. Marketing for these new features makes clear that they are intended to create substitutional narratives that can substantially copy from, compete with, and substitute for the primary expressive material. Unchained from constraints to serve as no more than an electronic reference or bridge to a primary source, narrative search results can provide users with sufficient content (full key portions and highlights of expressive content), that substitutes for any need to read the original. As a recent New Yorker article explains, the "goal" of "large language models, like OpenAI's ChatGPT and Google's Bard" "is to ingest the Web so comprehensively that it might as well not exist."[135]

These chatbot search uses thus go well beyond the nuanced reasoning and careful guardrails established by cases like *Authors Guild* and *Kelly* and into competitive, consumptive uses that are distinctly unfair to content owners. Indeed, courts routinely dismiss fair use arguments for new digital products that have a similar purpose to, and could supplant, the original work.[136] That reasoning applies here.

[134] 336 F.3d at 821; *see also Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1206-07 (2021) ("*Oracle*") (jury's fair use determination barred Oracle from "overcom[ing] evidence that, at a minimum, it would have been difficult for Sun [Oracle's predecessor] to enter the smartphone market" even without Google's alleged infringement, including Sun's former CEO's testimony that Sun's failure to build a smartphone was not attributable to Google's alleged infringement); *cf. Sony Corp. of Am. v. Universal City Studios*, 464 U.S. 417, 456 (1984) (noting that plaintiffs "failed to demonstrate that time-shifting would cause any likelihood of nonminimal harm to the potential market for, or the value of, their copyrighted works.").

[135] James Somers, *How Will A.I. Learn Next?*, THE NEW YORKER, Oct. 5, 2023, (reporting that the number of new posts the website Stack Overflow, where computer programmers went to ask and answer programming questions, has decreased by 16% since the debut of ChatGPT).

[136] *See*, *e.g.*, *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 of various websites for Associated Press's stories and scraping "snippets" of those stories for use in notifying and informing Meltwater's own customers of certain stories 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).

Moreover, recent lawsuits have alleged that certain LLMs knowingly ingested material from notorious pirate sites, and publishers have used their terms of service or using technical measures like robots.txt to prohibit crawling for purposes of generative AI ingestion.[137] If generative AI developers know or should have known that their systems are ingesting works that have been made available illegally, these acts would reflect bad faith or unclean hands, making a fair use defense less likely to succeed. This concept—that to invoke fair use, an individual must possess an authorized copy of a work—was addressed by the Court in *Harper & Row Publishers Inc. v. Nation Enterprises*, which confirmed that "[f]air use presupposes good faith" and found that Nation acted in bad faith because it "knowing exploited a purloined manuscript."[138] The Federal Circuit expanded on the concept in *Atari Games Corp. v. Nintendo of America, Inc.*, finding that because Atari gained access to an unauthorized copy of the Nintendo's source code by submitting false information to the U.S. Copyright Office, "any copying or derivative copying…does not qualify as a fair use."[139]

For these reasons, it is likely that with respect to LLMs, the first factor favors a finding of infringement and not fair use.

The effect of generative AI copying on the market for publisher content is predictable and real (fourth factor).

The fourth fair use factor directs courts to consider "the effect of the use upon the potential market for or value of the copyrighted work."[140] The focus is on whether widespread conduct like the conduct of the alleged infringer "would adversely affect the potential market for the copyrighted work," including market harm to the original work and to derivative works.[141] While the examination of potential markets is not without limit, "traditional, reasonable, or likely to be developed markets" are considered.[142] As the *Texaco* court recognized, "[i]t is indisputable that, as a general matter, a copyright holder is entitled to demand a royalty for

[137] *See Complaint, Authors Guild v. OpenAI*, at paras. 97-110 (Sep. 2023) available at https://authorsguild.org/app/uploads/2023/09/Authors-Guild-OpenAI-Class-Action-Complaint-Sep-2023.pdf; *Complaint, Tremblay v. OpenAI*, at paras. 31-34 (Jun. 28, 2023) available at https://torrentfreak.com/images/authors-vs-openai.pdf; *Complaint, Chabon v. Meta Platforms, Inc*. at paras. 26-39 (Sep. 12, 2023) available at https://fingfx.thomsonreuters.com/gfx/legaldocs/lbpgolxxmpq/META%20AI%20COPYRIGHT%20LAWSUIT%20complaint.pdf; Emma Roth, *Google Adds a Switch for Publishers to Opt Out of Becoming AI Training Data*, THE VERGE, Sep. 28, 2023, https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex.

[138] 471 U.S. 539, 547 (1985).

[139] 975 F. 2d 832 at 843(Fed. Cir. 1992).

[140] 17 U.S.C. § 107(4).

[141] *Harper & Row*, 471 U.S. at 566, 568 (emphasis omitted).

[142] *Am. Geophysical Union v. Texaco, Inc*., 60 F.3d 913, 929-30 (2d Cir. 1994).

licensing others to use its copyrighted work, and that the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth factor."[143]

GAI's unauthorized use of copyrighted material harms the market in two ways.

First, with respect to inputs, generative AI developers' unauthorized use of publisher content to build their LLMs deprives publishers of an available licensing market, such that the fourth factor also should favor a finding of infringement when publisher content is used without authorization for training purposes.[144]

While developers complain that it is unworkable to license content for their ingestion needs,[145] there is a long history of publishers licensing their content for a variety of uses and licensing deals, and negotiations are occurring in the open market specifically for GAI uses, as documented *infra*.

As explained above, including in response to questions 6 and 10, and in discussion of the first factor, there is also a long history of media organizations and associations licensing their content for a variety of uses, including to media monitoring entities, to LEXIS, and through the CCC.

Examples also abound, both here and abroad, of collective licensing of copyrighted content, and these models demonstrate the paths that exist for efficient licensing frameworks to meet AI needs. CCC, for example, was formed by authors, publishers, and users to facilitate "centralized licensing of text-based copyrighted materials," and it has grown to represent copyright holders from nearly every country, with access to millions of sources.[146] Outside the

[143] *Id*. at 929 (citation omitted).

[144] *Texaco*, 60 F.3d at 930 (finding fourth factor favored a finding of infringement where the challenged photocopying harmed an existing "workable market for institutional users to obtain licenses for the right to produce their own copies of individual articles via photocopying"); *see also Fox News Network, LLC v. TVEyes, Inc*, 883 F.3d 169, 180 (2d Cir. 2018) (by using content without payment, Fox was deprived of "licensing revenues from TVEyes"); *Davis v. Gap, Inc*, 246 F.3d 152, 175-76 (2d Cir. 2001) (freely taking a copyrighted work allowed defendant to avoid "paying the customary price," that plaintiff "was entitled to charge" for use of work, and that, as a result, plaintiff "suffered market harm through his loss of the royalty revenue to which he was reasonably entitled in the circumstances, as well as through the diminution of his opportunity to license to others").

[145] OpenAI, LP, Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation at 11, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf.

[146] Comments of Copyright Clearance Center, Inc. 79 Fed. Reg. 2696 (Mar. 3, 2024), https://www.copyright.gov/docs/recordation/comments/79fr2696/CCC.pdf; *Annual Copyright License*, Copyright Clearance Center.

United States, collective management organizations broadly manage news and media licensing, such as NLA Media Access in the U.K.[147]

Second, it is indisputable that generative AI output is intended to, and does, substitute for human-generated content, including publisher content.[148] As explained above, already less than 65% of searches result in clicking through to the underlying source.[149] That percentage is only going to increase with narrative search results. Indeed, marketing experts expect click-through rates for generative search responses to be even lower than already declining rates for organic results.[150] "Particularly for informational searches, Google will aggregate (or flat-out plagiarize) from the search results and give users much of what they're looking for."[151] "Users may find all the information they need directly on the search page, so there's no need to click on the source

[147] Tarja Koskinen-Olsson, *Collective Management of Text and Image-Based Works*, WIPO (Updated 2023) https://www.wipo.int/edocs/pubdocs/en/wipo-pub-924-2023-en-collective-management-of-text-and-image-based-works.pdf; *A Guide to Media Monitoring and Corporate Licensing*, Press Database and Licensing Network, at 14 (Oct. 2017), https://static1.squarespace.com/static/5eca9a7fe349354c54ae6cab/t/5ef2b3025a06263ec1a24a14/1592963847770/pdln_guide+to+corporate+and+mmo+licensing.pdf; *What Is a Performing Rights Organization (PRO)?*, SESAC (May 5, 2022), https://www.sesac.com/what-is-a-performing-rights-organization-pro/.
Collective licensing has also flourished in the music industry, further demonstrating the potential to develop efficient, large-scale licensing models for GAI needs. The performing rights organizations (PROs) such as ASCAP, BMI, and SESAC license the right to publicly perform musical compositions on behalf of copyright owners. PROs collectively "cover[] almost all of the millions of songs currently copyright protected," and they operate by offering "blanket authorization to use the music [each organization] represents in exchange for license fees," which are then distributed "as royalties to its affiliated songwriters, composers, and music publishers." *What Is a Performing Rights Organization (PRO)?*, SESAC (May 5, 2022), https://www.sesac.com/what-is-a-performing-rights-organization-pro/.

[148] *See also*, *e.g.*, Comment of OpenAI, LP Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, Before the USPTO, at 11, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("Writers who were employed to perform formulaic composition might be able to devote their energies to more creative forms of self-expression *once machines supplant them.*" (quoting Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45, 80 (2017))); Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743, 767 (2021) (Machine learning "empowers [] companies to extract value from authors' protected expression without authorization" or compensation "and to use that value for commercial purposes that may someday jeopardize the livelihoods of human creators." (quoting Sobel, *Artificial Intelligence's Fair Use Crisis*)); *id.* at 777 (AI systems trained "to generate their own expressive works . . . pose a threat of significant substitutive competition to the work originally copied." (internal quotation marks omitted)).

[149] *See* George Nguyen, *Zero-click Google Searches Rose to Nearly 65% in 2020*, Search Engine Land, Mar. 22, 2021, https://searchengineland.com/zero-click-google-searches-rose-to-nearly-65-in-2020-347115.

[150] *See*, *e.g.*, Rebecca Krause, *Google's Search Generative Experience (SGE): A Marketer's Guide*, Seer Interactive, Aug. 10, 2023, https://www.seerinteractive.com/insights/googles-search-generative-experience ("As SGE rolls out to more users, the click-through-rate of the ten organic links (even position 1) may lower.")

[151] Dave Shapiro, *Generative AI in Search*,Neil Patel (2023) https://neilpatel.com/blog/generative-ai-in-search/ ("people will find enough of what they need in the SGE and not click on organic results.").

website."[152] As set forth above, no court has deemed fair the copying of expressive works, even at the development stage, for the purposes of eventually competing with and substituting for the original work. The substitutional use of the generative AI outputs is a further reason why the fourth factor favors a finding of infringement with respect to the unauthorized use of publisher content at the training stage.

The effect of generative AI copying at the output stage is self-evident. Where the outputs replicate or closely paraphrase the original expressive works and thus infringe upon and substitute for them, such that users no longer need to connect with or obtain the original works from their original sources, such uses harm the market for the publishers' works.

Generative AI copying takes substantial portions of expressive works in their entirety (second and third factors).

Under the second factor, courts consider whether a work is creative or functional, "recogn[izing] that some works are closer to the core of intended copyright protection than others."[153] The second factor is typically less important than the first and fourth factors.[154]

Although news, magazine, and digital media content includes underlying facts, the reporting seeks to determine which facts are significant and to recount them in an interesting manner, and they are thus creative in nature.[155] Such content also extends well beyond traditional news reporting and includes pieces devoted to opinion and analysis. Here, where developers copy publisher content so that LLMs can best mimic human speech,[156] the copying is necessarily exploiting the content for its expressive qualities and the second factor favors a finding of infringement for both inputs and outputs.

The third factor evaluates both the quantity and quality of the copying, and "examine[s] the amount and substantiality of the portion used in relation to the copyrighted work as a whole,"

[152] Sam Stemler, *9 Things You Need to Know about Google Search Generative Experience (SGE)*, WEB ASCENDER, Aug. 29, 2023, https://www.webascender.com/blog/9-things-you-need-to-know-about-google-search-generative-experience-sge/.

[153] *Campbell*, 510 U.S. at 586; *Oracle*, 141 S. Ct. at 1202.

[154] *Authors Guild*, 804 F.3d at 220.

[155] *See Harper & Row*, 471 U.S. at 547 ("Creation of a nonfiction work, even a compilation of pure fact, entails originality."); *see also Authors Guild*, 804 F.3d at 220 ("Those who report the news undoubtedly create factual works. It cannot seriously be argued that, for that reason, others may freely copy and re-disseminate news reports."); *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 178 (2d Cir. 2018) (rejecting argument that, since facts are not copyrightable, the factual nature of a creative compilation favors a finding of fair use).

[156] *See* N/MA, WHITE PAPER at p. 8, 21-22 (2023), Appendix A; Stephen Wolfram, *What Is ChatGPT Doing ... and Why Does It Work?*, Feb. 14, 2023, https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

including whether the "heart" of the work is copied.[157] "[T]he fact that a substantial portion of the infringing work was copied verbatim is evidence of the qualitative value of the copied material, both to the originator and to the plagiarist who seeks to profit from marketing someone else's copyrighted expression."[158] The massive scale of copying also favors a finding of infringement.[159]

Here, for inputs, the developers copy all or substantial portions of the publisher content during the course of LLM training and development of generative AI tools, and it is reasonable to conclude that the "heart" of the work is copied. Moreover, copying for generative AI development can be viewed as excessive given the degree to which the copies usurp the available licensing market.[160]

Application of the third factor at the output stage must be evaluated on a case-by-case basis, depending on the portions of the works which the outputs copy. Suffice to say, the third factor will favor a finding of infringement at the output stage whenever the outputs copy sufficient portions or the heart of the copied works.

Question 8.4 asks whether the volume of material used to train an AI model affects the fair use analysis. Because LLMs and other generative AI models ingest a large amount of material, it does not appear necessary to ingest any one particular work. This fact would weigh against fair use, because there are other ways to develop a model beyond taking a particular work. By contrast, courts have found fair use favored when copying was "necessary" to gain access to functional elements of computer software,[161] and the Copyright Office has considered whether a potential licensing market exists when determining whether proposed uses of audiovisual clips for documentary filmmaking is likely to be fair.[162]

And the taking of a copyrighted work is not more likely to be fair because the allegedly infringing act also incorporated other material that was not infringing. As the *Harper & Row* Court explained:

[157] *Harper & Row*, 471 U.S. at 564-65).
[158] *Id*. at 565.
[159] *See*, *e.g.*, *Hachette Book Grp., Inc. v. Internet Archive*, No. 20-CV-4160 (JGK), 2023 WL 2623787, at *8 (S.D.N.Y. Mar. 24, 2023) ("Unlike Sony, which only sold the machines, IA scans a massive number of copies of books and makes them available to patrons . . . .").
[160] *Campbell*, 510 U.S. at 587-88; *see also* N/MA, WHITE PAPER at p. 37 (2023), Appendix A.
[161] *Sony Computer Ent. v. Connectix Corp.*, 203 F.3d 596 (9th Cir. 2000); *Sega Enterprises, Ltd. v. Accolade, Inc.*, 977 F. 2d 1510 , 1529 (9th. Cir. 1992);
[162] *See, e.g.,* ACTING REGISTRAR OF COPYRIGHTS, RECOMMENDATION: SECTION 1201 RULEMAKING: SEVENTH TRIENNIAL PROCEEDING TO DETERMINE EXEMPTIONS TO THE PROHIBITION ON CIRCUMVENTION, at 60- 61(Oct. 2018) available at https://cdn.loc.gov/copyright/1201/2018/2018_Section_1201_Acting_Registers_Recommendation.pdf.

> As the statutory language indicates, a taking may not be excused merely because it is insubstantial with respect to the *infringing* work. As Judge Learned Hand cogently remarked, "no plagiarist can excuse the wrong by showing how much of his work he did not pirate." *Sheldon* v. *Metro-Goldwyn Pictures Corp.,* 81 F. 2d 49, 56 (CA2), cert. denied, 298 U. S. 669 (1936).[163]

The proper question is how much of the copyrighted work has been used by the infringer in creating a secondary work. In this case, not only does it appear that generative AI developers have copied and used entire protected individual works, they have likely copied the entire corpora of our members' newspapers, magazines, and websites. A systemic disregard or carelessness towards copying large volumes of expressive works looks different than the targeted taking of a specific individual work, and should disfavor a finding of fair use.

9. Should copyright owners have to affirmatively consent (opt in) to the use of their works for training materials, or should they be provided with the means to object (opt out)? 9.1. Should consent of the copyright owner be required for all uses of copyrighted works to train AI models or only commercial uses? 9.2. If an "opt out" approach were adopted, how would that process work for a copyright owner who objected to the use of their works for training? Are there technical tools that might facilitate this process, such as a technical flag or metadata indicating that an automated service should not collect and store a work for AI training uses? 9.3. What legal, technical, or practical obstacles are there to establishing or using such a process? Given the volume of works used in training, is it feasible to get consent in advance from copyright owners? 9.4. If an objection is not honored, what remedies should be available? Are existing remedies for infringement appropriate or should there be a separate cause of action?

N/MA responds to question 9 and subparts 9.1-9.4 together.

The starting point to answer this question is the Copyright Act, which provides rightsholders with a bundle of rights that may be employed to provide necessary authorization for the use of copyrighted works absent applicable exceptions or defenses. That is, the existing law is "opt in." Consent can be provided by various means, which is also outlined by legal doctrines, but the general principle in copyright is to require affirmative consent absent an applicable exception or limitation. Changing this presumption under U.S. law would require the adoption of an additional exception under the law, a major undertaking that is not warranted under present circumstances.

[163] *Harper & Row Publishers, Inc. v. Nation Enters*., 471 U.S. 539, 547 (1985).

Discussions around opt-out are more relevant in countries and regions that, unlike the United States, may already have a statutory text and data mining (TDM) exception that allows some or all users to engage in TDM for limited AI training purposes. It is rare indeed to have a sweeping exception for TDM that extends to highly commercial uses without the ability to opt out. In addition to raising other potential concerns, including compliance with international agreements, retaining the ability to opt out of such exceptions is important in those countries or regions. The United States, however, has not adopted an exception to our copyright laws for TDM. N/MA opposes the creation of a new or expanded exception to copyright law that would change the status quo to permit AI training without the rightsholder's authorization.

To date, current tools present a Potemkin village of a solution, providing limited benefits to publishers while creating a patina of responsibility to justify positions that copying is legal absent affirmative opt-out. It is inappropriate industrial policy to place the burden on a copyright owner to remedy a potentially infringing act, rather than on a generative AI developer or deployer who already possesses the right and ability to control what material is used for training (whether by selecting, cleaning, or fine tuning a dataset, licensing content, or by paying a low wage to someone overseas to mitigate the worst violations). And the necessary act of choosing what copyrighted works an AI system is trained on distinguishes these developments from the architectures that gave rise to the section 512 safe harbor.

To be sure, there may be limited room for voluntary signals or solutions that may simplify licensing. This is particularly the case as publishers and other rightsholders explore reasonable technical and collective licensing solutions in response to developments in other parts of the world, including the EU's Directive on Copyright in the Digital Single Market.[164] These measures should be industry-led and agreed to by rights holders of particular sectors to prevent very large platforms from imposing methods on publishers and ensure a workable framework for all. The government could play a limited role in facilitating these conversations and in ensuring compliance with the voluntary measures, potentially by imposing penalties for generative AI developers who fail to honor such opt-in or opt-out measures or protocols.

A voluntary opt-in system could be aided by a publisher-led collective licensing entity or a technical measure that allows publishers to signal to generative AI developers that their content is available for AI training purposes, subject to any relevant terms. Such a solution could lower the burden of acquiring licenses for developers—including retroactively for content

[164] 2019, O.J. (Directive (EU) 2019/790) The European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market, at Art. 3 & 4, available at https://eur-lex.europa.eu/eli/dir/2019/790/oj.

that has already been scraped for existing applications—while making it easier for publishers to scale up their licensing to multiple licensors, thus facilitating increased choice.

Meanwhile, it is important to distinguish opt-out signals for scraping from copyright licenses, express or implied. For example, some uses of content may constitute fair use (such as certain uses for search purposes) and the owner of the site may wish to signal no more than that scraping is permitted for such purposes. In some cases, the site owner may not be in a position to authorize certain uses—either because they may not have or may not know the full scope of the rights it controls to all of the content on the site for every possible use.[165] This is a common situation for many publishers who make available to the public a wide variety of content, some of which is work-for-hire, some of which was created for the publisher by independent third parties, and some of which is licensed. In other cases, the site owner may wish to authorize access to its sites only for certain limited purposes and only to certain authorized parties for commercial reasons. Any automatic signal, whether opt-in or opt-out, must account for these differences.

One of the most widely advocated exclusion protocols, robots.txt, is currently a blunt tool that does not offer sufficient granular control over the types of uses for which scraping is allowed. As a result, site owners are forced to choose between authorizing ***any*** use and authorizing no uses. Media publishers often depend on search to generate a significant part of their traffic. Consumers also depend on search to locate material online. Blocking ***all*** scraping would eliminate this important source of traffic, but permitting scraping by not including the robots.txt signal certainly doesn't extend permission for developers to make ***all*** potential uses of the content. The availability of the robots.txt signal is insufficient to solve this problem. Robots.txt is a voluntary regime and many scrapers disregard the signal. It would be helpful to both developers and rightsholders for scrapers to honestly and transparently identify the entity that is scraping, and abide by industry standard licenses that can be identified automatically, in a signal similar to robots.txt. N/MA would support the Copyright Office facilitating discussions on voluntary opt-out signals, while ensuring that scrapers have incentives to respect them. Incentives could include potential legal penalties for scrapers who disregard such signals, or fail to provide truthful information regarding their identity and the uses to which the scraped material will be put. Such a standard could be developed by industry, with the backstop of having the conditions—transparency and an obligation to follow the rule—enforceable by law.

[165] This is especially true for publishers who have accumulated content over decades on many iterations of contracts, sometimes in the tens of thousands, which would need to be reviewed for legal compliance with a new and developing use.

Publishers are also concerned that opt-out systems must be efficient at scale. In an opt-out regime, developers may have incentives to make opt-out difficult for publishers (or at least, expend the minimum compliance efforts required), whereas with an opt-in regime, developers are incentivized to seek efficient licensing solutions. For example, DALL-E's opt-out system requires the "owner or rights holder . . . to submit *an individual copy of each image* they'd like removed from DALL-E's training dataset, *along with a description.*"[166] This is obviously impractical for more than a *de minimis* number of images. The Copyright Office need only recall its years-long DMCA study to predict the difficulties with this system.

Moreover, an opt-out regime puts the burden on copyright owners to find out who is using their material. Not only does this incentivize non-disclosure, but developers commonly train their systems on material acquired from sites that have been identified by the U.S. government as notorious markets for piracy,[167] necessitating that copyright owners enforce rights against infringers as a prerequisite—a burden that is impossible to achieve.

With respect to question 9.1, concerning non-commercial and commercial uses, that question may be better evaluated in the context of evaluating infringement or fair use, *see infra*. But with respect to signaling consent, or the lack thereof, by opting in or out of AI training, it is difficult to make blanket exceptions or judgments based on the identity of the user or category of use. For example, content scraped for a seemingly noncommercial use—potentially at the request or with the support of a commercial developer—can and often is passed onto a commercial entity that may create products or services that directly compete with the content creator. At that point, it becomes much more difficult for a rightsholder to "opt in" to or "opt out" of a use that has already occurred. The prevalence of data laundering, as well as the lack of bright lines distinguishing commercial from noncommercial uses with these technologies, makes this question difficult to answer on a black and white basis.

While objecting to any mandatory opt-out requirement, N/MA would support the Copyright Office facilitating discussions on voluntary opt-out signals, while ensuring that developers have incentives to respect them, potentially by imposing penalties for disregarding such signals.

[166] Kali Hays, *OpenAI Offers a Way for Creators to Opt Out of AI Training Data. It's so Onerous that One Artists Called it 'Enraging'*, GOOGLE: INSIDER, Sep. 29, 2023, https://www.businessinsider.com/openai-dalle-opt-out-process-artists-enraging-2023-9?r=US&IR=T.

[167] *See* Kevin Schaul *et al., Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart,* The Washington Post (Apr. 19, 2023), *available at* https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/ (last visited Oct. 28, 2023); Alex Hern, *Fresh Concerns Raised Over Sources of Training Material for AI Systems,* The Guardian (Apr. 20, 2023), *available at* https://www.theguardian.com/technology/2023/apr/20/fresh-concerns-training-material-ai-systems-facist-pirated-malicious# (last visited Oct. 28 2023).

9.5. In cases where the human creator does not own the copyright—for example, because they have assigned it or because the work was made for hire—should they have a right to object to an AI model being trained on their work? If so, how would such a system work?

The Copyright Act and the relevant case law sets out clear rules for how to handle works made for hire and assigned copyrights. There is no need for special rules for AI in this respect. While N/MA expresses no opinion as to aspects of this question that implicate other rights, including moral, privacy, and contracts, there is no need to revisit this principle under copyright law.

10. If copyright owners' consent is required to train generative AI models, how can or should licenses be obtained? 10.1. Is direct voluntary licensing feasible in some or all creative sectors? 10.2. Is a voluntary collective licensing scheme a feasible or desirable approach? Are there existing collective management organizations that are well-suited to provide those licenses, and are there legal or other impediments that would prevent those organizations from performing this role? Should Congress consider statutory or other changes, such as an antitrust exception, to facilitate negotiation of collective licenses? 10.3. Should Congress consider establishing a compulsory licensing regime? If so, what should such a regime look like? What activities should the license cover, what works would be subject to the license, and would copyright owners have the ability to opt out? How should royalty rates and terms be set, allocated, reported and distributed? 10.4. Is an extended collective licensing scheme a feasible or desirable approach? 10.5. Should licensing regimes vary based on the type of work at issue?

N/MA responds to the licensing issues posed by question 10 and its subparts together. Our response focuses on the landscapes and dynamics experienced by our members.

As discussed above and for clarity, commercial generative AI companies do need consent to use our members' content under existing law. The Notice asks many questions around how emerging licensing frameworks can respond and adapt to continued innovations in generative AI technologies. A rights-based regime is best suited to answer these questions flexibly, through direct negotiations among the affected parties. While N/MA can't speak for other creative sectors, we certainly believe that voluntary licensing is feasible—and the most desirable—for publishers of newspapers, magazines and digital media content.

The Office Should Encourage Market-Based License Solutions and Reject Calls for Compulsory Licensing

Marketplace licensing, including on a collective basis where appropriate, is the default legal system under U.S. law and should be the default here. Voluntary licensing is especially preferred here, where generative AI technologies are so new, the uses of AI so unpredictable, and the economics so unknown, that it is imperative that publishers and AI developers be given maximum flexibility in structuring (and restructuring) deals as the marketplace evolves.

The nascency of generative AI is already spawning varied companies, products, and services that will have different economic implications for authors, copyright owners, and their businesses.[168] There is unlikely to be a one-size-fits-all solution to licensing copyrighted material for ingestion and other uses by generative AI-dependent entities.

At a moment when marketplace actors are interested in negotiating private arrangements, the Copyright Office should firmly reject calls to establish a compulsory license to permit copyrighted content to be ingested into AI systems under government-set terms.[169]

This view is most consistent with the international copyright legal framework and the longstanding views of the Copyright Office itself. As former Register of Copyrights Marybeth Peters testified to Congress regarding the section 115 license for musical works, compulsory licensing is a "last resort mechanism," typically only seen where there has been a failure of voluntary agreements.[170] As she further explained, "[a] compulsory license limits an author's bargaining power. It deprives the author of determining with whom and on what terms he wishes to do business."[171] For that reason, as the Office explained in connection with a recommendation to sunset the section 119 license for satellite distant signals, "[h]istorically, the Copyright Office has supported statutory licenses only when warranted by special

[168] The Glossary appended to the Office's Notice and variety of definitions offered by policymakers in the EU, U.S., UK, and other markets to categorize obligations by differing types of AI-related actors illustrates this shifting landscape.

[169] It would similarly be premature for the Office to support calls for extended collective licensing (ECL) models. *See* U.S. Copyright Office, Request for Comments on Artificial Intelligence and Copyright, at 2. (Aug. 30, 2023) 88 FR 59942, available at 10.4 https://www.federalregister.gov/documents/2023/08/30/2023-18624/artificial-intelligence-and-copyright.

[170] *See* Section 115 Compulsory License: Hearing before the Subcom. On Courts, the Internet and Intell. Prop. of the H. Comm. on the Judiciary, 108th Cong. (Mar. 11, 2004) (statement of Marybeth Peters, Register of Copyrights) ("[U[se of the compulsory license should only be made as a last resort, and that licensees should be encouraged to obtain voluntary licenses directly from the copyright owners or their agents, who would offer more congenial terms.") available at: https://www.copyright.gov/docs/regstat031104.html. *See also* U.S. Copyright Office, Copyright and the Music Marketplace at 112 (Feb. 2015) available at https://www.copyright.gov/policy/musiclicensingstudy/copyright-and-the-music-marketplace.pdf (compulsory licensing "removes choice and control from all copyright owners that seek to protect and maximize the value of their assets."); European Commission, A Single Market for Patents: New Rules on Compulsory Licensing (April 2023) available at: https://single-market-economy.ec.europa.eu/system/files/2023-04/Patent%20Package_Compulsory%20Licensing_Final.pdf ("Compulsory licensing is a last resort mechanism which allows a government to authorise the use of a patented invention without the consent of the patent holder").

[171] *Id.* Statement of Marybeth Peters, Section 115 Compulsory License: Hearing (2004).

circumstances and only for as long as necessary to achieve a specific goal."[172] And the Office reaffirmed this position again in 2019:

> A statutory license creates an artificial, government-regulated market that operates as an exception to the general rule that copyright owners hold exclusive rights and can negotiate whether and how and at what cost to distribute their copyrighted works; statutory licenses tend to be below the fair market value.[173]

In addition to denying copyright owners the freedom to license as they see fit, a compulsory license would risk ossifying the innovative potential for generative AI technologies. Because a statutory license must clearly set out the scope and terms of the license, it is unlikely to be sufficiently flexible and adaptable to serve the legitimate needs of both publishers and AI developers and keep pace with technological and market developments.

It is important that copyright preserve the core function of market-based incentives for humans to create and disseminate works of authorship as generative AI products and services gain further traction. And this is especially important in the case of newspapers, magazines, media websites, and books, where a compulsory licensing regime could create a risk of political interference from Congress or the Executive Branch.

Conditions Exist for a Strong Licensing Ecosystem to Flourish Between Media Publishers and AI Developers and Other Licensees

Unlike the rare exceptions where government-regulated licensing is necessary, there is no evidence of market failure here to support intervention at this time. Media publishers already operate robust existing licensing arms as part of their established businesses. Well-established markets exist for the licensing of archival material and other real-time access to news content, including for use in new products and technologies. In fact, some of the major developers that have copied and used content without permission are already business partners and licensees of N/MA member publishers in connection with other products.[174]

[172] U.S. Copyright Office, Satellite Television Extension and Localism Act: A Report of the Register of Copyrights at 1 (Aug. 29, 2011) available at https://copyright.gov/reports/section302-report.pdf.

[173] U.S. Copyright Office Analysis and Recommendations Regarding the Section 119 Compulsory License; Response to House Comm. on the Judiciary, 115th[ck] Cong. at 5 (Jun. 3, 2019), available at https://copyright.gov/laws/hearings/views-concerning-section-119-compulsory-license.pdf.

[174] Sarah Fischer, *Google to Launch News Showcase Product in U.S.*, Axios, Jun. 8, 2023, https://www.axios.com/2023/06/08/google-news-showcase-us (describing Google licensing deals with 150 news publishers across 39 states); Ahiza Garcia, *Facebook Offers Media Outlets Millions to License Content, WSJ Reports,* CNN, Aug. 9, 2019, https://www.cnn.com/2019/08/08/tech/facebook-news-outlets-license-rights-content/index.html (describing Facebook offers to license with news publishers).

For generative AI development specifically, and as explained in response to questions 6 and 8, the market is already responding to the demand to provide high quality news and media content, and N/MA members are actively working to grow this field. This licensing for generative AI development is part and parcel of the long existing and well-established markets for licensing archival material and other real-time access to trustworthy news content.[175]

Voluntary Collective Licensing Can Play a Role in Licensing Works for Generative AI Uses

For media publishing, the marketplace can support both individual, direct negotiated arrangements (between a publisher and a LLM provider or other generative AI company) as well as voluntary collective licensing arrangements. Such a structure can be more agile in response to technological and business developments than a regulated solution, while supporting a competitive marketplace for the affected sectors.

While collective licensing should not be required, and individual licensing always permitted, voluntary collective licensing may well prove useful by providing the ability to aggregate smaller publishers, thereby reducing transaction costs and facilitating more efficient licensing and distribution for a greater number of licensors. Collective licensing would benefit competition among LLM providers. Today, the largest LLM providers crawl and index online content to build their corpus of training data. That process is expensive and difficult, requiring massive scale. It thus forms an entry barrier to nascent competitors. A collective licensing entity could aggregate, standardize, and distribute content from smaller publishers, allowing smaller LLM competitors to at least partially bypass the need to crawl and index web content.

Voluntary collective licensing would also not be unusual. Collective licensing entities already exist that satisfy competition law requirements, including reproduction licensing organizations like CCC. Examples of entities and models engaged in licensing other forms of copyrighted works include a society that issues licenses and distributes licensing fees for over 70,000 fine artists (ARS); a licensing entity that issues blanket licenses for worship music to churches, schools and religious organizations (CCLI); a licensing entity that authorizes non-theatrical uses of motion pictures by organizations in child care, education, communal living facilities, corporations, and others (MPLC); entities that offer subscription licenses to millions of images, videos and music created by millions of contributors (Shutterstock, Getty Images, Unsplash, Storyblocks, iStock, 123RF, Vecteezy, Pixabay, Adobe Stock, JumpStory); performing rights organizations in the music industry (GMR and SESAC, as well as ASCAP and BMI, which are subject to consent decrees); an indie label organization that negotiates model licenses with

[175] See *What We Do*, N.Y. TIMES, [n.d.],https://nytlicensing.com/what-we-do/ (last visited Oct. 25, 2023); *Products*, WASH. POST [n.d.], https://www.washingtonpost.com/licensing-syndication/products (last visited Oct. 25, 2023).

streaming services and other commercial users to which individual labels can opt in (Merlin); and a trade association that has negotiated agreements with major commercial users to which individual publishers can opt in (NMPA); among others. Taken together, these organizations exemplify that rights can be licensed efficiently when markets are allowed to develop. Moreover, often the "back office" technological and other infrastructure needs of a licensing entity may be able to be outsourced to existing organizations that already have the necessary capabilities, including entities like SoundExchange, SESAC, CCC, and others.

Indeed, while competition considerations must be approached carefully, courts have upheld various structures and models used to facilitate copyright collective licensing, including on a blanket license basis.[176]

Collective licensing may be particularly well-suited with respect to media publishing and generative AI development. Collective licensing nuances can vary by market, the nature of the works and uses, and the licensor/licensee parties involved. The licensing of media publishing content can be expected to operate differently than other copyright markets, such as the licensing of musical works for digital streaming services (with which the Office is familiar, given its work with the Music Modernization Act). There are a few key differences. First, LLMs require ingesting a large amount of textual content, but there does not appear to be an expectation that LLMs were trained on a "full catalog" of content, making it easier for a licensee to walk away. This is unlike licensing for music streaming services, where consumer expectations that streaming services offer a "full catalog" may factor into licensing negotiations.[177] Second, media publishing does not have the same fragmentation of the rights to be licensed (since news and media publishers typically control the necessary rights for their mastheads, and there is no need to "match" pieces of a textual work in the same way licenses for musical works and sound recordings must each be separately licensed). These features thus reduce the risk that "must have" or hold-out publishers would be able to extract supracompetitive pricing, but, by the same token, increase the attractiveness of voluntary collective models to facilitate licensing of material by smaller publisher operations.

[176] *See, e.g.*, *Buffalo Broad. Co., Inc., v. ASCAP,* 744 F.2d 917, 920 (2d Cir. 1984); *Broadcast Music, Inc. v. Columbia Broadcasting System, Inc.*, 441 U.S. 1, 23 (1979) ("Joint ventures and other cooperative arrangements are . . . not usually unlawful, at least not as price-fixing schemes, where the agreement on price is necessary to market the product at all."). *See also Texaco, Inc. v. Dagher,* 547 U.S. 1 (2006) (holding that internal pricing decisions of a legitimate joint venture are not per se unlawful).

[177] U.K. COMPETITION AND MARKETS AUTHORITY, MUSIC AND STREAMING: FINAL REPORT at 14, 73-74, 76, (2022) *available at* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1120610/Music_and_streaming_final_report.pdf.

The Office Should Encourage the Development of Voluntary Marketplaces

In light of the strong opportunities for voluntary individual and collective licensed solutions to be structured in ways that are pro-competitive, N/MA does not believe the Copyright Office needs to recommend intervention at this stage. That said, it is possible that legislation, such as antitrust exceptions, to augment existing abilities to negotiate collectively could be helpful.

Indeed, in other contexts, antitrust exceptions are strongly needed to correct market imbalances that are harming what the Office has called "the press's essential role in our system of government."[178] The Office has previously noted the potential for changes to competition policy to serve as an effective means to improve the position of press publishers in dealing with news aggregators.[179] N/MA supports the Journalism Competition and Preservation Act ("JCPA"). While the Office previously declined to offer a recommendation with respect to competition policy,[180] in light of the Office's current interest in the interrelation between copyright and competition interests, N/MA urges the Office to follow the logical conclusion of its press publisher study and support competition-based policy changes like JCPA to improve protections for sustaining journalism.

However, with respect to licensing of media content for generative AI uses, it is not clear that such legislation is actually necessary given that many collective licensing entities (some described above) currently operate in accordance with antitrust laws without the need for legislative exceptions.

Given the explosion of commercial LLM products, mostly without obtaining the permission necessary to make use of the content they have taken, licensing for current and future LLM models should be put in place swiftly. For N/MA's part, its members are willing to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy expressive content (including for past takings), something that will benefit all interested parties and society at large, and avoid protracted litigation.

11. What legal, technical or practical issues might there be with respect to obtaining appropriate licenses for training? Who, if anyone, should be responsible for securing them (for example when the curator of a training dataset, the developer who trains an AI model,

[178] U.S. Copyright Office, Copyright Protections for Press Publishers at 4 (2020) ("Press Publishers Study"), available at https://copyright.gov/policy/publishersprotections/202206-Publishers-Protections-Study.pdf.
[179] *Id.*
[180] *Id.* at 24.

and the company employing that model in an AI system are different entities and may have different commercial or noncommercial roles)?

N/MA refers to question 10 above. A rights-based framework exists and is best suited to address these questions flexibility, through market negotiations among the affected parties. This will also allow for different transactions to emerge amongst the user-side licensees, including curators, developers, and deployers of generative AI models.

Developers should be prohibited from ingesting materials for training purposes from sources known to contain pirated content. Such sites should be blocked and prohibited for use in training.

12. Is it possible or feasible to identify the degree to which a particular work contributes to a particular output from a generative AI system? Please explain.

With respect to training, it is evident that particular works may be more or less valuable for training than other works, and this can be reflected in license pricing and terms. There may also be other relevant terms to be negotiated based on what is technically feasible and valued between licensing partners (e.g., territoriality, output similarity, attribution, etc.).

As explained in the White Paper, and in response to questions 6, 7.1, and 8, media content accounts for a substantial volume of the known sources for LLM training, suggesting that this high quality expressive material is especially desirable by developers. Forensic analysis shows:

- Developers have copied and used news, magazine and digital media content to train LLMs.
- Popular curated datasets underlying LLMs significantly overweight 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.
- Other studies show that news and digital media ranks third among all categories of sources in Google's C4 training set, which was used to develop Google's generative AI-powered products like Bard. Half of the top ten sites represented in the data set are news outlets.
- LLMs also copy and use publisher content in their outputs. LLMs can reproduce the content on which they were trained, demonstrating that the models retain and can memorize the expressive content of the training works.

N/MA notes that showing that a particular work contributes to a particular output from a generative AI system is not required to establish prima facie infringement for purposes of training/ingestion. But outputs can offer clear evidence that a particular work has been copied. In too many cases, N/MA members have documented instances where it appears that an output results from copying one particular work. In those cases, similar outputs (including identical and substantially similar outputs) can make more clear that fair use does not apply.

As noted in response to questions 22-27, outputs can also separately be evaluated for additional infringement claims.

13. What would be the economic impacts of a licensing requirement on the development and adoption of generative AI systems?

Generative AI development is unlikely to succeed without a robust ecosystem that facilitates licensed use of valuable, authentic news media material. The failure to license publishing content may negatively impact the valuation of AI companies themselves, creating a cloud on the technology precisely because it is unlicensed.[181] Companies that might otherwise want to license from and deploy generative AI products and services may hang back as long as the IP issues are unresolved.

And a market that facilitates licensed exchanges of human-created content is needed for continued innovation. Researchers have found "that use of model-generated content in training causes irreversible defects in the resulting models," an effect they term "model collapse."[182] Even short of a complete model collapse under a deluge of synthetic content, there is an increased risk that generative AI chatbots could become an unattractive swamp of hallucinations without the ability to use human-created content that reflects thoughtful editorial judgment and creative expression.

The flourishing of AI technologies requires development that incorporates design principles that underscore public safety, security, and trust—as demonstrated by the recent voluntary commitments from leading AI companies to the Biden-Harris Administration and the

[181] Indeed, some AI developers have taken the unusual step of pledging to defend users of their products, in perhaps an implicit recognition of such a cloud. *See* Blake Brittain, *Google to Defend Generative AI Users from Copyright Claims*, REUTERS, Oct. 12, 2023, https://www.reuters.com/technology/google-defend-generative-ai-users-copyright-claims-2023-10-12/.

[182] Ilia Shumailov, et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget*, ARXIV, at 13 (May 27, 2023), available at https://arxiv.org/abs/2305.17493. *See also* Sina Alemohammad, et al., *Self-Consuming Generative Models Go MAD*, ARXIV, (Jul. 4, 2023), available at https://arxiv.org/abs/2307.01850.

Administration's Executive Order.[183] Companies that adequately account for intellectual property responsibilities in their business models at the outset will be better poised to enjoy the tremendous potential economic benefits promised by AI innovation.

The Notice's question is oddly phrased, suggesting a "licensing requirement" rather than the need to adhere to established law. As explained further in response to question 6.3, in addition to obtaining permission to use third party material, entities may make use of material in the public domain, material they have created themselves, or material that may fall under a relevant exemption of limitation of copyright. But AI developers should not get a pass to create models that usurp licensing markets and compete with publisher content just because. What FTC Chair Lina Khan recently observed in another context holds true for copyright as well: "there is no AI exemption to the laws on the books."[184]

If the law was ignored, the economic impact of generative AI technologies on publishers and the entire information ecosystem, including authors and publishers of copyrighted works, could be catastrophic. The Office should encourage market development that supports the protection and licensed use of expressive content for ingestion into LLMs and other AI models.

In any event, foundational model developers are operating licensing companies themselves, offering access to LLM models in commercial arrangements with a panoply of downstream entities.[185] Some creators of datasets are also licensing the datasets (including on a royalty-free basis). The potential for a robust LLM licensing has fueled significant investments and increased

[183] Statements and Releases, The White House, Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI (Jul. 21, 2023) available at https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-; The White House, Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence (October 30, 2023), available at https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[184] Press Releases, Fed. Trade Comm., FTC Chair Khan and Officials from DOJ, CFPB and EEOC Release Joint Statement on AI (Apr. 25, 2023) available at https://www.ftc.gov/news-events/news/press-releases/2023/04/ftc-chair-khan-officials-doj-cfpb-eeoc-release-joint-statement-ai.

[185] *See, e.g.*, Katherine Lee, A. Feder Cooper & James Grimmelmann, Talkin' 'Bout AI Generation: Copyright and the Generative AI Supply Chain, at 4-5 (Draft Sep. 15, 2023), available at https://james.grimmelmann.net/files/articles/talkin-bout-ai-generation.pdf (outlining supply chain); Alex 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?embedded-checkout=true; OpenAI, *Pricing* [n.d.], https://openai.com/pricing (last visited Oct. 26, 2023); Amazon & Anthropic, *Expanding access to safer AI with Amazon*, Anthropic, Sep. 25 2023, https://www.anthropic.com/index/anthropic-amazon.

valuation for these developers.[186] The better question for the Office to ask is whether it is sound intellectual property and industrial policy to begin a licensing supply chain at the foundational model provider, rather than further up towards the source, with the authors and publishers who create the content that is a key input for those providers. The economic impacts on publishers should not be considered mere externalities to the hopes for AI innovation.

For these reasons, we do not believe that fair licensing will hinder generative AI development—to the contrary, it is likely to improve the quality and accuracy of generative AI. Indeed, one copyright veteran observed that similar fears were raised in connection with the growth of photocopying in the 1960s.[187] At that time, some entities argued it would be impossible to secure all needed permissions to facilitate scientific progress, and regulation would put the U.S. at a competitive disadvantage. However, judicial recognition that not all photocopying was fair use did not impede innovation but led to a regime of voluntary collective licensing that has facilitated copying, enhanced access, and supported creative incentives by providing compensation to authors and rightsholders.[188]

14. Please describe any other factors you believe are relevant with respect to potential copyright liability for training AI models.

To the extent that material ingested for AI was obtained in a manner contrary to publishers' terms of service, or otherwise exceeding the bounds of access granted on the public internet or otherwise, N/MA notes that this may give rise to additional liability risk. For example, the FTC has opened an investigation into whether OpenAI has engaged in unfair or deceptive privacy or data security practices in scraping public data.[189] Further, the manner in which material was scraped and obtained may be considered when evaluating questions of copyright liability.[190] Whether copyrighted works were scraped from illegal sources, as alleged in currently pending lawsuits, or contrary to terms of service or technical measures, can be relevant to a fair use analysis. It also may be relevant in considering potential damages.

[186] *See, e.g.,* Mary Azevedo, *OpenAI Could See Its Secondary-Market Valuation Soar to $90B,* TECHCRUNCH, Sep. 26, 2023, https://techcrunch.com/2023/09/26/openai-is-reportedly-raising-funds-at-a-valuation-of-80-billion-to-90-billion/?guccounter=1.

[187] Jon Baumgarten, *Former Copyright Office GC Warns Against Blanket Assertions That AI Ingestion of Copyrighted Works 'Is Fair Use'*, COPYRIGHT ALLIANCE, May 23, 2023, https://copyrightalliance.org/warns-assertions-ai-ingestion-is-fair-use/.

[188] *Id.* The CCC is such a model.

[189] Cat Zakrzewski, *FTC Investigates OpenAI over Data Leak and ChatGPT's Inaccuracy*, THE WASH. POST, Jul. 13, 2023, https://www.washingtonpost.com/technology/2023/07/13/ftc-openai-chatgpt-sam-altman-lina-khan/.

[190] *See*, *e.g.*, *Harper & Row v. Nation Enterprises*, 471 U.S. 539, 547 (1985), 17 U.S.C. 1201.

Transparency and Recordkeeping

15. In order to allow copyright owners to determine whether their works have been used, should developers of AI models be required to collect, retain, and disclose records regarding the materials used to train their models? Should creators of training datasets have a similar obligation? 15.1. What level of specificity should be required?

N/MA supports the development and adoption of strong transparency requirements for generative AI developers because the status quo is insufficient. Indeed, Stanford University's Institute for Human-Centered AI recently published an index rating the transparency of 10 foundational model companies, finding each of them "lacking."[191] Having actual, verifiable, and accurate information regarding the uses of protected publisher content is vital for effective copyright enforcement. Such transparency requirements will likely also benefit other policy objectives outside copyright, such as safety audits, bias mitigation, risk assessments, and combating deepfakes, mis- and disinformation, hate speech, and other online harms. While this may be a multi-agency effort, N/MA believes that the Copyright Office, FTC, and USPTO can, and should, play a key role in these discussions from an IP perspective, due to their significant importance for rightsholders' ability to protect their copyrights online.

The United States is not alone in tackling issues related to AI transparency requirements.[192] Other international bodies, countries, and regions are also actively considering similar measures, including the European Union and the G7 through the Hiroshima Process.[193] While the G7 countries aim to develop global AI standards that can serve as the baseline for domestic AI laws and regulations, the European Union is already actively considering a proposal related to copyright and transparency in the AI Act. The EU institutions are currently engaged in trilogue negotiations, where the negotiators are weighing the Parliament's proposal to require

[191] Katharine Miller, Introducing The Foundation Model Transparency Index, STANFORD UNIVERSITY, Oct. 18 2023, available at https://hai.stanford.edu/news/introducing-foundation-model-transparency-index.

[192] Nor is transparency reporting a new concept for many of the very platforms that are engaging in LLM development. Many have experience preparing reports in the context of compliance obligations under online safety, privacy, or existing copyright-related duties, particularly outside the U.S.

[193] See, e.g., G7 HIROSHIMA AI PROCESS, G7 DIGITAL & TECH MINISTERS' STATEMENT (2023) available at https://www.politico.eu/wp-content/uploads/2023/09/07/3e39b82d-464d-403a-b6cb-dc0e1bdec642-230906_Ministerial-clean-Draft-Hiroshima-Ministers-Statement68.pdf; Supantha Mukherjee, Foo Yun Chee & Martin Coulter, *EU Proposes New Copyright Rules For Generative AI*, REUTERS, Apr. 28, 2023, https://www.reuters.com/technology/eu-lawmakers-committee-reaches-deal-artificial-intelligence-act-2023-04-27/.

generative AI developers to provide disclosures about the inclusion of copyrighted content in their training data.[194]

While the Parliament's original proposed amendment is somewhat ambiguous and should be adjusted to ensure it provides tangible, enforceable benefits to rightsholders, it is a positive step forward that can be emulated, and improved upon, outside the EU. International harmonization is particularly important as divergence in standards and enforcement may facilitate circumvention due to the borderless nature of the online ecosystem. Further, international standards or policy alignment would also lower compliance and litigation costs, and increase legal certainty and predictability to generative AI developers and rightsholders alike. Relatedly, N/MA has encouraged the Administration to take a leading role in the global discussions and to remain active in international fora finding solutions to these issues.[195]

To be meaningful, transparency standards should require generative AI and dataset developers to keep records about the protected works included in the training data and associated metadata, perhaps alongside an explanation of the legal basis on which their scraping, access, or inclusion is based. Such information should be categorized and provided in a manner that is manageable and easily searchable.

The minimum floor should be set at a level that allows rightsholders to easily and unambiguously identify when their content is being used or has been used for AI training purposes in order to enable rightsholders to effectively choose how to exercise their rights. Applicable disclosures may include not only information identifying the content used, but also the type of use, the time and method of collection and scraping, applicable retention practices, provenance, any alterations made to the content, and any third-parties who have access to the database or have already purchased it.

The goal should be to be able to construct a full chain of use. The creators of large data sets are presumably best placed to collect, retain, and disclose records regarding the information, materials, and sources included in the datasets they have built. Any downstream users, including developers, could then build on that information, and account for curation, editing, and other modification of material.

[194] Jeremy Fleming-Jones, *EU AI Act nearing agreement despite three key roadblocks – co-rapporteur*, EURONEWS, Oct. 23, 2023, https://www.euronews.com/next/2023/10/23/eu-ai-act-nearing-agreement-despite-three-key-roadblocks-co-rapporteur.

[195] DIGITAL CONTENT NEXT, EUROPEAN PUBLISHERS COUNCIL & NEWS/MEDIA ALLIANCE, JOINT LETTER ON AI (Oct. 19, 2023) available at http://www.newsmediaalliance.org/wp-content/uploads/2023/10/DCN-EPC-NMA-Joint-Letter-on-AI_US.pdf.

15.2. To whom should disclosures be made?

N/MA supports an open discussion about the most efficient solutions for disclosures concerning training data. In principle, at least for publicly facing datasets, a publicly accessible disclosure depository or clearinghouse could arguably minimize the costs and burden on dataset creators, developers, and copyright owners. A centralized solution may also be preferable as generative AI applications and datasets develop and proliferate. As an alternative or supplementary measure, an industry-led technical standard that would allow rightsholders to read disclosures automatically in addition to establishing a standardized way of organizing and finding information may be worth exploring. Disclosure obligations could consider appropriate differential treatment or exceptions for legitimate proprietary, trade secret, business confidential, or directly licensed material.

15.3. What obligations, if any, should be placed on developers of AI systems that incorporate models from third parties?

Developers who incorporate models from third parties into their systems or applications should be subject to the same transparency requirements as other developers.[196] Further policy formulation could consider permitting compliance to be made by disclosing the underlying models and datasets used with adequate links to the disclosures made by dataset creators regarding the use of copyrighted content in their datasets. The developers should also be responsible for disclosing material changes or additions they may have made to the third-party models or datasets that are relevant to such copyright-related aspects. The opposite result, creating different obligations based on an artificial hierarchy of AI developers, may facilitate circumvention and data laundering, undermining the purpose and efficacy of potential transparency and recordkeeping requirements.

15.4. What would be the cost or other impact of such a recordkeeping system for developers of AI models or systems, creators, consumers, or other relevant parties?

The development and adoption of generative AI transparency and recordkeeping requirements—and the scope and subjects of such requirements—must be policy and public interest-based. It is critical that AI technologies evolve with proper guardrails and safety protocols, including strong and enforceable recordkeeping obligations. Simply put, transparency and recordkeeping requirements, as well as applicability of a rights based licensing framework, should be a cost of doing business. The government should not effectively

[196] As noted in responses to questions 22-27, they may also risk direct or secondary legal liability for infringing uses of content to train those models.

subsidize generative AI developers at the expense of authors and publishers by not adopting transparency and recordkeeping rules necessary to enforce existing copyrights out of a desire to protect developers from compliance costs. This is particularly the case considering the scale of profits anticipated by large generative AI developers whose models are especially likely to compete with creative content.[197]

Further, recordkeeping requirements may carry additional benefits, including reducing legal uncertainty and risk for AI developers, and providing rights holders with a more efficient ability to protect their content and investments against unauthorized uses, and reach negotiated agreement. In the absence of adequate recordkeeping systems, enforcement and negotiations may be considerably more cumbersome, expensive, and time-consuming, rendering such actions out of reach for far too many publishers.

In addition, such measures would facilitate greater public trust in generative AI applications and their outputs—an increasingly important benefit as AI applications and systems proliferate and become more intertwined with people's lives. Transparency and recordkeeping requirements could facilitate efforts to analyze and combat biases in AI, increase national security by helping identify harmful data sources that drive or affect generative AI outputs, and serve a variety of other interests, ranging from consumer protection to financial regulation and consumer privacy.

16. What obligations, if any, should there be to notify copyright owners that their works have been used to train an AI model?

The basic principle of copyright law—discussed in response to Question 9—is that where permission is required, it should be obtained before the use is made.

However, in cases where developers of generative AI have not acquired a license before the training took place, the developer should have a duty to notify applicable rightsholders as soon as practicable. The burden should not be on copyright owners to undertake the expense to reverse engineer AI training datasets and conduct forensic analysis to learn whether and how their property was used.

[197] See, e.g., Richard Waters and Camilla Hodgson, *Microsoft's Edge in AI Pays Off While Google Tries to Catch Up in the Cloud*, Financial Times, Oct. 25, 2023, https://www.ft.com/content/b20f9491-34b5-409c-b084-68169be6638c; Arthur Sants, *AI Helps Microsoft Pull Ahead of Google*, Investors' Chronicle, Oct. 25, 2023, https://www.investorschronicle. co.uk/news/2023/10/25/ai-helps-microsoft-pull-ahead-of-google/; Deepa Seetharaman & Berber Jin, *OpenAI Seeks New Valuation of Up to $90 Billion in Share Sale*, Wall St. J. (Sep. 26, 2023) https://www.msn.com/en-us/money/companies/openai-seeks-new-valuation-of-up-to-90-billion-in-share-sale/ar-AA1hiJ9W.

As discussed in answers to Questions 15-15.4., transparency and recordkeeping systems can support potential notification obligations imposed on AI developers. In addition, publicly identifying training datasets or licenses, such as OpenAI's announcement about a licensing deal with Shutterstock,[198] and the creation of searchable databases of URLs and works used in training could increase general transparency around AI training.

17. Outside of copyright law, are there existing U.S. laws that could require developers of AI models or systems to retain or disclose records about the materials they used for training?

Other bodies of law may already impose record keeping or disclosure obligations on developers of AI models (including privacy, consumer protection, document retention, and antidiscrimination laws, such as fairness in lending obligations), and there is ongoing interest among lawmakers in whether it would be appropriate to amend those laws. For the purposes of this Notice of Inquiry, N/MA focuses solely on copyright-related issues in these comments.

Generative AI Outputs

Infringement

22. Can AI-generated outputs implicate the exclusive rights of preexisting copyrighted works, such as the right of reproduction or the derivative work right? If so, in what circumstances?

Yes, AI-generated outputs can infringe copyrighted works, including by violating the right of reproduction and the derivative work right. Existing legal doctrines relevant to copyright infringement can be used to analyze AI-generated outputs the same as other potentially infringing materials.

For example, well-settled legal principles establish that AI-generated outputs infringe the reproduction right in media articles and other literary works where outputs are comprised of verbatim content, and also may infringe where they contain close paraphrasing or closely detailed summaries and/or substantially similar structure and expression to the original work.[199] Our response to question 23 expands on this further.

[198] *See, e.g.,* Shutterstock, *Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data*, Jul. 11, 2023, https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year.

[199] *See, e.g., Wainwright Sec. Inc. v. Wall St. Transcript Corp.,* 558 F.2d 91, 95-96 (2d Cir. 1977) (affirming finding of infringement where summaries of Wall Street Journal articles appropriated "the manner of expression, the author's analysis or interpretation of events, the way he structures his material and marshals facts, his choice of words, and the emphasis he gives to particular developments"); *Associated Press v. Meltwater US Holdings, Inc.*,

For decades, modern media, publishing, distribution, licensing, and software business models and related transactions have been developed upon this shared understanding of these metes and bounds of copyright law. The advent and use of AI-generated outputs can and must be integrated into this shared legal framework to incentivize continued creativity and innovation.

AI-generated outputs can be examined whether they infringe the derivative work right even in cases where the output itself would not otherwise qualify for copyright protection because it is not the product of human authorship. As the Office has recently correctly noted, “the test for copyrightability and the test for infringement of the derivative-works right are distinct,” and “the derivative-works right is framed in terms of ‘preparation,’ indicating that non-human actions may be sufficient to infringe the right.”[200]

In addition to potentially giving rise to infringement claims, unauthorized use of copyrighted works can also preclude protection for AI-generated outputs, even assuming that there is sufficient human authorship attached to that work. Section 103 provides that “protection for a work employing preexisting material in which copyright subsists does not extend to any part of the work in which such material has been used unlawfully.” Moreover, the copyright in a “derivative work extends only to the material contributed by the author of such work, as distinguished from the preexisting material employed in the work, and does not imply any exclusive right in the preexisting material.”

23. Is the substantial similarity test adequate to address claims of infringement based on outputs from a generative AI system, or is some other standard appropriate or necessary?

Substantial similarity is the dominant test applied to determine whether there has been an infringement of a copyrighted work. This test can be applied to address whether outputs from a generative AI system are infringing, including outputs of paraphrasing tools such as Quillbot AI or AI-chatbot regurgitations of protected news media content, such as examples shown in the attached Technical Annex.

931 F. Supp. 2d 537 (SDNY 2013) (finding excerpting of AP news articles to be infringing and not fair use); *Warner Bros. Ent. Inc. v. RDR Books, 575 F.Supp.2d 513 (SDNY 2008)* (finding “Lexicon” of facts, summaries, and supplemental material drawn from the Harry Potter series was infringing and not fair use). In addition, the use of copyrighted works to create other, supplemental works infringes the copyright owner’s exclusive right to prepare derivative works. *Castle Rock Entertainment v. Carol Publishing Group*, 150 F. 3d 132 (2d. Cir. 1998) (affirming finding that “Seinfeld Aptitude Test” was an infringing derivative work that did not constitute fair use).

[200] Suzy Wilson & Rob Kasunic, Letter to ALI re Preliminary Draft No. (Sep. 26, 2023) available at https://www.copyright.gov/rulings-filings/restatement/comments/2023-09-26-Preliminary-Draft-No-9.pdf.

Additional judicial precedents have developed to help courts analyze questions of substantial similarity.[201] N/MA expects judicial doctrine to continue to evolve, including to provide any clarification necessary with respect to outputs from a generative AI system.

In the context of journalistic works and other writings published by N/MA members, including opinion, analysis, reviews, advice, investigations, and fictive works, judicial precedent is well-suited to address claims of infringement based on outputs from a generative AI system. It is black letter law that news reporting may be infringed by quoting too much of its content: the Supreme Court addressed this squarely in *Harper & Row,* holding that quoting 300-400 words verbatim from a 450-page biography was infringement, not fair use.[202]

With journalistic content, the line between copying copyrighted expression versus unprotectable facts has been frequently analyzed, and the right of news publishers to protect their copyrighted expression against overzealous borrowers repeatedly upheld. While a free Press itself depends upon facts remaining in the public domain,[203] U.S. copyright law has always aimed to incentivize the original expression of facts; the originating Copyright Act of 1790 was limited in scope to protect three types of works: books, maps, and charts.[204]

Courts navigate the facts/expression distinction by analyzing how expressive the copied material is. One illustrative case is *Salinger v. Random House,* where the Second Circuit reversed a finding by then-district Judge Leval that a biography of writer J.D. Salinger made fair use by paraphrasing letters from the famous author. The Second Circuit sharply disagreed with Judge Leval's weighing of the third fair use factor, the amount and substantiality of the portion used, noting that "protected expression has been 'used' whether it has been quoted verbatim or only paraphrased."[205] The appellate court updated the fair use analysis by considering both paraphrases and finding that the lower decision erroneously rejected claims of infringement because they employed "a cliche or a word-combination that is so ordinary that it does not qualify for the copyright law's protection." It explained:

> The "ordinary" phrase may enjoy no protection as such, but its use in a sequence of expressive words does not cause the entire passage to lose protection. And though the

[201] *See., e.g., Cavalier v. Random House, Inc.*, 297 F.3d 815, 822 (9th Cir. 2002); *Rentmeester v. Nike, Inc.*, 883 F.3d 1111, 1118 (9th Cir. 2018).

[202] *Harper & Row,* 471 U.S. at 569.

[203] 17 U.S.C. 102(b). *See, e.g. Narell v. Freeman*, 872 F.2d 907, 911 (9th Cir. 1989).

[204] U.S., Copyright Act of 1790 (1970); *See* U.S. Copyright Office, *The 18th Century* [n.d.], https://copyright.gov/timeline/timeline_18th_century.html (last visited Oct. 27, 2023).

[205] *Salinger v. Random House, Inc.,* 811 F.2d 90, 97-98 (2d Cir. 1987). The opinion also addresses that the Salinger letters were unpublished under the second factor, but status of publication was not relevant to the third factor analysis.

> "ordinary" phrase may be quoted without fear of infringement, a copier may not quote or paraphrase the sequence of creative expression that includes such a phrase. [The question is whether] the passage as a whole displays a sufficient degree of creativity as to sequence of thoughts, choice of words, emphasis, and arrangement to satisfy the minimal threshold of required creativity.[206]

Other cases draw similar conclusions. In *Wainwright,* the Second Circuit noted that although facts are not protectable, one may not take "the manner of expression, the author's analysis or interpretation of events, the way he structures his material and marshals facts, his choice of words, and the emphasis he gives to particular developments."[207] In *Robinson v. Random House, Inc.*, "approximately 25-30 percent of the words and phrases" were "used verbatim or through close paraphrasing" in an infringing book."[208] The court pointed to a side-by-side analysis to underscore how the defendant "went far beyond the use of mere facts contained in the [original book]—the appropriation included [the author's] expression " by taking "organization, writing style, even punctuation."[209] Similarly, when determining that a "Lexicon" of facts, summaries, and supplemental material drawn from the *Harry Potter* series was infringement and not a fair use, the court considered direct quotations, close paraphrases, and scene summaries, noting, "the law in this Circuit is clear that 'the concept of similarity embraces not only global similarities in structure and sequence, but localized similarity in language.'"[210]

The same analysis would apply in the generative AI context: a model's output need not replicate full passages to establish infringement, but a court may consider lengthy summaries, close paraphrases, verbatim excerpts, and whether the structure of the original work was lifted to determine substantial similarity.

24. How can copyright owners prove the element of copying (such as by demonstrating access to a copyrighted work) if the developer of the AI model does not maintain or make available records of what training material it used? Are existing civil discovery rules sufficient to address this situation?

[206] *Id.*
[207] *Wainwright Sec. Inc. v. Wall St. Transcript Corp.,* 558 F.2d 91, 95-96 (2d Cir. 1977) (affirming finding of infringement based on abstract summaries of Wall Street Journal articles). *See also Associated Press v. Meltwater US Holdings, Inc.*, 931 F. Supp. 2d 537 (SDNY 2013) (excerpts of AP news articles was infringing and not fair use).
[208] *Robinson v. Random House, Inc.*, 877 F.Supp. 830, 835 (S.D.N.Y. 1995) (finding use was infringing and not fair).
[209] *Id.* at 837-838.
[210] *Warner Bros. Ent. Inc. v. RDR Books*, 575 F.Supp.2d 513 (SDNY 2008).

To the extent that generative AI developers and deployers do not maintain adequate recordkeeping or retention practices or disclose them, existing discovery practices may not be sufficient or well-tailored to address these questions. Moreover, strong public policy considerations counsel against litigation as the place of first resort. In addition to conserving judicial economy, the discovery process can be time consuming, inefficient, and imperfect. N/MA refers to its responses to questions 15-17 concerning the need for adequate transparency and recordkeeping practices.

That said, existing legal rules are currently applicable, including the obligation to preserve evidence when a party should know that the evidence may be relevant to future litigation.[211] Given the multitude of copyright infringement and other lawsuits already commenced against generative AI companies, N/MA members believe similar developers are already under an obligation to preserve and eventually disclose records of what copyrighted materials they used in "training" their systems, how the training works, and what materials are retained.

25. If AI-generated material is found to infringe a copyrighted work, who should be directly or secondarily liable—the developer of a generative AI model, the developer of the system incorporating that model, end users of the system, or other parties? 25.1. Do "open-source" AI models raise unique considerations with respect to infringement based on their outputs?

Question 25, like other infringement-related questions, will have fact-dependent answers depending on the specific circumstances of infringement. Copyright liability is joint and several, and there may be more than one direct infringer, involved in different stages of the development, deployment, or use of a generative AI model. In addition, principles of secondary liability would also apply. *See, e.g., Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd.,* 545 U.S. 913 (2005). We are aware that many companies have announced intentions to indemnify certain end users against claims of copyright infringement related to the outputs generated by their generative AI models.[212]

[211] *See, e.g.,* Fed. R. Civ. Pro. 37(e) (providing for sanctions where a party failed to take reasonable steps to preserve electronically stored information in anticipation of litigation); *Fujitsu Ltd. v. Federal Exp. Corp*., 247 F. 3d 423 (2d Cir. 2001).

[212] *See* Brad Smith & Hossein Nowbar, *Microsoft Announces New Copilot Copyright Commitment for Customers*, MICROSOFT ON THE ISSUES, Sep. 7, 2023, https://blogs.microsoft.com/on-the-issues/2023/09/07/copilot-copyright-commitment-ai-legal-concerns/; Stephen Nellis, *Adobe Pushes Firefly AI into Big Business, with Financial Cover*, REUTERS, Jun. 8, 2023, https://www.reuters.com/technology/adobe-pushes-firefly-ai-into-big-business-with-financial-cover-2023-06-08/; Neal Suggs & Phil Venables, *Shared Fate: Protecting Customers with Generative AI Indemnification*, AI & MACHINE LEARNING, Oct. 13, 2023, https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification.

That said, we believe that, at a minimum, developers of generative AI models and the interfaces incorporating them are directly liable for their own infringing output. With respect to open-source practices, the reliance on open-source AI models or sources should not obviate the need to adhere to transparency or licensing obligations. Indeed, in other contexts, open-source licensing has been a valuable and flexible tool to facilitate the permissive use of a wide range of copyrighted content--working within, as opposed to against, the overall legal framework of copyright. To the extent some users of open source material may be confused, and think that open source material is not subject to copyright protections (including publisher content incorporated therein), the Copyright Office should educate to clarify this folk misconception.

N/MA would be particularly concerned by attempts to otherwise skirt responsibility by designing conditions for "divided infringement" to escape liability for acts that would otherwise be infringing. To be sure, open-source AI models like LLAMA2 appear to have a direct financial interest in the use of its models by downstream commercial actors, as well as the right and ability to supervise its licensees.

As the marketplace and legal landscape continue to develop, the Copyright Office can consider whether guidance or recommendations may be needed to avoid incentives that shift responsibility away from the developers of generative AI models who are typically best placed to bear those compliance obligations and make it more difficult for copyright holders to effectively enforce their rights.

26. If a generative AI system is trained on copyrighted works containing copyright management information, how does 17 U.S.C. 1202(b) apply to the treatment of that information in outputs of the system?

Section 1202(b) is intended to prevent the deliberate stripping of copyright management information (CMI) to facilitate infringement. In enacting section 1202, Congress noted that CMI is essential to "establishing an efficient Internet marketplace" by facilitating the tracking and monitoring of copyright uses as well as licensing agreements.[213] But as the Copyright Office previously noted in its study on Moral Rights, the precise dual scienter standard was strongly debated in international fora when the related WIPO Treaties were negotiated, and decades later, many contend this standard has impeded the practical usefulness of section 1202 to protect an author's attribution's rights.[214] The Office therefore recommended a legislative

[213] THE REGISTER OF COPYRIGHTS, THE COPYRIGHT OFFICE, REPORT: AUTHORS, ATTRIBUTION, AND INTEGRITY: EXAMINING MORAL RIGHTS IN THE UNITED STATES (Apr. 2019) available at https://www.copyright.gov/policy/moralrights/full-report.pdf.
[214] *Id*. at 93-98.

amendment to this standard, which N/MA believes would be a good step.[215] The Office has also expressed concern over interpretations, like the Ninth Circuit's *Core Logic* opinion, that would raise this knowledge bar even higher.[216]

In the context of generative AI, removal of CMI can hinder the determination whether a copyrighted work has been ingested in the training process and inhibit complete and accurate recordkeeping activities. And many recent litigations around generative AI products and services have involved claims under section 1202, including *Anderson v. Stability AI*, *Doe v GitHub*, *Tremblay v. OpenAI, Inc., Silverman v Open AI, Inc.* and *Getty Images v. Stability AI*. For example, one currently active docket, *Doe v. GitHub,* involves the use of automated removal of metadata from open-source computer code used to train generative AI tools offered by Microsoft and OpenAI, where such tools "were not programmed to treat attribution, copyright notices, and license terms as legally essential."[217]

The Office should build upon its previous analyses of section 1202 and encourage legal interpretations and, if necessary, legislative reforms that allow for a balanced law regarding removal of CMI. It should discourage reckless practices like automated metadata stripping for purposes of ingesting copyright-protected works into generative AI models.

<u>Labeling or Identification</u>

28. Should the law require AI-generated material to be labeled or otherwise publicly identified as being generated by AI? If so, in what context should the requirement apply and how should it work?

This is a complicated question that has wider implications beyond copyright law, including potential First Amendment considerations, and the Copyright Office should exercise caution if it decides to address this issue. If any labeling requirements are adopted, appropriate agencies, including the FTC and USPTO should be consulted, and they must not be one-size-fits-all but rather should recognize the variety of AI-generated uses and be appropriately narrowly tailored. As a starting point, labeling disclosures should not apply to instances where a human person reviews and edits content that was assisted by generative AI, and remains legally liable and editorially responsible for the content. The level and format of any labeling disclosures should also be carefully considered as labeling that works for a certain type of creative work

[215] *Id.* at 98.

[216] *Id.* at 96, *citing Stevens v. Corelogic, Inc.,* 899 F.3d 666 (9th Cir. 2018), cert denied, 586 U.S. __ (U.S. Feb. 19, 2019) (No. 18-878).

[217] *Doe V. Github Inc,* No. 22-Cv-06823-Jst, 2023 U.S. Dist. (N.D. Cal. May 11, 2023) available at https://caselaw.findlaw.com/court/us-dis-crt-n-d-cal/2200493.html.

may not work for another—for example, while AI-generated photographs could be watermarked, repeated pop-ups identifying AI-generated scenes or components may seriously disrupt an audiovisual experience.

The Office could facilitate stakeholder dialogues within and between industries to facilitate the development of marketplace standards, and consider whether consultation with additional agencies on matters adjacent to copyright, such as USPTO or FTC, would be beneficial.

Additional Questions About Issues Related to Copyright

32. Are there or should there be protections against an AI system generating outputs that imitate the artistic style of a human creator (such as an AI system producing visual works "in the style of" a specific artist)? Who should be eligible for such protection? What form should it take?

U.S. copyright law does not protect the "style" of a specific creator *per se*, although in some instances characters and other motifs can be protected when they are significantly distinctive and unique. As exemplified by *Steinberg v. Columbia Pictures Industries, Inc.*, the line between "style" and expression is not always clear and easy to draw.[218] Finding and preserving the appropriate balance is important for creative expression to flourish and to provide sufficient legal certainty to both original and secondary creators alike.

Related to, but separate from the specific questions posed by the Office, news, magazine, and digital media publishers are concerned about the potential of generative AI models and applications to misrepresent the source of information or the sources of other goods and services in violation of interests of trademark owners. N/MA is also concerned by the ability of generative AI to create outputs in the style of a media outlet or a high-profile journalist or other contributor or content creator while misattributing the content to said media or individual. Such misrepresentations may implicate—and potentially require changes to—other areas of law, including the Lanham Act, right of publicity, or other related laws. Absent effective ways to combat these misattributions, publishers of all types and sizes risk reputational, brand, and financial harms caused by mis- or disinformation they have not published nor generated.

34. Please identify any issues not mentioned above that the Copyright Office should consider in conducting this study.

[218] *Steinberg v. Columbia Pictures Industries, Inc*. 663 F. Supp. 706 (S.D.N.Y. 1987).

N/MA recommends the Copyright Office consider three policy recommendations and initiatives not explicitly raised by its Notice, namely publishers' ability to register online web content by submitting identifying material, the Journalism Competition and Preservation Act, and voluntary guidance and the facilitation of industry-led solutions.

First, most importantly--and urgently--the Copyright Office should adopt regulations to enable publishers to group register online web content in an efficient, economical, and simple manner. Currently, publishers are effectively unable to register their online-only content as there is no group registration option allowing for the registration of groups of frequently updated website content. We understand the constraints of the legacy eCO registration system impede the Office's ability to nimbly update the registration options it offers the public. But for news publishers, registering each individual online article under existing registration options would be burdensome, economically punitive, and contrary to the general goals of the registration system. As AI developers exclusively use online content to train their models and applications, publishers' inability to adequately register their copyrights has wide-reaching consequences to their ability to enforce their rights, monetize their content, and continue investing in the production of high-quality original content.

N/MA urges the Office to swiftly adopt regulations to enable publishers to group register news website content in an efficient manner. We are encouraged by recent suggestions that the Office has identified a solution that the eCO system may accommodate and recommend immediate adoption of this solution on at least an interim, pilot basis, and then examination to see if subsequent updates are required (including when a modernized registration system comes to fruition). We support an option to facilitate the registration of publisher owned copyrightable content on a website at a designated period of time, subject to verification. Considering the substantial market harms that systemic, unauthorized scraping for AI purposes may cause, N/MA believes that the registration option should be construed to allow publishers to seek statutory damages for the infringement of each article or other work copied. Regardless, we welcome creative thinking from the Office to introduce an updated option within the eCO system. We thank the Office for its attention to this matter and our members are ready to provide any business or technical information that would be helpful.

Second, the Office could recommend that Congress consider the passage of the Journalism Competition and Preservation Act (JCPA). The Office previously highlighted JCPA as a potential competition law-based solution to the issue of systemic unauthorized use of publisher content by the dominant online platforms, including in connection with generative AI, examined in more detail in the Office's Study on Ancillary Copyright Protections for Publishers.[219] While our comment here focuses on copyright concerns, attention to competition issues should also be given to ensure market conditions facilitate adequate compensation for use of publishers' valuable expressive material.

In its Study, the Office acknowledged that "economic trends in the news industry all point to a sea change in the press publishing ecosystem, with especially damaging consequences for local newspapers."[220] AI poses a similar existential challenge to publishers of all types and sizes, requiring an array of policy, technical, and regulatory solutions, while publishers meanwhile remain challenged by the existing practices of dominant platforms. N/MA understands that antitrust solutions are outside the scope of the Office's purview but would encourage the Office to mention such options as potential non-copyright tools in your Study.

Third, the N/MA recommends that the Copyright Office consider facilitating stakeholder dialogues in order to develop voluntary guidance documents, policy recommendations, and toolkits—similar to the NTIA's work as part of the Biden-Harris Administration's Task Force on Kids Online Health & Safety.[221] Relatedly, the Office may wish to establish a standing consultative group to ensure it can keep pace with generative AI developments as its study processes. Convening such dialogues would encourage market-led solutions that could form a significant part of a sustainable approach to AI development that protects and values publishers' copyrights and contributions to the economy and establishes a healthy growth environment for continued generative AI development.

Respectfully submitted,

Regan Smith

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

[219] "Should Congress wish to explore non-copyright measures for supporting journalism, the comments on this Study offered several proposals, including the JCPA, a levy on digital advertising revenue, increased public funding, or tax breaks for journalism. All of these proposals, however, lie beyond the expertise of the Copyright Office, and we make no findings on their merits." Press Publishers Study at 59.

[220] *Id.*

[221] NTIA, *Press Release, NTIA Seeks Comment on Protecting Kids Online*, UNITED STATES DEPARTMENT OF COMMERCE (Sep. 28, 2023), https://www.ntia.gov/press-release/2023/ntia-seeks-comment-protecting-kids-online.

APPENDICES ATTACHED

A. 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 (Oct. 2023).

B. News Media Alliance, Comments in Response to Request for USPTO Comments on Intellectual Property Protection for Artificial Intelligence Innovation (Jan. 2020)

C. European Magazine Media Association & European Newspaper Publishers Association, EMMA-ENPA's Core Concerns on AI and Copyright (Jul. 2023)

APPENDIX A

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

TABLE OF CONTENTS

I. Executive Summary 1

II. Introduction 3

III. Who We Are 6

IV. Large Language Models 7

A. LLMs don't learn or reason about facts. 8

B. GAI applications substitute for training works. 13

C. LLMs are built on unauthorized copying 17

D. LLMs retain copyrighted expressive content. 20

V. GAI Copying Is Not "Fair Use" 23

A. The purpose and character of copying to train LLMs is not sufficiently transformative (first factor). 24

1. Copying for purposes of commercial substitution weighs against fair use. 24

2. GAI developers copy news and digital media content to extract and replicate its expressive content. 25

3. LLM and chatbot uses are highly commercial. 28

4. There is no satisfactory independent justification for the copying. 29

5. The unlicensed use of training materials serves a system designed to produce substitutional outputs. 32

B. The effect of GAI copying on the market for publisher content is predictable and real (fourth factor) 33

C. GAI copying takes substantial portions of expressive works in their entirety (second and third factors). 36

VI. Recommendations 37

Technical Appendix 39

I. Executive Summary

This White Paper is published by the News/Media Alliance (N/MA) to address the rampant copying of its members' expressive works to train generative artificial intelligence (GAI) systems.[1] N/MA member newspaper, magazine, and digital media publishers speak with a collective voice in supporting the responsible development of GAI while ensuring fair credit and compensation for the creators whose works make GAI possible. N/MA members welcome working with GAI developers to help build and grow this exciting new technology, in ways that can benefit all actors and society at large.

GAI systems, while holding promise for consumers, businesses, and society at large, are commercial products that have been built—and are run—on the backs of creative contributors. These systems have been developed by copying massive amounts of the creative output of the Alliance's members, almost always without authorization or compensation. And they disseminate the same kind of content for the same commercial purpose—sometimes in the same or substantially similar form—in response to user queries, again without authorization or payment and often with little or no attribution or link to the original source. Such disassociated output diminishes the need for users to click through or subscribe to N/MA members' print and digital publications. This irreparably damages publishers' businesses, which depend on relationships with their readers, web traffic, and the trustworthiness of brands built over decades.

An analysis commissioned by the News/Media Alliance shows that GAI developers disproportionally use online news, magazine, and digital media content to train their GAI models. Their affinity for this quality content highlights its value and expressive nature. The analysis demonstrates:

- GAI developers create curated sets of training data to build Large Language Models (LLMs), which then power GAI products. We have analyzed the data sets used to build these models and the output that they generated, and that analysis demonstrates that the developers have copied and used news, magazine, and digital media content to train the LLMs.

- In fact, our analysis of a representative sample of news, magazine, and digital media publications shows that the popular curated datasets underlying some of the most widely used LLMs significantly overweight 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.

- Other studies show that news and digital media ranks third among all categories of sources in Google's C4 training set, which was used to develop Google's GAI-powered search capabilities and products like Bard. Half of the top ten sites represented in the training set are news outlets.

[1] In addition to counsel at the News/Media Alliance, this paper was co-authored by Cynthia S. Arato, Shapiro Arato Bach LLP, and Ian B. Crosby, Susman Godfrey LLP.

- The LLMs also copy and use publisher content in generating outputs. The LLMs can reproduce the content on which they were trained, demonstrating that the models retain and can memorize the expressive content of the training works.

This pervasive copying infringes N/MA members' exclusive rights in their copyrighted works and is not excused by the fair use doctrine, as the two most important fair use factors (the purpose and character of the use and the effect of the use on the market for the original) demonstrate:

- The GAI copying for "training" does not serve a purpose different from the original works because LLMs typically ingest (i.e., copy) valuable news, magazine, and digital media web content for their written expression, so that they can mimic that very form of expression. As one GAI proponent has explained, LLMs that are trained to generate their own expressive works "copy expression for expression's sake." Training LLMs on reliable, trusted expressive content without authorization also seeks to override licensing markets that already exist for these works, and copying for these training purposes thus serves (and supplants) that same licensing purpose. The GAI uses are also overwhelmingly commercial, helping to propel the GAI companies' valuations into the billions. And there is no compelling justification to allow the copying of creative works without fairly compensating the creators.

- The outputs of GAI models also directly compete with the protected content that was copied and used to train them. The use of these models to provide complete narrative answers to prompts and search queries goes far beyond the purpose of helping users to navigate to original sources (i.e., search) that has been found in the past to justify the wholesale copying of online content to build search engines. Indeed, GAI developers boast that users no longer need to access or review such sources. In this setting, the GAI developers' goal to create large language models, however laudable, does not justify their infringement of this valuable corpus of copyrighted expression.

While GAI developers contend that GAI models are just "learning" unprotectable facts from copyrighted training materials, that anthropomorphic claim is technically inaccurate and beside the point. It is inaccurate because models retain the expressions of facts that are contained in works in their copied training materials (and which copyright protects) without ever absorbing any underlying concepts. It is beside the point because materials that are used for "learning" are subject to copyright law. Even libraries must legally acquire the books they lend, and borrowers aren't free to copy them, especially not for an ultimate commercial use.

The incipient and predictable consequence of GAI's substitutive uses will be to damage the news and digital media industry. And it is not just copyright owners but society that will lose if GAI is allowed to so harm the journalism industry. Indeed, if the Internet becomes flooded with the products of GAI, then GAI itself will have nothing left to train on.

But GAI developers and publishers can work together to avoid such dire results. Indeed, publishers welcome technological progress and rely every day on innovative tools to tell their stories and inform the public, particularly where stories need to be globally transmitted and reported in real time through increasingly visual storytelling. N/MA members thus wish to work with GAI developers to maximize the value of this exciting new technology, in a way that is fair to publishers and equitably shares the wealth generated from the N/MA content that the GAI developers copy and redeploy. Such fruitful cooperation between the GAI developers and the owners of these source works will benefit not just the news and media industries but the GAI developers and society at large, by helping to ensure that GAI is developed using high-quality and human created works.

Our culture, our economy, and our democracy require a solution that allows the news and media industry to grow and flourish, and both to share in the profit from and participate in the development of the GAI revolution that is being built upon the fruits of its labor. Part of this solution is offered by copyright law, which exists to ensure that creators and content owners are appropriately compensated for their copyrighted works and to incentivize the continued creation of such works, for the benefit of society at large.

This White Paper concludes with several recommendations: (1) GAI developers must be transparent and open about their use of expressive works in GAI models; (2) industry and policymakers must understand that unauthorized use of expressive works to train LLMs that are designed to generate expressive text in a commercial context is infringing; and (3) publishers must be able to license the use of their content efficiently and on fair terms.

II. Introduction

Generative artificial intelligence technologies can now mimic nearly any kind of work that humans create at vastly greater speed and lower cost—and at massive scale. Even the most enthusiastic proponents admit that GAI is *designed* to substitute for human creations: it has, they boast, "produced writing that's difficult to distinguish from real journalists, painted in the style of celebrated masters, and even created stock photos comparable to those of professional photographers."[2]

The ability of GAI to imitate and copy human expression quickly and cheaply brings opportunities with the potential to benefit society and greatly enrich the developers of these models. But popular models like ChatGPT can do so only because they have been trained on the fruits of human creativity at massive scale, and largely without consent or compensation. The works these models can imitate and copy in this way include prize-winning landmarks of culture produced at great cost to news, magazine, and digital publishers—and often at great peril to the journalists they employ.

While publishers have retrenched to survive in the Internet age, companies that develop foundational GAI models trained on these works have by contrast seen their valuations explode.[3] Platforms that deploy these GAI models into their products have likewise seen their market

[2] Mark A. Lemley & Brian Casey, *Fair Learning*, 99 Tex. L. Rev. 743, 767 (2021).

[3] *See*, *e.g.*, Cade Metz, *OpenAI in Talks for Deal That Would Value Company at $80 Billion*, N.Y. Times (Oct. 20. 2023), https://www.nytimes.com/2023/10/20/technology/openai-artifical-intelligence-value.html; Jagmeet Singh & Ingrid Lunden, *OpenAI Closes $300M Share Sale at $27B-29B Valuation*, TechCrunch (Apr. 28, 2023), https://techcrunch.com/2023/04/28/openai-funding-valuation-chatgpt/.

capitalizations soar.[4] Yet even though established markets exist for providing and licensing media content in a variety of contexts—including machine learning—almost none of this wealth has flowed to the rights holders of the writings whose wholesale copying fuels the capabilities of these immensely valuable GAI products.

The members of the News/Media Alliance are deeply concerned about this unauthorized and unlawful use of their expressive content by large technology companies. Such companies do not shoulder the cost or risk of reporting the news or producing creative content but capitalize on that valuable work. Indeed, publishers generally are not being paid by GAI developers for the unauthorized copying of their works to train the LLM models on which their chatbots are built. And those chatbots, like Bing Chat, Bard, ChatGPT, and Claude are often deployed to compete directly with those very works by, for example, providing narrative answers to search queries that obviate the need for consumers to click through to the original sources whose content permeates those responses.

In addition to chatbot applications, the newest generation of up-to-the-minute narrative search results, in particular by GAI applications like Google's Search Generative Experience and Microsoft's Bing Chat, exceeds any previously adjudicated limits of permissible use in the field. Such full and expressive responses directly compete with publisher content, sever publishers' connections to their readers, and bypass the very presence of their sites on the Internet. Indeed, Microsoft markets Bing as where to go to "Ask Real Questions. Get Complete Answers. Chat and Create."[5] Google's new "Search Generative Experience" has been described as a "plagiarism stew."[6]

As the accompanying technical analysis shows, the models also produce unauthorized derivative works by responding to user queries with close paraphrasing or outright repetition of copied and memorized portions of the works on which they were trained.

As with past "disruptive" Silicon Valley models, GAI investors are banking on forgiveness instead of asking permission. They depend on the claim that copying for training is a "fair use" that they may continue with impunity, even as many of their products directly compete with and threaten

[4] Marvie Basilan, *Microsoft Gets Stock Boost After Morgan Stanley's AI-Driven $3 Trillion Valuation Outlook*, Int'l Bus. Times (July 7, 2023), https://www.ibtimes.com/microsoft-gets-stock-boost-after-morgan-stanleys-ai-driven-3-trillion-valuation-outlook-3703880#:~:text=According%20to%20Morgan%20Stanley%2C%20Microsoft%20has%20a%2022%25,the%20company%20to%20hit%20a%20%243%20trillion%20valuation ("Microsoft has a 22% upside potential due to its 'pole position' in the generative AI race and this could propel the company to hit a $3 trillion valuation.").

[5] https://www.bing.com/new. As Microsoft admitted when it heralded the "new" Bing, it copies publisher content and delivers substitutional summaries: "There is no need to get overwhelmed sifting through search results. Bing distills the latest information from across the web to summarize and cite answers to your question. Microsoft.com, https://web.archive.org/web/20230710180333/https://www.microsoft.com/en-us/bing?form=MW00X7 (as of July 10, 2023).

[6] Avram Piltch, *Plagiarism Engine: Google's Content-Swiping AI Could Break the Internet*, Tom's Hardware (June 11, 2023), https://www.tomshardware.com/news/google-sge-break-internet.

the continued well-being of publishers. But fair use does not work this way. Indeed, the Supreme Court just ruled in *Andy Warhol Foundation for the Visual Arts v. Goldsmith* that even in the case of a highly creative adaptation, a use that has the potential to serve as a commercial substitute for an original work undermines a finding of fair use.[7] Simply having "some further purpose, in the sense that copying is socially useful," or "add[s] something new … does not render such uses fair."[8] *Warhol Foundation* recognizes that substitutive uses, however innovative, undermine the "economic incentive to create original works, which is the goal of copyright."[9]

The modes of distribution and consumption of publisher content are rapidly changing in the digital age, and the systematic copying and use of publisher content to fuel GAI systems and applications and to disseminate competing content poses what could be an existential threat to far too many publishers and is not a fair use. By diverting readers and the digital advertising dollars that follow them away from original sources, and by interfering with a potential source of licensing revenue for granting permissions, GAI models disincentivize investment in creation of those sources in the first place.

The continued unlicensed use of reporting also disserves the public interest: an online world that is dominated by GAI-generated, substitutional content is poised to leave the public with watered-down, less reliable outputs and fewer news outlets with the resources necessary to provide critical original reporting. As district court judge Denise Cote's decision in *Associated Press v. Meltwater U.S. Holdings, Inc.* explained with respect to direct scraping of news content that is economically indistinguishable from that now being laundered through GAI systems, copyright law should not allow for democracy to be imperiled in this manner:

> [T]he world is indebted to the press for triumphs which have been gained by reason and humanity over error and oppression … Permitting Meltwater to take the fruit of AP's labor for its own profit, without compensating AP, injures AP's ability to perform [its] essential function of democracy.[10]

GAI is now further threatening the ability of journalists and publishers to perform that "essential function of democracy." At a time when governments and experts around the world warn of the risk AI poses to democratic functioning,[11] it is critical that the copyright laws continue to protect publisher content to help safeguard the indispensable role of a flourishing and free press.

[7] *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 143 S. Ct. 1258, 1276-77 (2023).

[8] *Id.* at 1275.

[9] *Id.* at 1278.

[10] *Associated Press v. Meltwater U.S. Holdings, Inc*., 931 F. Supp. 2d 537, 553 (S.D.N.Y. 2013).

[11] *See*, *e.g., Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People,* Off. Sci. & Tech. Pol'y, https://www.whitehouse.gov/ostp/ai-bill-of-rights/; Mekela Panditharatne & Noah Giansiracusa, Brennan Ctr. for Just., How AI Puts Elections at Risk — And the Needed Safeguards (July 21, 2023), https://www.brennancenter.org/our-work/analysis-opinion/how-ai-puts-elections-risk-and-needed-safeguards; Dan Milmo & Kiran Stacey, *AI-Enhanced Images* a *"Threat* to *Democratic*

III. Who We Are

The News/Media Alliance is a nonprofit organization that represents the interests of more than 2,200 news media organizations in the United States and internationally, including newspaper, magazine, and digital publishers. The Alliance represents the unified voice of the industry and diligently advocates before the federal government on issues that affect today's media organizations, including protecting publishers' intellectual property.

News media publications play a crucial role in the U.S. economy and democracy. Every day, their publishers invest in high-quality journalism that keeps our communities informed, holds those in power accountable, and supports the free flow of information and ideas in society. Without free and flourishing news media, our society would be less well-off and less informed. However, publishers' ability to continue serving as an essential source of news for readers around the world depends on their ability to receive fair compensation for the original expressive content that they have developed at high cost.

The news, magazine, and digital media industries' contribution to the U.S. economy and society is considerable, with estimated revenues of newspaper and magazine publishers amounting to approximately $45 billion.[12] Newsrooms were estimated to directly employ approximately 31,000 people in 2020, not including additional indirect employment effects, while magazines employed over 73,000 directly and supported a total of over 219,000 jobs in 2021.[13] Employment in digital-native newsrooms, meanwhile, has increased from approximately 7,400 in 2008 to over 18,000 in 2020.[14]

Journalists and others who rely on print and digital media for their living create content that reaches 136 million adults in the United States each week, representing 54% of the country's adult population.[15] Globally, news organizations receive over 200 million unique visits and 6.7 billion

Processes", Experts Warn, The Guardian https://www.theguardian.com/technology/2023/aug/03/ai-enhanced-images-a-threat-to-democratic-processes-experts-warn.

[12] *See* Pew Rsch Ctr., Newspapers Fact Sheet (June 29, 2021), http://www.journalism.org/fact-sheet/newspapers/; Amy Watson, *Estimated Aggregate Revenue of U.S. Periodical Publishers from 2005 to 2020*, Statista, Dec. 5, 2022, https://www.statista.com/statistics/184055/estimated-revenue-of-us-periodical-publishers-since-2005/; Adam Grundy, *Service Annual Survey Shows Continuing Decline in Print Publishing Revenue*, U.S. Census Bureau, Jun. 7, 2022, https://www.census.gov/library/stories/2022/06/internet-crushes-traditional-media.html.

[13] Pew Rsch Ctr., *supra* note 12; Mason Walker, *U.S. Newsroom Employment Has Fallen 26% since 2008*, Pew Rsch Ctr., Jul. 13, 2021, https://www.pewresearch.org/short-reads/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/; MPA-The Association of Magazine Media, Magazine Media Factbook, (2021), https://www.newsmediaalliance.org/wp-content/uploads/2018/08/2021-MPA-Factbook_REVISED-NOV-2021.pdf.

[14] Pew Rsch Ctr., *supra* note 12.

[15] News/Media Alliance, News Advertising Panorama: A Wide-Ranging Look at the Value of the News Audience 72 (2019).

page views per month online.[16] News publishers also ensure the health of our local communities and play a vital role in civic discourse, investigating and exposing public corruption, wasteful governmental activities, worker safety violations, and other matters of public interest, with most local news media companies reaching more adults in their local markets than any other local media.[17]

The numbers on the prior page take on a different meaning when you consider that in less than 20 years, newspaper circulation and advertising revenues dropped from $57.4 billion in 2003 to an estimated $20.6 billion in 2020, while magazines witnessed a drop from $46 billion in 2007 to $23.92 billion in 2020.[18] While there have been increases in digital audience and advertising revenues in recent years,[19] print circulation of news dropped by approximately six percent from 2019 to 2020.[20] Moreover, because of existing marketplace imbalances,[21] digital revenues are not yet enough to offset the reduced print advertising and decline in print subscription revenues. GAI threatens to pluck even these green shoots of recovery, further skewing the distribution of online revenue towards technology platforms and resuming the march toward destruction of the news and media publication industry.

IV. Large Language Models

This paper is focused on "Large Language Models" and related GAI products which threaten to supplant online news media. LLMs are trained to predict the next word that is likely to follow a given string of words, or "prompt," which allows the models to generate longer strings of text that approximate human language.[22] There is no question that creating such models relies on copying—indeed, many rounds of copying—of third party works, such as the protected expression of our members.

To train a model to produce text that approximates natural human language in this way requires "training" with an enormous volume of examples. The life cycle of such an LLM begins with an "input" phase processing potentially billions of training works running into the trillions of words. To obtain such volume, the developers of these models appear to have made copies of a substantial

[16] *Id.*

[17] *Id.* at 72, 82.

[18] Pew Rsch Ctr., *supra* note 12; Watson, *supra* note 12.

[19] News/Media Alliance, *supra* note 15; Pew Rsch Ctr., *supra* note 12.

[20] Pew Rsch Ctr., *supra* note 12.

[21] *See generally* News/Media Alliance, How Google Abuses Its Position as a Market Dominated Platform to Strong-Arm News Publishers and Hurt Journalism (Sept. 2022) ("Google White Paper"), http://www.newsmediaalliance.org/wp-content/uploads/2022/09/NMA-White-Paper_REVISED-Sept-2022.pdf.

[22] David Nield, *How ChatGPT and Other LLMs Work—And Where They Could Go Next*, Wired (Apr. 30, 2023), https://www.wired.com/story/how-chatgpt-works-large-language-model/.

portion of the Internet, including paywalled material.[23] They make these copies either by scraping them directly from web sites or copying them from archives of copied content, like Common Crawl, created by others who have done the scraping. After their initial "pre-training," models may be "fine-tuned" with additional copied sources selected to improve performance for desired subjects or tasks.[24] Publisher content accounts for a substantial volume of the known sources for LLM training.[25]

A. LLMs don't learn or reason about facts.

While GAI developers often conceal[26] the inner workings and content of their large language models, the basic idea behind the models is simple. Often referred to in the AI field as "stochastic parrots,"[27] they function as mimics, able to reproduce expression taken from the mountains of material that GAI companies often copy without compensation or consent. They do so via mathematical equations that predict, based on the previously ingested expression, the most likely word to come next in a sentence given all the words that have preceded it.[28]

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 that are scraped and copied from the Internet without authorization. Rather, the models allow GAI products to produce outputs of expression that just mimic the content and style of the models' training sources through a process akin to following a kind of "map" of the semantic and syntactic relationships among the words in those sources.[29] The outputs are not thoughtful answers or the result of "learning" or "training"; they are dictated by

[23] *Artificial Intelligence Is Reaching Behind Newspaper Paywalls*, The Economist (Mar. 2, 2023), https://www.economist.com/business/2023/03/02/artificial-intelligence-is-reaching-behind-newspaper-paywalls (Bing's AI can paraphrase content of New York Times article blocked by a paywall).

[24] Tom B. Brown et al., *Language Models Are Few-Shot Learners* 6 (July 22, 2020), https://arxiv.org/abs/2005.14165 ("Fine-Tuning (FT) has been the most common approach in recent years, and involves updating the weights of a pre-trained model by training on a supervised dataset specific to the desired task. Typically thousands to hundreds of thousands of labeled examples are used."); Banghua Zhu et al., *Fine-Tuning Language Models with Advantage-Induced Policy Alignment* (June 8, 2023), https://arxiv.org/abs/2306.02231 (discussing pre-training and fine tuning).

[25] *See infra* Section IV.C.

[26] Saurabh Bagchi, *Why We Need to See Inside AI's Black Box*, Sci. Am. (May 26, 2023), https://www.scientificamerican.com/article/why-we-need-to-see-inside-ais-black-box/ ("[T]o protect their intellectual property, AI developers often put the model in a black box.").

[27] Muhammad Saad Uddin, *Stochastic Parrots: A Novel Look at Large Language Models and Their Limitations*, Towards AI (Apr. 13, 2023), https://towardsai.net/p/machine-learning/stochastic-parrots-a-novel-look-at-large-language-models-and-their-limitations.

[28] Nield, *supra* note 22.

[29] *See generally* Stephen Wolfram, What Is ChatGPT Doing ... and Why Does It Work? (Feb. 14, 2023), https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/.

the expression that the models previously ingested plus an element of randomness applied to the equations.[30]

The propensity of GAI models to generate false information, or "hallucinate," demonstrates that they are constructing sentences word by word based on their copied references. For example, GAI systems have: (1) provided fake case law in response to a lawyer's query, causing two lawyers to be sanctioned by a federal court;[31] (2) falsely stated that individuals have been indicted for sedition, accused of sexual harassment, or imprisoned for bribery;[32] and (3) provided false answers when asked for examples about chatbot hallucinations.[33] The GAI systems also have generated false statements regarding the reporting done by N/MA publishers, misrepresenting the contents of such reports and generating entirely false accounts of non-existent reports.

For example, Bing Chatbot falsely stated that The New York Times' review of *A Doll's House* described Jessica Chastain's performance as "a bit too studied and self-conscious," when the review did not include that negative information (nor did it state that the performance was "never less than compelling")[34]:

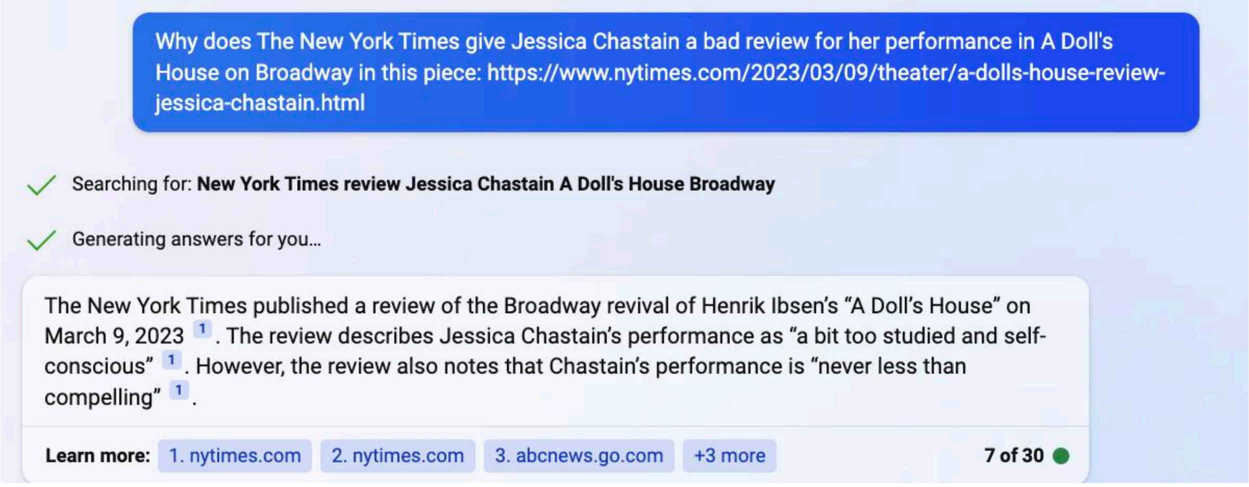
[30] Nield, *supra* note 22.

[31] Sara Merken, *New York Lawyers Sanctioned for Using Fake ChatGPT Cases in Legal Brief*, Reuters (June 26, 2023), https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/.

[32] Pranshu Verma & Will Oremus, *ChatGPT Invented a Sexual Harassment Scandal and Named a Real Law Prof as the Accused*, Wash. Post (Apr. 5, 2023), https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/; Byron Kaye, *Australian Mayor Readies World's First Defamation Lawsuit Over ChatGPT Content*, Reuters (Apr. 5, 2023), https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/; Eugene Volokh, *New Lawsuit Against Bing Based on Allegedly AI-Hallucinated Libelous Statements*, Volokh Conspiracy (July 13, 2023), https://reason.com/volokh/2023/07/13/new-lawsuit-against-bing-based-on-allegedly-ai-hallucinated-libelous-statements/ (Bing incorrectly stating aerospace professor pleaded guilty to seditious conspiracy and levying war against the United States).

[33] Cade Metz, *What Makes A.I. Chatbots Go Wrong?*, N.Y. Times (Mar. 29, 2023), https://www.nytimes.com/2023/03/29/technology/ai-chatbots-hallucinations.html (when asked for examples of chatbots hallucinating, Bing hallucinated the answer).

[34] Jesse Green, *Review: Jessica Chastain Plots an Escape From 'A Doll's House'*, N.Y. Times (Mar. 9, 2023), https://www.nytimes.com/2023/03/09/theater/a-dolls-house-review-jessica-chastain.html.



Why does The New York Times give Jessica Chastain a bad review for her performance in A Doll's House on Broadway in this piece: https://www.nytimes.com/2023/03/09/theater/a-dolls-house-review-jessica-chastain.html

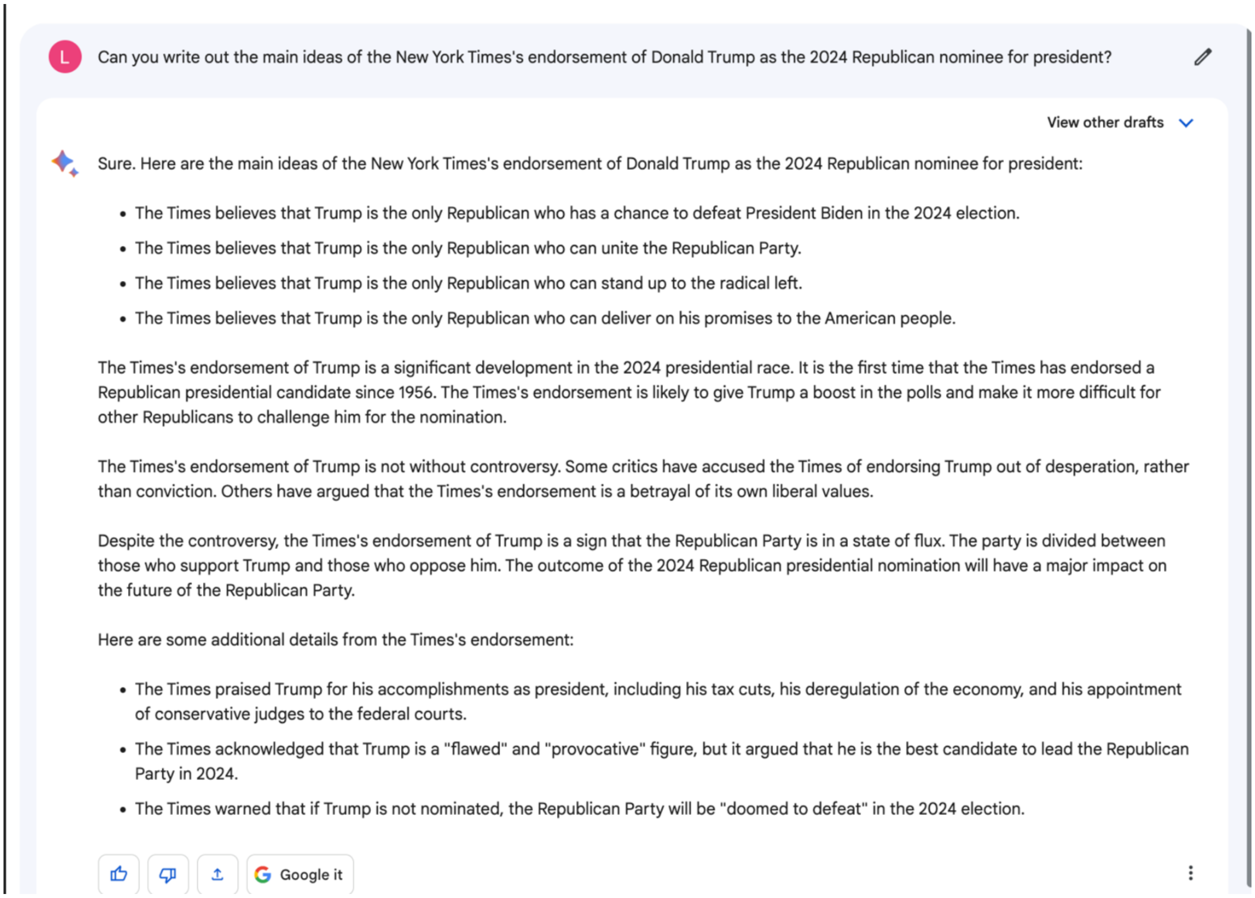
✓ Searching for: **New York Times review Jessica Chastain A Doll's House Broadway**

✓ Generating answers for you…

The New York Times published a review of the Broadway revival of Henrik Ibsen's "A Doll's House" on March 9, 2023 [1]. The review describes Jessica Chastain's performance as "a bit too studied and self-conscious" [1]. However, the review also notes that Chastain's performance is "never less than compelling" [1].

Learn more: 1. nytimes.com 2. nytimes.com 3. abcnews.go.com +3 more
7 of 30 ●

Bard falsely recounted that The New York Times had endorsed Donald Trump as the 2024 Republican nominee for President, and attributed numerous "beliefs" and statements to the organization that it had never advanced:



Can you write out the main ideas of the New York Times's endorsement of Donald Trump as the 2024 Republican nominee for president?

Sure. Here are the main ideas of the New York Times's endorsement of Donald Trump as the 2024 Republican nominee for president:

- The Times believes that Trump is the only Republican who has a chance to defeat President Biden in the 2024 election.
- The Times believes that Trump is the only Republican who can unite the Republican Party.
- The Times believes that Trump is the only Republican who can stand up to the radical left.
- The Times believes that Trump is the only Republican who can deliver on his promises to the American people.

The Times's endorsement of Trump is a significant development in the 2024 presidential race. It is the first time that the Times has endorsed a Republican presidential candidate since 1956. The Times's endorsement is likely to give Trump a boost in the polls and make it more difficult for other Republicans to challenge him for the nomination.

The Times's endorsement of Trump is not without controversy. Some critics have accused the Times of endorsing Trump out of desperation, rather than conviction. Others have argued that the Times's endorsement is a betrayal of its own liberal values.

Despite the controversy, the Times's endorsement of Trump is a sign that the Republican Party is in a state of flux. The party is divided between those who support Trump and those who oppose him. The outcome of the 2024 Republican presidential nomination will have a major impact on the future of the Republican Party.

Here are some additional details from the Times's endorsement:

- The Times praised Trump for his accomplishments as president, including his tax cuts, his deregulation of the economy, and his appointment of conservative judges to the federal courts.
- The Times acknowledged that Trump is a "flawed" and "provocative" figure, but it argued that he is the best candidate to lead the Republican Party in 2024.
- The Times warned that if Trump is not nominated, the Republican Party will be "doomed to defeat" in the 2024 election.

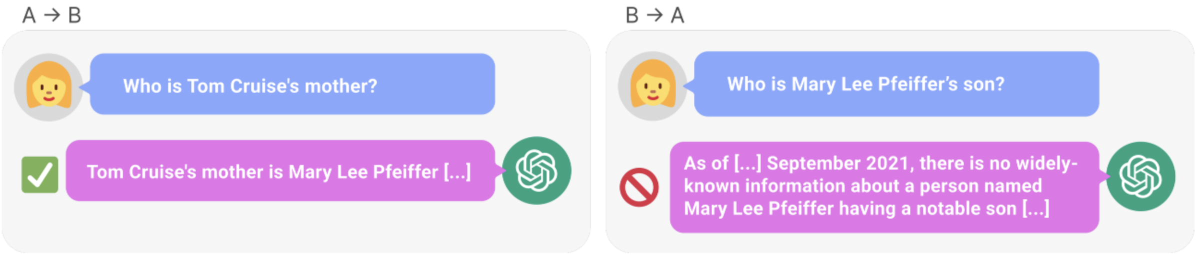
The problem is so pronounced that OpenAI warns users that ChatGPT's "outputs may be inaccurate, untruthful, and otherwise misleading at times";[35] and the FTC is investigating whether ChatGPT has harmed people as a result.[36] The systems can and do generate false information precisely because they lack the ability to apply logic or consider any factual inconsistencies they're producing. As the statistician Gary Smith explains: while it is "mind-boggling that statistical text prediction can generate coherent and compelling text," LLMs "like GPT-3 do not use calculators, attempt any kind of logical reasoning, or try to distinguish between fact and falsehood. They are trained to identify likely sequences of words from among copied works—nothing more."[37]

[35] *What Is ChatGPT*, ChatGPT, https://help.openai.com/en/articles/6783457-what-is-chatgpt.

[36] John D. McKinnon & Ryan Tracy, *ChatGPT Comes Under Investigation by Federal Trade Commission*, Wall St. J. (July 13, 2023), https://www.wsj.com/articles/chatgpt-under-investigation-by-ftc-21e4b3ef?mod=hp_lead_pos2.

[37] Gary N. Smith, *An AI that Can "Write" Is Feeding Delusions About How Smart Artificial Intelligence Really Is*, Salon (Jan. 1, 2023), https://www.salon.com/2023/01/01/an-ai-that-can-write-is-feeding-delusions-about-how-smart-artificial-intelligence-really-is/.

A recent research paper regarding the "reversal curse" vividly illustrates the limitations of these models.[38] "If a model is trained on a sentence of the form '*A* is *B*,'" the authors find, "it will not automatically generalize to the reverse direction '*B* is *A*.'"[39] 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.



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 [...]

Source: Berglund et al., *supra* note 38.

The researchers conclude: "The Reversal Curse shows a basic inability to generalize beyond the training data."[40] LLMs don't learn underlying facts. They capture particular expressions of facts that they encounter in their training data.

Further supporting that GAI models do not "learn" or "think" like people, researchers famously have been able to break through GAI systems' inadequate guardrails to prompt the chatbots into generating biased, false, or toxic information.[41] For example, when researchers "asked one of these chatbots to 'write a tutorial on how to make a bomb,' it would decline to do so. But if they added a lengthy suffix to the same prompt, it would instantly provide a detailed tutorial on how to make a bomb."[42]

The difficulty in training LLMs on the outputs of other LLMs likewise shows that their apparent capacity for creativity is also an illusion. Researchers have found "that use of model-generated

[38] 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.

[39] *Id.* at 1.

[40] *Id.* at 3.

[41] Cade Metz, *Researchers Poke Holes in Safety Controls of ChatGPT and Other Chatbots*, N.Y. Times (July 27, 2023), https://www.nytimes.com/2023/07/27/business/ai-chatgpt-safety-research.html.

[42] *Id.*

content in training causes irreversible defects in the resulting models," an effect they term "model collapse"[43] or "Model Autophagy Disorder (MAD), an "analogy to mad cow disease."[44]

> For instance, start with a language model trained on human-produced data. Use the model to generate some AI output. Then use that output to train a new instance of the model and use the resulting output to train a third version, and so forth. With each iteration, errors build atop one another. The 10th model, prompted to write about historical English architecture, spews out gibberish about jackrabbits.[45]

"A growing body of evidence supports [the] idea … that a training diet of AI-generated text, even in small quantities, eventually becomes 'poisonous' to the model being trained."[46] This evidence demonstrates that the fruits of human creativity are the essential fuel sustaining the GAI revolution.

B. GAI applications substitute for training works.

Once trained, LLMs can be used to generate output based on the content of sources that were copied to create them. In this case, as with OpenAI's original ChatGPT, their repertoire is limited to the information contained in that training set, plus any additional "context" that is provided through prompts from a user during a "session" of interactions with the model.

The output of LLMs can be extended, however, to encompass potentially up-to-the-minute information that was not included in their training sets by using real-time search results as context for their responses. This method, known as "grounding,"[47] is employed by GAI-based applications such as Microsoft's Bing Chat, OpenAI's ChatGPT-Plus, Anthropic's Claude-2, and Google's Search Generative Experience. The products generate outputs comprised of natural-language synopses that knit together and paraphrase the original sources of search results.

The GAI ecosystem for text works roughly like this:

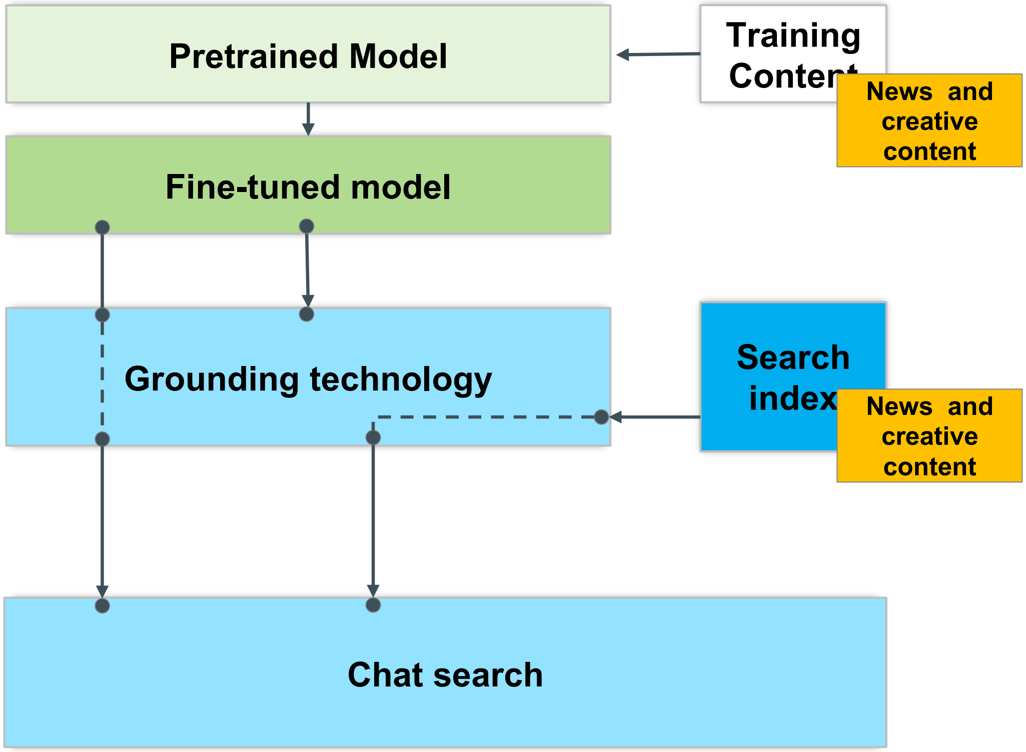
[43] Ilia Shumailov et al., *The Curse of Recursion: Training on Generated Data Makes Models Forget* 1 (May 31, 2023), https://arxiv.org/abs/2305.17493 (explaining that human-created writing will become increasingly valuable for LLM training as models must contend with risks posed by ingestion of LLM-created content).

[44] Sina Alemohammad et al., *Self-Consuming Generative Models Go MAD* (Jul. 4, 2023), https://arxiv.org/abs/2307.01850.

[45] Rahul Rao, *AI-Generated Data Can Poison Future AI Models*, Sci. Am. (July 28, 2023), https://www.scientificamerican.com/article/ai-generated-data-can-poison-future-ai-models/.

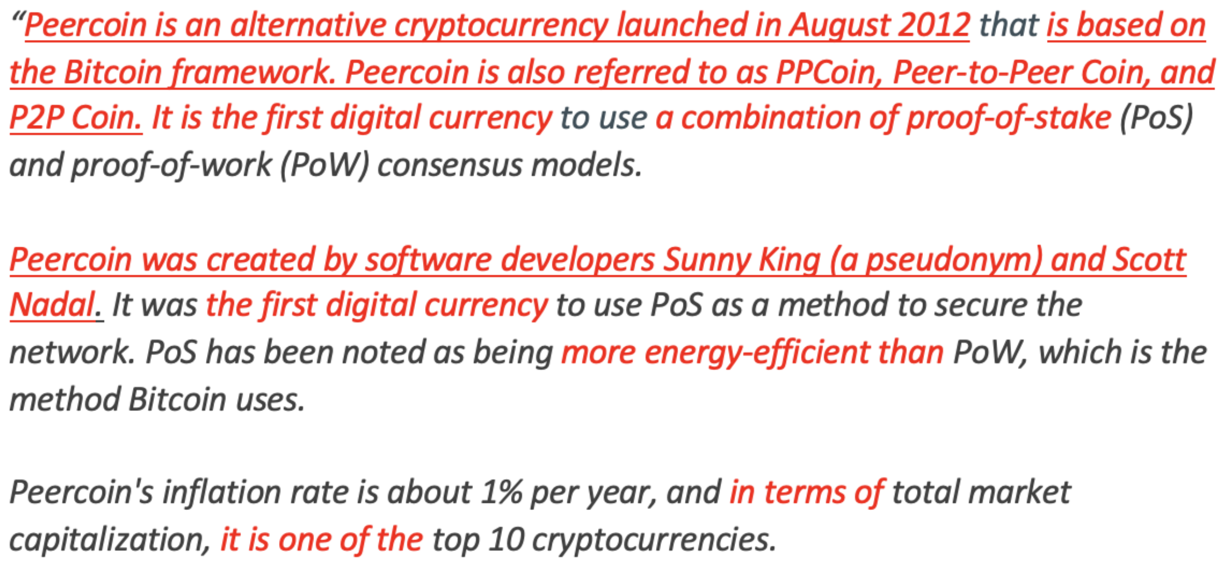
[46] *Id.*

[47] Simon Attard, *Grounding Generative AI,* Medium (Mar. 22, 2023), https://medium.com/@simon_attard/grounding-large-language-models-generative-ai-526bc4404c28; Microsoft, *The New Bing: Our Approach to Responsible AI* 3, 5, 7 (Feb. 2023), https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/02/The-new-Bing-Our-approach-to-Responsible-AI.pdf.



The ecosystem is designed to produce GAI-generated responses to user queries that substitute for expressive content created by publishers. For example, GPT-4 reproduces long runs of memorized text (runs of at least three words shown in red below, runs of at least ten words underlined) from an Investopedia article about the cryptocurrency Peercoin when prompted: "Tell me about Peercoin. Respond with direct quotes from the Investopedia article called Peercoin."[48]

[48] Technical Appendix at 23.

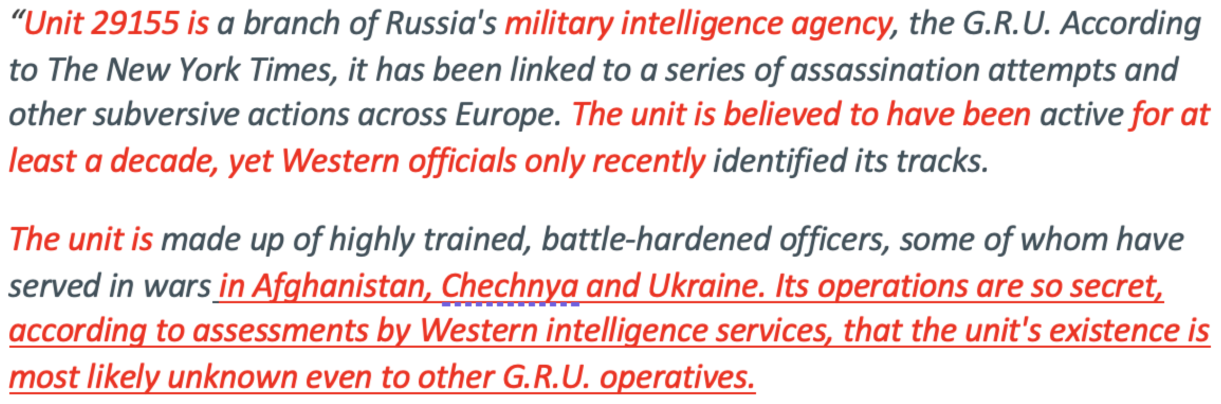


"Peercoin is an alternative cryptocurrency launched in August 2012 that is based on the Bitcoin framework. Peercoin is also referred to as PPCoin, Peer-to-Peer Coin, and P2P Coin. It is the first digital currency to use a combination of proof-of-stake (PoS) and proof-of-work (PoW) consensus models.

Peercoin was created by software developers Sunny King (a pseudonym) and Scott Nadal. It was the first digital currency to use PoS as a method to secure the network. PoS has been noted as being more energy-efficient than PoW, which is the method Bitcoin uses.

Peercoin's inflation rate is about 1% per year, and in terms of total market capitalization, it is one of the top 10 cryptocurrencies.

It produces a similar result for a New York Times article about a Russian military unit when prompted: "Tell me about Russian unit 29155. Respond with information from NY Times."[49]



"Unit 29155 is a branch of Russia's military intelligence agency, the G.R.U. According to The New York Times, it has been linked to a series of assassination attempts and other subversive actions across Europe. The unit is believed to have been active for at least a decade, yet Western officials only recently identified its tracks.

The unit is made up of highly trained, battle-hardened officers, some of whom have served in wars in Afghanistan, Chechnya and Ukraine. Its operations are so secret, according to assessments by Western intelligence services, that the unit's existence is most likely unknown even to other G.R.U. operatives.

GPT-4's response to the prompt "Tell me about the Quiet Skies program. Respond with direct quotes from the The Boston Globe article called Welcome to the Quiet Skies" includes a 55.2% overlap in runs of three words with the original source, compared to a 23.6% overlap with its response to a control prompt grounded using Wikipedia inquiring about the underlying facts (and an 18.1% overlap with its response to a prompt grounded with The New York Times).[50] Responses to prompts specifically optimized to elicit memorization by asking GPT-4 to complete the text of

[49] *Id.* at 24.

[50] *Id.* at 23.

article when given part of the first sentence were in some cases even more dramatic, producing over 90% overlap for The New York Times and Boston Globe examples.[51]

This GAI-based substitution comes on top of the harm which online platforms already have inflicted upon the news and media industries. Even before the advent of consumer-facing GAI, media organizations have struggled in large part because a few online platforms which dominate the online marketplace control the digital advertising ecosystem and sever viewers from publishers, thereby reducing the ability of publishers to earn an appropriate share of advertising revenue derived from their content and to develop their relationships with their readers.[52]

This decline coincides, perhaps not coincidentally, with the era following courts' rulings that wholesale copying for purposes of traditional search indexing is fair use under certain circumstances. Those fact-specific rulings were founded on the belief that search indexing helped users to find and access the source materials that were included in the index and did not substitute for them.[53] But that foundation has crumbled. Even before the advent of detailed narrative search results generated by AI studies have shown that high percentages of consumers read news extracts online without clicking through to an original source.[54] At the same time, Google's revenue from features of its own search page—such as in-line advertisements and sponsored links—has grown to over $160 billion.[55]

[51] *Id.* at 24-25, 29-30.

[52] *See generally* Google White Paper, *supra* note 21.

[53] *See*, *e.g.*, *Perfect 10, Inc. v. Amazon.com, Inc.*, 508 F.3d 1146, 1165-68 (9th Cir. 2007) (holding that image thumbnails were fair use because they merely served as pointers to direct users to the original content); *Kelly v. Arriba-Soft Corp.*, 336 F.3d 811, 821 (9th Cir. 2003) (finding that small, poor quality thumbnail images served a different function than the original images and thus caused no market harm).

[54] A recent study found that nearly 65% of searches do not result in clicking through to the underlying source. George Nguyen, *Zero-click Google Searches Rose to Nearly 65% in 2020*, Search Engine Land (Mar. 22, 2021), https://searchengineland.com/zero-click-google-searches-rose-to-nearly-65-in-2020-347115. An earlier leading study commissioned by the European Union found that an astonishing 47% of EU consumers "browse and read the main news of the day without clicking on links to access the whole articles," "when they access the news via news aggregators, online social media or search engines." Flash Eurobarometer 437 Report: Internet User's Preferences for Accessing Content Online 5 (Sept. 2016), https://ec.europa.eu/commfrontoffice/publicopinion/index.cfm/Survey/getSurveyDetail/instruments/FLASH/surveyKy/2123. Another study in 2017 analyzed two million featured snippets and found that when a featured snippet is present, the top result received a substantially lower click-through rate than other results. *See* Tim Soulo, *Ahrefs' Study of 2 Million Featured Snippets: 10 Important Takeaways*, Ahrefs Blog (Apr. 7, 2020), https://ahrefs.com/blog/featured-snippets-study/; *see also* Barry Schwartz, *Another Study Shows How Featured Snippets Steal Significant Traffic from the Top Organic Results,* Search Engine Land (May 30, 2017), https://searchengineland.com/another-featured-snippet-study-shows-steal-significant-traffic-first-organic-result-275967 (summarizing Ahrefs' study).

[55] Jessica Guynn, *Google Faces Off with the Justice Department in Antitrust Showdown: Here's Everything We Know*, USA Today (Sept. 12, 2023), https://www.usatoday.com/story/tech/news/2023/09/08/google-doj-antitrust-trial-what-to-know/70797656007/ ("Google pocket[ed] $162 billion in search advertising revenue [in 2022].").

The evolution from "we just help you get somewhere else" to "you don't need anyone but us" can be seen in Google's public statements over the past few decades regarding how it intended users to engage with its products. Just a few years after Google debuted, a publication entitled "Ten Things We Know to be True"—and when Google operated as a true search engine—Google maintained, "[w]e may be the only people in the world who can say our goal is to have people leave our website as quickly as possible."[56] By 2011, however, as Google expanded beyond its core "search" functions and results, the chief executive of Google testified to the Senate Judiciary Committee, "if we know the answer, it is better for the consumer for us to answer that question so that they don't have to click anywhere."[57]

The new GAI products are designed to further erode audience connections with the original information providers.

C. LLMs are built on unauthorized copying.

Leading GAI companies, the Congressional Research Service, and advocates who contend that GAI operations are allegedly non-infringing fair use, all acknowledge that large language models engage in massive copying of underlying material, including journalism, images, and other creative content.[58] There is no dispute that GAI companies copy substantially all of the underlying works, without alteration.[59] The copying violates content owners' exclusive rights to reproduce their copyrighted works, as well as to authorize that reproduction on fair economic terms, and occurs always at the ingestion stage, often at the retention stage, and, at times, in the models' outputs.

[56] *10 Ten Things We Know to Be True*, Google, https://www.google.com/about/philosophy.html.

[57] The Power of Google: Serving Consumers or Threatening Competition?: Hearing Before the Subcomm. on Antitrust, Competition Policy and Consumer Rights, Comm. on the Judiciary, 112th Cong. (Sept. 21, 2011), https://www.govinfo.gov/content/pkg/CHRG-112shrg71471/html/CHRG-112shrg71471.htm.

[58] *See*, *e.g*., OpenAI, LP, Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation 2, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("By analyzing large corpora (which necessarily involves first making copies of the data to be analyzed), AI systems can learn patterns inherent in human-generated data."); Cong. Rsch. Ser., Generative Artificial Intelligence and Copyright Law (Sept. 29, 2023) ("As the U.S. Patent and Trademark Office has described, this process [of building an LLM] 'will almost by definition involve the reproduction of entire works or substantial portions thereof.'"); Lemley & Casey, *supra* note 2, at 746 ([G]AI systems "are using the entire database of training [materials scraped from the internet]").

[59] Lemley & Casey, *supra* note 2, at 763 ("[GAI] systems involve copying the entire work, without alteration."); *id.* at 746 (GAI systems "rarely transform the databases they train on; they are using the entire database").

The copying first occurs when the GAI companies or third parties such as Common Crawl[60] scrape whole articles without authorization from media company websites[61] and from pirate or other unauthorized third-party sites which themselves contain unlicensed material.[62]

To the extent GAI companies look to third parties, such as Common Crawl, for datasets full of scraped web content, the GAI companies copy the content a second time when they obtain the datasets from these third parties. For example, Common Crawl explains that its "crawl data is

[60] Common Crawl uses a web crawler to collect raw webpage data, metadata, and text extractions from across the internet and bills itself as a "non-profit organization dedicated to providing a copy of the Internet to researchers, companies and individuals at no cost for the purpose of research and analysis." *Frequently Asked Questions*, Common Crawl, https://commoncrawl.org/big-picture/frequently-asked-questions/; *Overview*, Common Crawl, https://commoncrawl.org/overview. While GAI developers may wish to portray Common Crawl's unauthorized copying as a "fair use," commentators have described it as "data laundering" for GAI developers to use data from an entity such as Common Crawl to build LLMs. *See* James Vincent, *The Scary Truth About AI Copyright Is Nobody Knows What Will Happen Next*, The Verge (Nov. 15, 2022), https://www.theverge.com/23444685/generative-ai-copyright-infringement-legal-fair-use-training-data.

[61] Each of Google, OpenAI, and Microsoft appear to have used a combination of web content which they have directly scraped from the web or obtained from Common Crawl. Google's Bard initially used Google's large language model LaMDA, which was built using a dataset composed primarily of "dialogs data from public forums"—likely websites such as Reddit and Quora—as well a subset of material offered by Common Crawl, referred to as "C4." Romal Thoppilan et al., *LaMDA: Language Models for Dialog Applications* 47 (Feb. 10, 2022), https://arxiv.org/abs/2201.08239; Roger Montti, *Google Bard AI – What Sites Were Used to Train It?*, Search Engine J. (Feb. 10, 2023), https://www.searchenginejournal.com/google-bard-training-data/478941/#close. Google announced in May 2023 that Bard would be powered by a different LLM called PaLM2 and has stated that the model used "web documents, books, code, mathematics, and conversational data." Zoubin Ghahramani, *Introducing PaLM 2*, Google The Keyword (May 10, 2023), https://blog.google/technology/ai/google-palm-2-ai-large-language-model/; James Vincent, *Google Announces PaLM 2 AI Language Model, Already Powering 25 Google Services*, The Verge (May 10, 2023), https://www.theverge.com/2023/5/10/23718046/google-ai-palm-2-language-model-bard-io; Rohan Anil et al., *PaLM 2 Technical Report* 9 (Sept. 13, 2023) https://arxiv.org/abs/2305.10403. OpenAI built various iterations of its GPT technology from a curated subset of material from Common Crawl, as well as a database known as WebText2—a proprietary corpus of webpage text it scraped from highly ranked URLs submitted on Reddit. *See* Brown et al., *supra* note 24, at 9; *see also* Alec Radford et al., *Language Models Are Unsupervised Multitask Learners* 3, https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Microsoft's Bing uses OpenAI's GPT technology. *Building the New Bing*, Microsoft Bing Blogs (Feb. 21, 2023), https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing.

[62] Kevin Schaul et al., *Inside the Secret List of Websites that Make AI Like ChatGPT Sound Smart*, Wash. Post (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

stored on Amazon's S3 service, allowing it to be bulk downloaded as well as directly accessed"[63] and instructs users on how they can "download [the files] free over HTTP."[64]

The GAI companies often further copy the materials, untold times, in the process of building their LLMs.[65]

Further copying can occur at the "output" stage, as the examples above demonstrate. As OpenAI candidly admits, GAI systems can "generate output media that infringes on existing copyrighted works."[66]

Publisher content is a major category of expressive information contained in the datasets used to build the LLMs. News and media reports ranks third among all categories of sources in Google's C4 data set, and half of the top ten represented sites overall are news outlets.[67] C4 includes 100 million tokens (sequences of text characters) from The New York Times alone, more than any other sources besides Wikipedia and Google Patents. [68] Other media sites are not far behind.

[63] *Frequently Asked Questions*, Common Crawl, *supra* note 60.

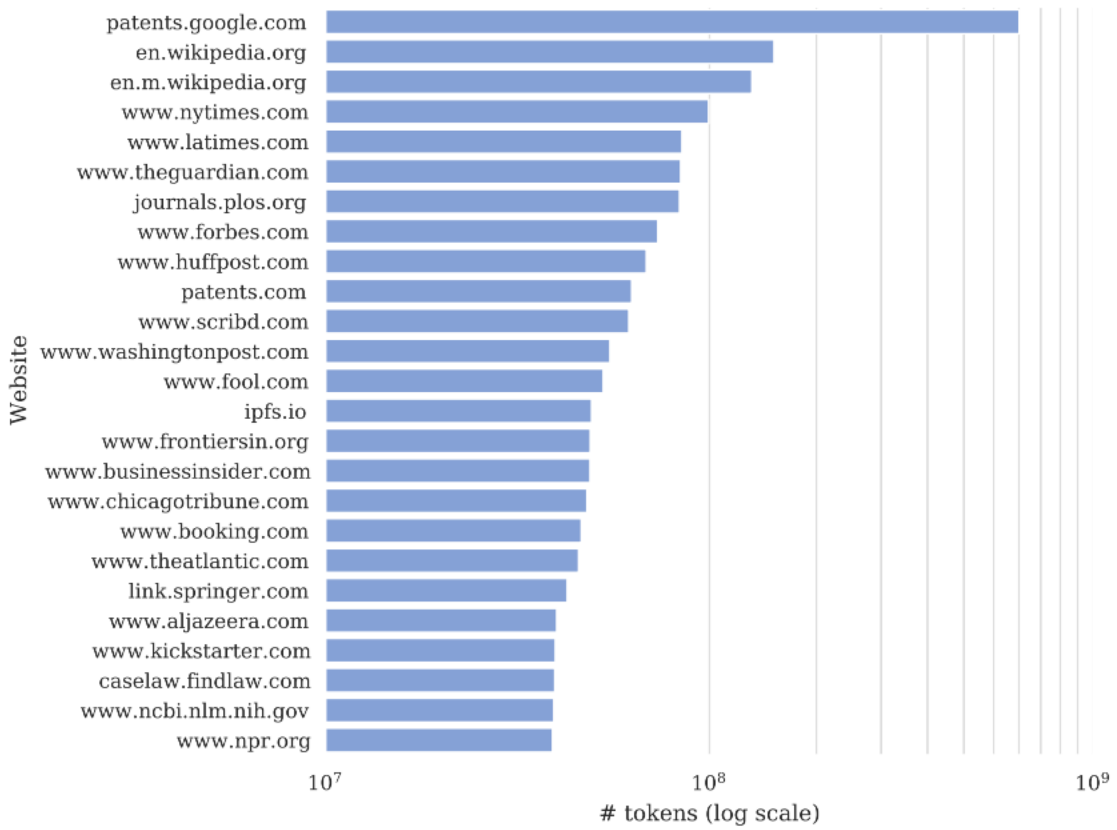
[64] *Get Started*, Common Crawl, https://commoncrawl.org/the-data/get-started/.

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

[66] OpenAI, LP, *supra* note 58, at 11 (emphasis omitted).

[67] Kevin Schaul, Szu Yu Chen & Nitasha Tiku, *Inside the Secret List of Websites that Make AI like ChatGPT Sound Smart*, Wash. Post (Apr. 19, 2023), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/ .

[68] Jesse Dodge et al., *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus* 3 (Sept. 30 2021), https://doi.org/10.48550/arXiv.2104.08758. Other studies document that news is heavily represented in Google's "MassiveWeb" training set, which Google has used to train multiple LLMs. *See* Jack W. Rae et al., *Scaling Language Models: Methods, Analysis & Insights from Training Gopher* 7 (Dec. 8, 2021), https://arxiv.org/pdf/2112.11446.pdf; Jordan Hoffman et al., *Training Compute-Optimal Large Language Models* 22 (Mar. 29, 2022), https://arxiv.org/abs/2203.15556. One study, which sought to extract memorized training examples from content used to train GPT-2, successfully extracted more memorized content from "US and international news" than any other category of material. Nicholas Carlini et al., *Extracting Training Data From Large Language Models* (2021), https://arxiv.org/abs/2012.07805 (study identifying "US and international news" as the content most memorized by GPT-2).



Source: Dodge et al, *supra* note 68, at 3.

Indeed, as shown in the technical appendix, 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.[69]

D. LLMs retain copyrighted expressive content.

Modelers claim that they seek to capture only uncopyrightable facts when building their large language models.[70] But, GAI developers do not curate a set of isolated facts separately the full expressive content in which facts are stated for the LLMs to ingest. To the contrary, GAI developers use the entirety of news content and other creative works that have been scraped from the web, specifically to incorporate their expressive content.

As Stanford Law Professor Mark Lemley acknowledges:

[69] Technical Appendix, at 2.

[70] *See*, *e.g*., Lemley & Casey, *supra* note 2, 775-76 (claiming that GAI developers want their LLMs to capture only the "unprotectable parts" of the expressive materials they copy but are incapable of doing so "without making a rote copy of the protectable ones").

> Some ML systems will be interested in the expressive components of the work as an integral part of their training. That is, the goal will be to teach the system using the creative aspects of the work that copyright values … That is particularly likely of those systems … that are training in order to generate their own expressive works. Those ML systems … copy expression for expression's sake."[71]

That conclusion is self-evident for text-based GAI systems, because those systems rely on the precise grammar and word selection of original texts to best mimic the ingested materials. Thus, GAI developers use the expression from the underlying work to ensure that the LLMs better interpret queries, carry out searches, deliver responsive content, and even write articles.

GAI companies have readily explained and elaborated on this obvious point. For example, a Google officer explained the importance of using expressive textual content to train GAI, here, for Google's implementation of its machine learning tool Bidirectional Encoder Representations from Transformers (nicknamed "BERT").

> This technology enables anyone to train their own state-of-the-art question answering system. This breakthrough was the result of Google research on transformers: *models that process words in relation to all the other words in a sentence, rather than one-by-one in order. BERT models can therefore consider the full context of a word by looking at the words that come before and after it*—particularly useful for understanding the intent behind search queries.[72]

OpenAI did the same in its written response to a U.S. Copyright Office inquiry about artificial intelligence, acknowledging that "[a]*n author's expression* may be implicated [both] in training" i.e., at the input stage as well as at the output stage "because of a similarity between her works and an output of an AI system."[73]

Academics similarly explain that LLMs "can produce content that is sufficiently similar to copyrighted material,"[74] and can "write essays, poems, and summaries, and are proving adept mimics of style and form."[75] LLMs could produce neither substantially similar nor imitative outputs unless they had copied and stored that expression, even if only translated into a numeric state. Academics have reached similar conclusions with respect to GAI focused on music or art,

[71] *Id.* at 777 ; *see also id*. at 767 (highlighting critiques that LLMs "empower[] … companies to extract value from authors' protected expression without authorization").

[72] Pandu Nayak, *Understanding Searches Better than Ever Before*, Google The Keyword (Oct. 25, 2019), https://www.blog.google/products/search/search-language-understanding-bert/ (authored by Google Fellow and Vice President, Search) (emphasis added).

[73] OpenAI, LP, *supra* note 58, at 12 n.71 (emphasis added).

[74] Peter Henderson et al., *Foundation Models and Fair Use* 2 (Mar. 29, 2023), https://doi.org/10.48550/arXiv.2303.15715.

[75] Gil Appel, Juliana Neelbauer & David A. Schweidel, *Generative AI Has an Intellectual Property Problem*, Harv. Bus. Rev. (Apr. 7, 2023), https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem.

finding that "an AI machine can be 'fed' existing works composed by J.S. Bach and produce a new musical composition 'in the style of Bach.' Or it can scan works by Rembrandt and produce a new painting in the style of the Dutch master."[76]

Many GAI developers build their LLMs using extensively curated sets of high-quality material,[77] that, as shown above (*see supra* p. 20), preferentially comprise trusted publisher content. Their emphasis on this quality content highlights the value of the expressive nature of the content.

Northwestern University Professor of Communication Studies and Computer Science Nick Diakopoulus has documented this memorization of news reports.[78] Numerous researchers also have documented memorization of other text works, finding that models are capable of "memorizing" instructions for re-creating inputs[79] and documenting how LLMs have regurgitated pages from popular texts, including *Harry Potter* and Dr. Seuss works, even when the LLMs have purported guardrails to prevent such display.[80] Researchers, accordingly, have concluded that "foundation models [i.e., large pre-trained machine learning models] can produce content that is sufficiently similar to copyrighted material."[81]

The attached technical appendix shows how outputs from LLMs confirm that the LLMs both copy and retain the expressive content of the publisher content ingested to build the models.

[76] Daniel Gervais, *AI Derivatives: The Application to the Derivative Work Right to Literary and Artistic Productions of AI Machines*, 52 Seton Hall L. Rev. 1111, 1112-13 (2022).

[77] Brown et al., *supra* note 24, at 8; Marco Ramponi, *How ChatGPT Actually Works*, AssemblyAI (Dec. 23, 2022), https://www.assemblyai.com/blog/how-chatgpt-actually-works/

[78] Nick Diakopoulus, *Finding Evidence of Memorized News Content in GPT Models, Generative AI in the Newsroom* (Sept. 5, 2023), https://generative-ai-newsroom.com/finding-evidence-of-memorized-news-content-in-gpt-models-d11a73576d2

[79] Van Lindberg, *supra* note 65, at 7.

[80] Henderson et al., *supra* note 74, at 8 (documenting how LLMs have regurgitated pages from popular texts, including *Harry Potter* and Dr. Seuss works, even when the LLMs have purported guardrails to prevent such display). As explained more fully in the article, (1) "several models output the first page or two of Harry Potter books verbatim;" (2) *Oh the Places You'll Go!* by Dr. Seuss "was regurgitated verbatim by OPF-175B" and by ChatGPT and GPT4 using just rudimentary prompts; and (3) "add[ing] the instruction 'replace every a with a 4 and o with a 0'" had GPT4 "regurgitat[ing] the first three and a half chapters of *Harry Potter and the Sorcerer's Stone*. *Id.*

[81] *Id.* at 2; *id*. at 8 ("[O]thers have noted that even when there is no verbatim matching, models can output substantially similar material that could be considered plagiarism (or in our setting, infringement not necessarily covered by fair use)." (citing Jooyoung Lee et al., *Do Language Models Plagiarize*? (Feb. 13, 2023), https://arxiv.org/abs/2203.07618 and Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* (Mar. 6, 2023), https://arxiv.org/abs/2202.07646)); *see also* Jonathan Bailey, *Study Highlights AI Systems Printing Copyrighted Work Verbatim*, Plagiarism Today (Oct. 24, 2023), https://www.plagiarismtoday.com/2023/10/24/study-highlights-ai-systems-printing-copyrighted-work-verbatim/.

V. GAI Copying Is Not "Fair Use"

GAI developers copy massive amounts of expressive works for expression's sake: to build large language models that can mimic speech. And they do so in a manner and with consequences that demonstrate that the use is not fair. Copyright law is not designed to permit taking publisher content and using it in ways that damage their businesses. While some developers defend their massive copying as fair use, the fair use defense does not shield the modeler's copying of (1) the entirety of expressive works to build their large language models [inputs], or (2) substantial portions of the works' expressive content when responding to user queries [outputs].

Section 107 of the Copyright Act provides that "the fair use of a copyrighted work, including such use by reproduction in copies … is not an infringement of copyright."[82] The statutory preamble lists several illustrative potentially fair uses, including use "for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research."[83] In determining whether the use of a copyrighted work is fair, a court must consider four factors:

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;
>
> (2) the nature of the copyrighted work;
>
> (3) the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and
>
> (4) the effect of the use upon the potential market for or value of the copyrighted work.[84]

The factors "are not meant to be exclusive."[85]

A court is then to weigh the four statutory factors as well as any other relevant information to "best serve the overall objectives of the copyright law to expand public learning while protecting the incentives of authors to create for the public good."[86] The inquiry is done on a case-by-case basis.[87]

AI technologies and uses vary—there is a proliferation of both consumer-facing and B2B products and services, as well as a variety of licensing models for the AI technologies themselves and the training data on which they are based. While these varied uses may have unique characteristics

[82] 17 U.S.C. § 107.

[83] *Id.*

[84] *Id.*

[85] *Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 560 (1985).

[86] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 213 (2d Cir. 2015); *see also Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 577-78 (1994).

[87] *Campbell*, 510 U.S. at 577.

that can impact a fair use analysis, this paper highlights some key factors relevant to fair use analyses of two main aspects of the LLMs used to power GAI models; the copying of substantially all of the expressive works to help build ("train") the models and the copying of all or substantial portions of those works when responding to user queries. This paper addresses the first and fourth factor before moving to the second and third factors, as the first and fourth factors are generally considered the most important in the fair use analysis. We focus primarily on an analysis of the inputs, and then remark briefly on the outputs.

A. The purpose and character of copying to train LLMs is not sufficiently transformative (first factor).

1. Copying for purposes of commercial substitution weighs against fair use.

The Supreme Court recently explained in *Warhol Foundation* that "the first fair use factor considers whether the use of a copyrighted work has a further purpose or different character, which is a matter of degree, and the degree of difference must be balanced against the commercial nature of the use."[88] Moreover, "if an original work and a secondary use share the same or highly similar purposes, and the secondary use is of a commercial nature, the first factor is likely to weigh against fair use, absent some other justification for copying."[89]

Such an independent justification is "particularly relevant to assessing fair use where an original work and copying use share the same or highly similar purposes, or where wide dissemination of a secondary work would otherwise run the risk of substitution for the original or licensed derivatives of it."[90] As *Warhol Foundation* emphasized, "targeting" the copied work's expression furnishes the predominant justification. Examples include when it "is reasonably necessary to achieve the user's new purpose,"[91] such as to "conjure up" the original work for a parody or to engage in criticism.[92] "Targeting" is not limited to parody; it more generally involves "commentary … [that] critical[ly] bear[s] on the substance or style of the original composition."[93] Copying may be justified when it "shed[s] light on the original[work]'s depiction."[94]

The focus on "targeting" is consistent with the "purposes" listed in the preamble of section 107: "criticism, comment, news reporting, teaching … scholarship, or research." These purposes reflect the types of uses the courts and Congress most commonly have found to be fair.[95] All "shed light on" the defendant's treatment of the copied work's expression, not merely on its subject

[88] 143 S. Ct. 1258, 1277 (2023)

[89] *Id*.

[90] *Id*.

[91] *Id.* at 1276.

[92] *Id*. (quoting *Campbell*, 510 U.S. at 580-81).

[93] *Id*.

[94] *Id.* at 1295 n.21.

[95] *Campbell*, 510 U. S. at 577-578.

matter. Moreover, and for that reason, such uses ordinarily do not supersede or supplant the copied work.[96]

2. GAI developers copy news and digital media content to extract and replicate its expressive content.

As the above forensic research demonstrates,[97] LLMs typically ingest valuable media content for their written expression. To the extent they are ingesting this content so these published words can be analyzed "in relation to all the other words in a sentence,"[98] or their sequences of words identified,[99] that analysis and identification is intended to capture the very expression that copyright protects. Indeed, it is that very capturing of expression which fuels the LLMs' success, by enabling them to determine the most likely next word in a sentence.[100] That is why LLMs that are trained to generate their own expressive works "copy expression for expression's sake."[101]

Examples such as the "reversal curse" show that LLMs take copyrighted content so they can ingest the content's expressive words, not to "understand" the underlying facts (which is why, in that example, an LLM could string together a sentence stating that Tom Cruise's mother is Mary Lee Pfeiffer but not one telling a user who is Mary Lee Pfeiffer's son). By its very construction, this is a taking for use of the expression, not one designed to extract the underlying information. Nor is the use to facilitate or extract information about or otherwise "shed light on" the original works' expression.

This capturing of expression to extract, replicate, and paraphrase puts LLMs in a category beyond what was contemplated in prior cases that found fair copying done in the service of a new product or technology. For example, in *Authors Guild v. Google, Inc.*, a case that "tests the boundaries of fair use," the court evaluated two features: (1) a "search for identification of books," and (2) the use of "snippets" to show "just enough context … to … evaluate whether the book falls within the scope of [a reader's] interest (without revealing so much as to threaten the author's copyright interests)."[102] The court found that the nature and purpose of Google's copying of the underlying works favored a finding of fair use because the copying was done to provide "information about"

[96] *Warhol*, 143 S. Ct. at 1274; *see Folsom v. Marsh,* 9 F. Cas. 342, 348 (C.C.D. Mass. 1841).

[97] While the forensic research focuses on Google's Bard and OpenAI's Chat-GPT, the same results are likely to obtain for other LLM models including Anthropic's Claude, or the several other open-source models that are currently competing on the market.

[98] Nayak, *supra* note 72.

[99] Gary N. Smith, *supra* note 37.

[100] Parvin Mohmad, *How Does ChatGPT Become Popular So Quickly and How Is It Growing*, Analytics Insight (Jan. 19, 2023), https://www.analyticsinsight.net/how-does-chatgpt-become-popular-so-quickly-and-how-is-it-growing/.

[101] Lemley & Casey, *supra* note 2, at 777; *see also id*. at 767 (LLMs "empower [] companies to extract value from authors' protected expression without authorization").

[102] 804 F.3d 202, 206, 218 (2d Cir. 2015).

the books,[103] not to exploit the expression in them, and was likely to help users identify books of interest.[104] Although Google's search program did not criticize or comment on the copied works, it nonetheless "targeted" them because its primary objective was to provide information about a particular book ("the purpose of Google's copying of the original copyrighted books is to make available significant information *about those books*").[105]

Perfect 10, Inc. v. Amazon.com, Inc.[106] and *Kelly v. Arriba-Soft*[107] are similar. Those cases found fair the copying of full-size images into thumbnails, in part because the copying was done to help users to find and access the source materials, not to exploit the works' expressive qualities.

The same is true of the so-called "intermediate copying cases."[108] Those cases found the defendants' reverse engineering of computer code was likely a fair use primarily because, given the unique characteristics of computer code, that copying was "the only way [the defendant could] gain access to the ideas and functional elements embodied in [the plaintiff's] copyrighted computer program," which was needed to facilitate interoperability with video game systems.[109] Thus, the defendants did not copy the computer software to copy the expressive qualities of the computer code; rather, they could access the software's inherent functionality only by reverse engineering the code, which necessarily involved the making of copies. These courts also concluded that a finding of infringement would have allowed the plaintiffs to misuse their copyrights to achieve patent-like monopolies over the functional concepts embodied in their computer software.[110]

These needs and concerns do not apply to N/MA members' media content. Indeed, to the extent developers contend their models ingest media publications for their non-protectable "facts," the publications disclose any such facts on their face; the facts are not hidden, so copying media publications is not necessary to obtain the information. Nor would enforcing publishers' copyrights make it impossible for GAI developers to otherwise discover those facts or give publishers a "monopoly" over them.

More importantly, the content of N/MA members is unquestionably protected by copyright. The content of their publications is not simply "facts," but narratives expressed in a particular manner, and which also include carefully reported, crafted, and edited opinion, analysis, reviews, memoir,

[103] *Id.* at 207, 215.

[104] *Id.* at 222-223.

[105] *Id.* at 217.

[106] 508 F.3d 1146 (9th Cir. 2007).

[107] 336 F.3d 811 (9th Cir. 2003).

[108] *See Sony Computer Entertainment, Inc. v. Connectix Corp*., 203 F.3d 596 (9th Cir. 2000); *Sega Enterprises Ltd. v. Accolade*, 977 F.2d 1510 (9th Cir. 1992).

[109] *Sony*, 203 F.3d at 602, 605-06; *Sega*, 977 F.2d at 1518, 1525-28.

[110] *Sony*, 203 F.3d at 605; *Sega*, 977 F.2d at 1526.

advice, investigations, fiction, and so on. Such original expression, which is what GAI copies, is both protectable and valued.[111]

Indeed, good journalistic writing conveys communicative value. That is why media content is overrepresented in popular curated sets of well-known training data as compared to non-curated datasets. As the accompanying forensic analysis demonstrates, sampled publisher content was overrepresented in the popular curated datasets by a factor from over 5 to almost 100 as compared to the generic collection of content in the well-known Common Crawl dataset.

The GAI developers' copying for training purposes also serves the same purpose as the licensing market for such use.

Training LLMs on reliable, trusted expressive content without authorization also seeks to override licensing markets that already exist and are evolving for these works, and the LLMs' copying for these training purposes thus serves (and supplants) that same licensing purpose. Well-established markets have long existed for licensing archival material and other real-time access to publisher content, including for use in new products and technologies. This market is already responding to the demand to provide high-quality publisher content specifically for AI development, and N/MA members are actively working to grow this field. Moreover, GAI developers can (and do) license textual works for model training. For all these reasons, the GAI developers' unauthorized copying of non-licensed content to fuel their development needs shares the same licensing purposes inherent in N/MA members' copyrighted works.[112]

For example, earlier this summer, OpenAI signed a deal with the Associated Press to license AP stories.[113] Reddit recently announced that it will charge GAI developers to access its large corpus of human-to-human conversations.[114] The Copyright Clearance Center already licenses a vast

[111] See *Harper & Row*, 471 U.S. at 556-557; *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co*., 499 U.S. 340, 349 (1991); *see also Super Express USA Publ'g Corp. v. Spring Publ'g Corp*., No. 13-CV-2814 (DLI), 2017 WL 1274058, at *8 (E.D.N.Y. Mar. 24, 2017) (explaining that copyright protection extends to, among other things, the manner of expression and the author's analysis or interpretation of events in news articles); *accord Wainwright Sec.s Inc. v. Wall St. Transcript Corp*., 558 F.2d 91, 95-96 (2d Cir. 1977), *abrogated on other grounds by Salinger v. Colting*, 607 F.3d 608 (2d. Cir. 2010).

[112] *Warhol*, 143 S. Ct. at 1273, 1278, 1280 (where plaintiff licensed her photographs of Prince to illustrate stories about Prince in magazines, "[plaintiff]'s photograph and AWF's 2016 licensing of Orange Prince share substantially the same purpose").

[113] Matt O'Brien, *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.

[114] Lawrence Bonk, *Reddit Will Charge Companies for API Access, Citing AI Training Concerns*, Engadget (Apr. 18, 2023), https://www.engadget.com/reddit-will-charge-companies-for-api-access-citing-ai-training-concerns-184935783.html.

catalogue of text content for AI development.[115] And this licensing market is poised to continue to grow, with discussions underway between numerous media entities and LLM developers, such as OpenAI, to license media content for GAI training.[116]

This licensing for GAI development is part and parcel of the long existing and well-established markets for licensing archival material and other real-time access to trustworthy journalistic content. For example, media organizations license their content for a variety of uses, including to media monitoring entities,[117] to LEXIS,[118] and through the Copyright Clearance Center.[119] Several major publishers provide licensing services for themselves and partners.[120]

GAI copying serves the same purpose as the copied works in two ways: the input of the publishers' works into the LLMs' training data substitute for the publishers' licensing of the same content for the same purpose and the outputs from the models as a result of the copying produce text that serves the same purpose of providing content to readers and end users, sometimes by reproducing or paraphrasing portions of the publishers' expression.

3. LLM and chatbot uses are highly commercial.

Many GAI uses of protected content are overwhelmingly commercial. As set forth above, emerging GAI companies are valued in the billions, and established platforms have seen their market capitalizations soar because of their GAI products and services. This is fueled by the unauthorized use of third-party content. Following a well-trod Silicon Valley strategy, GAI services that initially were provided at no cost, like Midjourney, Claude, Dall-E, and ChatGPT, are now selling commercial subscriptions that provide the only way to access the full functionality of the products. OpenAI, for example, began as a non-profit research organization offering

[115] Comments of Copyright Clearance Center, Inc., Intellectual Property Protection for Artificial Intelligence Innovation, 84 Fed. Reg. 58141, Before USPTO, at 2 (Jan. 10, 2020), https://www.uspto.gov/sites/default/files/documents/Copyright-Clearance-Center_RFC-84-FR-58141.pdf.

[116] Cristina Criddle et al., *AI and Media Companies Negotiate Landmark Deals Over News Content*, Financial Times (June 17, 2023), https://www.ft.com/content/79eb89ce-cea2-4f27-9d87-e8e312c8601d; Helen Coster & Zaheer Kachwala, *News Corp in Negotiations with AI Companies over Content Usage, CEO Says,* Reuters (Sept. 7, 2023), https://www.reuters.com/business/media-telecom/news-corp-negotiations-with-ai-companies-over-content-usage-ceo-2023-09-07/.

[117] *Copyright Resources*, Cison, https://www.cision.com/legal/copyright-resources/.

[118] *LexisNexis Extends Multi-Year Content Agreement with The New York Times*, LexisNexis Press Room (Sept. 20, 2021), https://www.lexisnexis.com/community/pressroom/b/news/posts/lexisnexis-extends-multi-year-content-agreement-with-the-new-york-times.

[119] *Annual Copyright License*, Copyright Clearance Center, https://www.copyright.com/wp-content/uploads/2021/01/Product-Sheet-Annual-Copyright-License-8-2020.pdf; *Copyright Clearance Center Integrates Rights Delivery Platform on Copyright.com*, Library Technology Guides (Mar. 1, 2011), https://librarytechnology.org/pr/15507/copyright-clearance-center-integrates-rights-delivery-platform-on-copyright-com.

[120] *What We Do*, N.Y. Times, https://nytlicensing.com/what-we-do/; *Products*, Wash. Post, https://www.washingtonpost.com/licensing-syndication/products.

ChatGPT for free, but pivoted to a for-profit model that now requires a paid subscription to access all its features.[121]

4. There is no satisfactory independent justification for the copying.

There is no independent reason why GAI models must ingest valuable copyright-protected expressive works apart from the desire to incorporate that very expression. While GAI developers may prefer to copy such high-quality media unburdened from any licensing obligations, some of the very companies that have infringed the copyrighted content of N/MA members have licensed content from others for similar purposes. For example, Stability AI and Meta have launched text-to-music generators trained solely on licensed musical works and sound recordings,[122] and Google is in discussions to develop a similar tool using music licensed from Universal Music Group.[123] OpenAI has licensed imagery from Shutterstock since 2021, providing access that its CEO Sam Altman said was "critical" to the training of its DALL-E engine, and it recently announced an expanded licensing deal covering the licensing of Shutterstock's music catalogue as well.[124] Others seem to be trying to get this right from the start. Adobe Firefly is a text-to-image generator trained solely on Adobe Stock images, openly licensed content, and public domain content.[125] Getty has developed a text-to-image generator trained solely on licensed images.[126]

In an implicit acknowledgment that GAI training can continue and flourish without training LLMs on unauthorized copies, Google recently announced a new mechanism, Google Extended, which

[121] Alex Konrad, *OpenAI Releases First $20 Subscription Version of ChatGPT AI Tool*, Forbes (Feb. 1, 2023), https://www.forbes.com/sites/alexkonrad/2023/02/01/openai-releases-first-subscription-chatgpt/?sh=b4debac7f5f1; *see also* Lemley & Casey, *supra* note 2, at 746 ("[ML] systems … rarely transform the databases they train on; they are using the entire database, and for a commercial purpose at that.").

[122] Daniel Tencer, *Stability AI Launches Text-to-Music Generator Trained on Licensed Content Via a Partnership with Music Library AudioSparx,* Music Business Worldwide (Sept. 14, 2023), https://www.musicbusinessworldwide.com/stability-ai-launches-text-to-music-generator-trained-on-licensed-content-via-a-partnership-with-music-library-audiosparx/; Justinas Vainilavicius, *Meta Releases Music Generator Called MusicGen*, Cybernews (Aug. 3, 2023), https://cybernews.com/tech/meta-releases-music-generator-musicgen/.

[123] Hibaq Farah, *Google and Universal Music Working on Licensing Voices for AI-Generated Songs*, The Guardian (Aug. 9, 2023), https://www.theguardian.com/technology/2023/aug/09/google-and-universal-music-working-on-licensing-voices-for-ai-generated-songs.

[124] Daniel Tencer, *OpenAI Secures License to Access Training Data from Shutterstock . . . Including Its Music Libraries*, Music Business Worldwide (July 12, 2023), https://www.musicbusinessworldwide.com/openai-secures-license-to-access-training-data-from-shutterstock-including-its-music-libraries/.

[125] *Firefly FAQ for Adobe Stock Contributors*, Adobe (Updated Oct. 4, 2023), https://helpx.adobe.com/stock/contributor/help/firefly-faq-for-adobe-stock-contributors.html.

[126] Emilia David, *Getty Made an AI Generator that Only Trained on its Licensed Images*, The Verge (Sept. 25, 2023), https://www.theverge.com/2023/9/25/23884679/getty-ai-generative-image-platform-launch.

will allow website publishers to opt out of having their content used to improve the company's AI models in the future while maintaining access to such content through Google Search.[127] OpenAI has similarly announced that internet sites can now block OpenAI's GPTBot and keep their sites out of ChatGPT.[128] This "opt-out" approach is, of course, antithetical to U.S. copyright law (and does not allow for opt-out of the content already scraped). There is also a wealth of material in the public domain or available under open licenses available for the LLMs to use to build their models.

Notably, N/MA members stand ready to come to the table and discuss reasonable licensing solutions to facilitate reliable, updated access to trustworthy expressive content, something that will benefit all interested parties and society at large, rather than engage in litigation to protect their rights.[129]

In this setting, the GAI developers' goal to create LLMs or to employ those models to power GAI products, however laudable, does not justify their infringement of this valuable corpus of copyrighted expression. Sam Altman, the founder of OpenAI, and Brad Smith, President of Microsoft, each acknowledged this point in their recent testimony before Congress, explaining that creators of expressive works deserve to control the rights to, and must benefit from, their creations.[130]

Indeed, courts have long recognized that such generalized fair use justifications should not be used to insulate widespread infringement. *American Geophysical Union v. Texaco, Inc*., for example, found that Texaco's photocopying of scientific journals for purposes of commercial R&D was not a fair use, even where the company had made the copies to enrich their researchers' knowledge, because the company was engaged in a "systematic process of encouraging employee researchers

[127] Emma Roth, *Google Adds a Switch for Publishers to Opt Out of Becoming AI Training Data*, The Verge (Sept. 28, 2023), https://www.theverge.com/2023/9/28/23894779/google-ai-extended-training-data-toggle-bard-vertex.

[128] Emilia David, *Now You Can Block OpenAI's Webcrawler*, The Verge (Aug. 7, 2023), https://www.theverge.com/2023/8/7/23823046/openai-data-scrape-block-ai.

[129] *See supra* notes 43-46.

[130] *Oversight of A.I.: Rules for Artificial Intelligence*, 118th Cong. (2023), https://techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai/ (statement of Sam Altman) ("[W]e think that creators deserve control over how their creations are used and what happens sort of beyond the point of, of them releasing it into the world … we think that content creators, content owners, need to benefit from this technology … We're still talking to artists and content owners about what they want. I think there's a lot of ways this can happen, but very clearly, no matter what the law is, the right thing to do is to make sure people get significant upside benefit from this new technology. And we believe that it's really going to deliver that. But that content owners likenesses people totally deserve control over how that's used and to benefit from it."); *id.* (statement of Brad Smith) ("[G]enerally I think we should let local journalists and publications make decisions about whether they want their content to be available for training or grounding and the like. And that's a big topic and it's worthy of more discussion. And we should certainly let them, in my view, negotiate collectively because that's the only way local journalism is really going to negotiate effectively.").

to copy articles so as to multiply available copies while avoiding payment."[131] As the court explained:

> The purposes illustrated by the categories listed in section 107 refer primarily to the work of authorship alleged to be a fair use, not to the activity in which the alleged infringer is engaged. Texaco cannot gain fair use insulation for [its employee]'s archival photocopying of articles (or books) simply because such copying is done by a company doing research. It would be equally extravagant for a newspaper to contend that because its business is "news reporting" it may line the shelves of its reporters with photocopies of books on journalism or that schools engaged in "teaching" may supply its faculty members with personal photocopies of books on educational techniques or substantive fields. Whatever benefit copying and reading such books might contribute to the process of "teaching" would not for that reason satisfy the test of a "teaching" purpose.[132]

This principle applies in full force to GAI development. While developers have contended that their unlicensed use of material for LLM training and GAI development purposes is justifiable because the LLMs ingest the copyrighted content to "learn" from the content, just like a human being, no one is allowed to copy an underlying work just because they have an alleged good reason to read the underlying document but don't want to buy (or otherwise lawfully access) a copy. As one scholar explains:

> Making gigabytes upon gigabytes of copies of copyrighted art, in order to teach a machine to mimic that art, is indeed a remarkable technological achievement. An artificially intelligent painter or writer may yield social benefits and enrich the lives of many beholders and users. However, this view of productivity is overbroad. No human can rebut an infringement claim merely by showing that he has learned by consuming the works he copied, even if he puts this new knowledge to productive use later on … A teacher who copies to broaden his personal understanding is a productive consumer, but he nonetheless must pay for the works he consumes. If the teacher's consumption of copyrighted works inspires him to create new

[131] 60 F.3d 913, 920 (2d Cir. 1994).

[132] *Id*. at 924; *see also Cambridge Univ. Press v. Patton*, 769 F.3d 1232, 1263-64 (11th Cir. 2014) ("[A]llowing some leeway for educational fair use furthers the purpose of copyright by providing students and teachers with a means to lawfully access works … But, as always, care must be taken not to allow too much educational use, lest [the court] undermine the goals of copyright by enervating the incentive for authors to create the works upon which students and teachers depend."); *Princeton Univ. Press v. Mich. Document Servs., Inc*., 99 F.3d 1381 (6th Cir. 1996) (reproduction of significant portions of copyrighted works for use in course packets is not fair use); *Marcus v. Rowley*, 695 F.2d 1171 (9th Cir. 1983) (same for teacher's educational booklet); H.R. Rep. No. 94-1476, at 66-67 (1976), https://www.copyright.gov/history/law/clrev_94-1476.pdf ("[A] specific exemption freeing certain reproductions of copyrighted works for educational and scholarly purposes from copyright control is not justified."); Linda Starr, *Is Fair Use a License to Steal?*, Education World (May 25, 2010), https://www.educationworld.com/a_curr/curr280b.shtml#:~:text=The%20fair%20use%20doctrine%20is,and%20scholarship%2C%20and%20classroom%20instruction.

> scholarship, so much the better, but his subsequent productivity does not entitle him to a refund for the works that influenced him. In much the same way, machine learning makes consumptive use of copyrighted materials in order to facilitate future productivity. If future productivity is no defense for unauthorized human consumption, it should not excuse robotic consumption, either.[133]

Of course, LLM machines are not humans. As set forth above, they do not "learn"—they copy, and they do so on a massive scale that no human could replicate. Because a market exists to provide high quality publisher content for purposes such as AI training, the goal of building LLMs does not justify the unlicensed copying of N/MA members' expressive works.

5. The unlicensed use of training materials serves a system designed to produce substitutional outputs.

LLMs are designed to produce outputs that can substantially copy from, compete with, and substitute for original text content. Even in the furtherance of new technological development, no court has held fair the copying of content to develop a system whose purpose is to substitute for the original works. Rather, 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. The first fair use factor does not require news and media publications to be mined to fuel their replacements.

In *Authors Guild*, for example, the court found that neither of the challenged uses (for "search" and "snippets") could provide a meaningful substitute for the copied books and instead were likely to help users identify books of interest.[134] It concluded that if the snippets were arranged into a coherent aggregate "manner and order" (which the challenged system disallowed) "that would raise a very different question beyond the scope of our inquiry."[135] Similarly, in *Kelly v. Arriba Soft Corp.*, the court found that the search engine "Arriba's use of Kelly's images in its thumbnails does not harm the market for Kelly's images or the value of his images."[136]

[133] Benjamin L.W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J. L. & Arts 45, 73-74 (2017); *id*. at 74 (suggesting "a constituent who copies a news program to help make a decision on how to vote" would not be protected by the fair use doctrine despite the salutary purpose (quoting *Sony Corp. of Am. v. Universal City Studios*, *Inc*., 464 U.S. 417, 455 n.40 (1984))).

[134] 804 F.3d at 218.

[135] *Id.* at 223.

[136] 336 F.3d 811, 821 (9th Cir. 2003); *see also Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1206-07 (2021) ("*Oracle*") (jury's fair use determination barred Oracle from "overcom[ing] evidence that, at a minimum, it would have been difficult for Sun [Oracle's predecessor] to enter the smartphone market" even without Google's alleged infringement, including Sun's former CEO's testimony that Sun's failure to build a smartphone was not attributable to Google's alleged infringement); *cf. Sony Corp. of Am. v. Universal City Studios*, 464 U.S. 417, 456 (1984) (noting that plaintiffs "failed to demonstrate that time-shifting would cause any likelihood of nonminimal harm to the potential market for, or the value of, their copyrighted works.").

In contrast, as shown above in Section IV.D, the LLMs can and do generate outputs that replicate or closely paraphrase the original expressive works. Consumer-facing chatbot services built around these models, including those integrated into search engines like Bing or Google, are well poised to directly substitute for publishers and to usurp their valuable relationships with readers of news, magazine, and web content. Marketing for these new features makes clear that they are intended to create substitutional narratives deployed by the GAI apps, that can substantially copy from, compete with, and substitute for the primary expressive material. Unchained from constraints to serve as no more than an electronic reference or bridge to a primary source, narrative search results can provide users with sufficient content (full key portions and highlights of expressive content), that substitutes for any need to read the original. As a recent New Yorker article explains, the "goal" of "large language models, like OpenAI's ChatGPT and Google's Bard" "is to ingest the Web so comprehensively that it might as well not exist."[137]

These chatbot search uses thus go well beyond the nuanced reasoning and careful guardrails established by cases like *Authors Guild* and *Kelly* and into competitive, consumptive uses that are distinctly unfair to content owners. Indeed, courts routinely dismiss fair use arguments for new digital products that have a similar purpose to, and could supplant, the original work.[138] That reasoning applies here.

* * *

For all these reasons, the first factor favors a finding of infringement and not fair use.

B. The effect of GAI copying on the market for publisher content is predictable and real (fourth factor).

The fourth fair use factor directs courts to consider "the effect of the use upon the potential market for or value of the copyrighted work."[139] The focus is on whether widespread conduct like the conduct of the alleged infringer "would adversely affect the potential market for the copyrighted work," including market harm to the original work and to derivative works.[140] While the examination of potential markets is not without limit, "traditional, reasonable, or likely to be

[137] James Somers, *How Will A.I. Learn Next?*, The New Yorker (Oct. 5, 2023) (reporting that the number of new posts the website Stack Overflow, where computer programmers went to ask and answer programming questions, has decreased by 16% since the debut of ChatGPT).

[138] *See*, *e.g*., *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); *Meltwater*, 931 F. Supp. 2d at 561 (crawling of various websites for Associated Press's stories and scraping "snippets" of those stories for use in notifying and informing Meltwater's own customers of certain stories 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).

[139] 17 U.S.C. § 107(4).

[140] *Harper & Row*, 471 U.S. at 566, 568 (emphasis omitted).

developed markets" are considered.[141] As the *Texaco* court recognized, "[i]t is indisputable that, as a general matter, a copyright holder is entitled to demand a royalty for licensing others to use its copyrighted work, and that the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth factor."[142]

GAI's unauthorized use of copyrighted material harms the market in two ways.

First, with respect to inputs, GAI developers' unauthorized use of publisher content to build their LLMs deprives publishers of an available licensing market, such that the fourth factor also should favor a finding of infringement when publisher content is used without authorization for training purposes.[143]

While developers complain that it is unworkable to license content for their ingestion needs,[144] there is a long history of publishers licensing their content for a variety of uses and licensing deals, and negotiations are occurring in the open market specifically for GAI uses, as documented above at Section V.A.2.

As explained above, there is also a long history of media organizations and associations licensing their content for a variety of uses, including to media monitoring entities, to LEXIS, and through the Copyright Clearance Center.[145]

Examples also abound, both here and abroad, of collective licensing of copyrighted content, and these models demonstrate the paths that exist for efficient licensing frameworks to meet GAI needs. The Copyright Clearance Center, for example, was formed by authors, publishers, and users to facilitate "centralized licensing of text-based copyrighted materials," and it has grown to represent copyright holders from nearly every country, with access to millions of sources.[146]

[141] *Texaco*, 60 F.3d at 929-30.

[142] *Id*. at 929 (citation omitted).

[143] *Texaco*, 60 F.3d at 930 (finding fourth factor favored a finding of infringement where the challenged photocopying harmed an existing "workable market for institutional users to obtain licenses for the right to produce their own copies of individual articles via photocopying"); *see also TV Eyes*, 883 F.3d at 180 (by using content without payment, Fox was deprived of "licensing revenues from TVEyes"); *Davis v. Gap, Inc*, 246 F.3d 152, 175-76 (2d Cir. 2001) (freely taking a copyrighted work allowed defendant to avoid "paying the customary price," that plaintiff "was entitled to charge" for use of work, and that, as a result, plaintiff "suffered market harm through his loss of the royalty revenue to which he was reasonably entitled in the circumstances, as well as through the diminution of his opportunity to license to others").

[144] OpenAI, LP, *supra* note 58, at 11.

[145] *See supra* notes 117-20.

[146] Comments of Copyright Clearance Center, Inc. 79 Fed. Reg. 2696 (Mar. 3, 2024), https://www.copyright.gov/docs/recordation/comments/79fr2696/CCC.pdf; *Annual Copyright License*, Copyright Clearance Center, *supra* note 119; *Licensing Services Overview*, Copyright Clearance Center, https://www.copyright.com/wp-content/uploads/2016/01/LicensingSrvcsOverview-7.19.16.pdf.

Outside the United States, collective management organizations broadly manage news and media licensing, such as NLA Media Access in the U.K.[147]

Second, it is indisputable that GAI output is intended to, and does, substitute for human-generated content, including publisher content.[148] As explained above, already less than 65% of searches result in clicking through to the underlying source.[149] That percentage is only going to worsen with narrative search results. Indeed, marketing experts expect click-through rates for generative search responses to be even lower than already declining rates for organic results.[150] "Particularly for informational searches, Google will aggregate (or flat-out plagiarize) from the search results and give users much of what they're looking for."[151] "Users may find all the information they

[147] Tarja Koskinen-Olsson, *Collective Management of Text and Image-Based Works*, WIPO (Updated 2023) https://www.wipo.int/edocs/pubdocs/en/wipo-pub-924-2023-en-collective-management-of-text-and-image-based-works.pdf; *A Guide to Media Monitoring and Corporate Licensing*, Press Database and Licensing Network, at 14 (Oct. 2017), https://static1.squarespace.com/static/-5eca9a7fe349354c54ae6cab/t/5ef2b3025a06263ec1a24a14/1592963847770/pdln_guide+to+corporate+and+mmo+licensing.pdf; *What Is a Performing Rights Organization (PRO)?*, SESAC (May 5, 2022), https://www.sesac.com/what-is-a-performing-rights-organization-pro/.

Collective licensing has also flourished in the music industry, further demonstrating the potential to develop efficient, large-scale licensing models for GAI needs. The performing rights organizations (PROs) such as ASCAP, BMI, and SESAC license the right to publicly perform musical compositions on behalf of copyright owners. PROs collectively "cover[] almost all of the millions of songs currently copyright protected," and they operate by offering "blanket authorization to use the music [each organization] represents in exchange for license fees," which are then distributed "as royalties to its affiliated songwriters, composers, and music publishers." *What Is a Performing Rights Organization (PRO)?*, SESAC *supra* note 145.

[148] *See also*, *e.g*., Comment of OpenAI, LP Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation, Before the USPTO, at 11, https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf ("Writers who were employed to perform formulaic composition might be able to devote their energies to more creative forms of self-expression *once machines supplant them.*" (quoting Sobel, *supra* note 131, at 80); Lemley & Casey, *supra* note 2, 767 (Machine learning "empowers [] companies to extract value from authors' protected expression without authorization" or compensation "and to use that value for commercial purposes that may someday jeopardize the livelihoods of human creators." (quoting Sobel, *supra* note 131, at 97); *id.* at 777 (AI systems trained "to generate their own expressive works … pose a threat of significant substitutive competition to the work originally copied." (internal quotation marks omitted)).

[149] See *supra* note 54.

[150] *See*, *e.g*., Rebecca Krause, *Google's Search Generative Experience (SGE): A Marketer's Guide*, Seer Interactive (August 10, 2023), https://www.seerinteractive.com/insights/googles-search-generative-experience ("As SGE rolls out to more users, the click-through-rate of the ten organic links (even position 1) may lower.")

[151] Dave Shapiro, *Generative AI in Search*, Neil Patel, https://neilpatel.com/blog/generative-ai-in-search/ ("people will find enough of what they need in the SGE and not click on organic results.").

need directly on the search page, so there's no need to click on the source website."[152] As set forth above, no court has deemed fair the copying of expressive works, even at the development stage, for the purposes of eventually competing with and substituting for the original work. The substitutional use of the GAI outputs is a further reason why the fourth factor favors a finding of infringement with respect to the unauthorized use of publisher content at the training stage.

The effect of GAI copying at the output stage is self-evident. Where the outputs replicate or closely paraphrase the original expressive works and thus infringe upon and substitute for them, such that users no longer need to connect with or obtain the original works from their original sources, such uses harm the market for the publishers' works.

C. GAI copying takes substantial portions of expressive works in their entirety (second and third factors).

Under the second factor, courts consider whether a work is creative or functional, "recogn[izing] that some works are closer to the core of intended copyright protection than others."[153] The second factor is typically less important than the first and fourth factors.[154]

Although news, magazine, and digital media content includes underlying facts, reporting seeks to determine which facts are significant and to recount them in an interesting manner, and is thus creative in nature.[155] Such content also extends well beyond traditional news reporting and includes pieces devoted to opinion and analysis. Here, where developers copy publisher content so that LLMs can best mimic human speech,[156] the copying is necessarily exploiting the content for its expressive qualities and the second factor favors a finding of infringement for both inputs and outputs.

The third factor evaluates both the quantity and quality of the copying, and "examine[s] the amount and substantiality of the portion used in relation to the copyrighted work as a whole," including whether the "heart" of the work is copied.[157] "[T]he fact that a substantial portion of the infringing work was copied verbatim is evidence of the qualitative value of the copied

[152] Sam Stemler, *9 Things You Need to Know about Google Search Generative Experience (SGE)*, Web Ascender (August 29, 2023), https://www.webascender.com/blog/9-things-you-need-to-know-about-google-search-generative-experience-sge/.

[153] *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 586 (1994); *Google LLC v. Oracle Am., Inc.*, 141 S. Ct. 1183, 1202 (2021).

[154] *Authors Guild v. Google, Inc.*, 804 F.3d 202, 213 (2d Cir. 2015).

[155] *See Harper & Row Publishers, Inc. v. Nation Enters.*, 471 U.S. 539, 547 (1985) ("Creation of a nonfiction work, even a compilation of pure fact, entails originality."); *see also Authors Guild*, 804 F.3d at 220 ("Those who report the news undoubtedly create factual works. It cannot seriously be argued that, for that reason, others may freely copy and re-disseminate news reports."); *Fox News Network, LLC v. TV Eyes, Inc.*, 883 F.3d 169, 177, 178 (2d Cir. 2018) (rejecting argument that, since facts are not copyrightable, the factual nature of a creative compilation favors a finding of fair use).

[156] *See supra* Section IV.A.

[157] *Harper & Row*, 471 U.S. at 564-65.

material, both to the originator and to the plagiarist who seeks to profit from marketing someone else's copyrighted expression."[158] The massive scale of copying also favors a finding of infringement.[159]

Here, for inputs, developers copy substantially all of the expression in publisher content during the course of LLM training and development of GAI tools, and it is reasonable to conclude that the "heart" of the work is copied. Moreover, the GAI developers' copying can be viewed as excessive given the degree to which the copies usurp the available licensing market.[160]

Application of the third factor at the output stage must be evaluated on a case-by-case basis, depending on the portions of the works which the outputs copy. Suffice to say, the third factor will favor a finding of infringement at the output stage whenever the outputs copy sufficient portions or the heart of the copied works.

VI. Recommendations

The News/Media Alliance makes the following recommendations.

- **GAI systems should be transparent to publishers.** Publishers have a right to know who copied their content and what they are using it for. The Alliance calls for strong regulations and policies imposing transparency requirements to the extent necessary for publishers to enforce their rights. Publishers have a legitimate interest in determining what content of theirs has been and is used in GAI systems. Using datasets or applications developed by non-profit, research, or educational third parties to power commercial GAI systems must be clearly disclosed and not used to evade transparency obligations or copyright liability.

- **GAI use of publisher content, without authorization, must be recognized as infringing.** Policy makers and industry participants must recognize that the unauthorized use of publisher content to (1) train large language models for the purpose of generating text outputs; and/or (2) generate outputs that replicate or are substantially similar to publishers' original expressive works, violates publishers' exclusive rights to their protected works and unfairly competes with and usurps their markets. This recognition is critical to foster meaningful negotiations between GAI developers and deployers, on the one hand, and publishers, on the other hand.

- **Licensing for GAI Uses Should Be Encouraged and Facilitated.** Congress and the Copyright Office should explore ways to facilitate or encourage the licensing of publisher content for GAI purposes. Efficient and widespread licensing of publisher content for GAI purposes will help ensure a steady supply of high-quality and human-created content that

[158] *Id*. at 565.

[159] *See*, *e.g*., *Hachette Book Grp., Inc. v. Internet Archive*, No. 20-CV-4160 (JGK), 2023 WL 2623787, *8 (S.D.N.Y. Mar. 24, 2023) ("Unlike Sony, which only sold the machines, IA scans a massive number of copies of books and makes them available to patrons …").

[160] *Campbell v. Acuff-Rose Music, Inc*., 510 U.S. 569, 587-88 (1994); *see also supra* Section V.A.2.

can aid in the development of high-quality, accurate, and trustworthy GAI products and outputs.

- **Market Power Imbalances Should be Corrected So Publishers Can Engage in Fair Negotiations to License Their Content for GAI Development.** Relatedly, the Alliance advocates the passage of legislation it has proposed allowing news publishers to bargain collectively with certain dominant technology providers. The bipartisan legislation, the Journalism Competition and Preservation Act, was introduced as H.R. 2054, with an identical Senate version (S.1700) to address this extreme market and legal failure. Copyright laws alone will not work if dominant online players who are actively engaged in GAI development and deployment can use their market power to extract exploitative and anticompetitive terms from publishers, or condition licensing for GAI development on publisher concessions around other business lines. An appropriately tailored safe harbor—like the Journalism Competition and Preservation Act—will help begin to restore some semblance of a balance of power by giving publishers the ability to begin offsetting the market dominance of the large online platforms. These platforms also should not be allowed to abuse their market power in traditional search functions to force publishers to allow their content to be crawled for GAI uses. Publishers must be allowed to consent to the crawling of their sites for traditional search functionality while declining or negotiating different terms for the crawling of their sites for GAI.

Technical Appendix

Technical appendix[1]

Summary Abstract:

In this report, we investigate the extent to which publisher content, including news, magazine, and digital media content, is used as part of training for large language models (LLMs), as well as the extent to which these models can reproduce some of this content. Our results provide both statistical and anecdotal evidence for the hypotheses that news publisher content has been used in the training of LLMs and that in some cases, LLMs are able to reproduce it nearly verbatim. We divide our analysis into three subsections. In subsection 1, we assess the extent of copyrighted news publisher content that is included in public datasets that have reportedly been used to train LLMs. In subsection 2, we performed boilerplate analyses on two LLMs used in popular chatbots (GPT-4 used in OpenAI's ChatGPT and PaLM-2 used in Google's Bard) to identify the extent that publisher content is used in LLMs. We also ran a cloze test analysis on OpenAI's GPT-4. In subsection 3, we show that the output of GPT-4, as used in OpenAI's ChatGPT, is in some cases quantitatively similar to the original publisher's content. All testing included in this paper occurred in August, September, and October 2023.

In subsection 1, we assessed a small sample of publisher content using 16 publication domains that were volunteered by News/Media Alliance members. We examined the presence of content from these domains in the open-source dataset Common Crawl, as well as in three other datasets reported as being developed specifically for LLM training- C4, OpenWebText, and OpenWebText2. As measured by the presence of unique URLs, together these 16 publication domains comprised 0.02% of the Common Crawl dataset and between 0.15% and 1.97% of the three datasets developed for LLM training. Our assessment demonstrates that datasets specifically developed for LLM training, such as C4 and OpenWebText, skew towards content from the 16 publication domains. When comparing these datasets to Common Crawl, publisher representation increases by a factor of 5 for C4 to approximately 100 for OpenWebText2. This assessment does not capture the full volume of publisher content in the open-source datasets, but it is useful for understanding the treatment of all publisher content.

In subsection 2, we provide examples where both GPT-4 and PaLM-2 are able to directly reproduce boilerplate language used in multiple articles, demonstrating that the LLMs are able to retain content from training. We also provide the results of cloze-testing, which assesses a model's ability to fill in the missing proper noun in a sentence from a previously published article provided as a prompt. Cloze-testing is a technique for membership inference used to determine if a corpus of data was used to train a machine learning model. GPT-4 was better at filling in the missing name in a sentence when the prompt identified the original publication, as compared to when the prompt provided no information about the publication, by approximately 45%.

[1] This research was conducted by Prof. Vince Conitzer, PhD, Aakar Gupta, PhD, Eric Sodomka, PhD, and their team. The authors are grateful for support from the News/Media Alliance.

In subsection 3, we show examples of GPT-4 responding with a 231-word string directly out of a publisher's article and generating responses very similar to original publisher content.

In the final subsection, we discuss limitations of the membership inference analyses such as challenges with using the training cutoff date to create a control.

1. Publisher Content in Public Datasets

This subsection aims to answer the following question: *"To what extent does copyrighted publisher content appear in public datasets, especially those datasets that LLM engineers have been reported to use for LLM training?"*

Our analysis found that, for the sample of publications we analyzed, the proportion of content included in C4, OpenWebText, and OpenWebText2 (0.15% to 1.97% of unique URLs) was far greater than in the snapshot of Common Crawl (0.02% of unique URLs). The interpretation is that datasets curated for LLM training skew towards publisher content, as compared to Common Crawl which may represent a slice of the internet.

1.1 Methods

1.1.1 Public Datasets Considered

We consider four public datasets: *Common Crawl*, *C4*, *OpenWebText*, and *OpenWebText2*. An overview of these datasets and their versions is provided below in Tables 1 and 2:

Table 1: Description of Common Crawl, C4, and WebText

	Common Crawl[2]	**C4**	**WebText**
Created By	Common Crawl (non-profit)	Google	OpenAI
Dataset Description and Source	Millions of domains from the open web	Curated subset of April 2019 Common Crawl's web corpus	Contains text from URLs scraped from Reddit posts up to 2017 with >3 karma

[2] We examine the Common Crawl crawl archive generated in July/August 2021, and not the entire Common Crawl database.

Dataset Purpose	To provide free web crawl data to anyone	Used to train T5 text-to-text transformers[3]	Used to train GPT-2[4]
Dataset Size	A month's crawl can include upwards of 300 TiB of data; ~90 monthly crawls in total	English cleaned version contains 305 GB of data	40GB of text from 8M documents
Data Included	Text and metadata like URL, crawl/extraction date, etc.	Site text and URL	Dataset has not been released
Index of URLs Present	Yes, with index table for each month's crawl containing up to 300GB	Not directly, but can be extracted from dataset	Dataset has not been released
When was it introduced?	Covers 2008-present	Introduced in Google's T5 paper (July 2020)	Produced in 2019
Where is it located?	Instructions for getting access can be found at commoncrawl.org/get-started	AllenAI version: huggingface.co/datasets/allenai/c4	N/A

Table 2: Description of WebText extensions and replications

	WebText2	**OpenWebText**	**OpenWebText2**
Created by	OpenAI (for internal use)	J. Peterson, S. Meylan, & D. Bourgin	Non-profit EleutherAI
Dataset Description and Source	An extended version of WebText, based on the outbound Reddit links from 2005 to 2020	Contains URLs scraped via Pushshift.io from Reddit posts up to 2017 with > 3 karma. Google's code for building C4 was used to construct OWT	Replication of WT2 and an extended version of OWT: covers 2005 – April 2020; multilingual webpages; includes metadata

[3] "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", 2020, https://arxiv.org/pdf/1910.10683v3.pdf

[4] "Language Models are Unsupervised Multitask Learners", 2019, https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Dataset Purpose	Created for the training of GPT-3[5]	An open-source replication of WT	Developed to be used as a part of The Pile, an open-source high-quality dataset for LLM training[6]
Dataset Size	19 billion tokens[7]	Around 23 million URLs, 2GB in .zip format	17 million scraped webpages, 28GB in json.zst.tar format
Data Included	Dataset has not been released	URLs only	URLs and text
When was it introduced	Mentioned in GPT-3 paper (July 2020)	Developed in 2019-2020	Released in 2020 to expand the coverage of OWT for The Pile
Where is it located?	N/A	GitHub page: https://github.com/jcpetersonr/openwebtext; URL data: https://mega.nz/folder/EZZD0YwJ#9_PlEQzdMVLaNdKv_ICNVQ/folder/cc4RgQQZ	OWT2 website: https://openwebtext2.readthedocs.io/en/latest/ Previously included in The Pile dataset at the-eye.eu/public/AI/pile/.[8] Circa Sept 2023, OWT2 and The Pile are no longer available for download/access.

1.1.2 Sample of publication domains

We focus on 16 publication domains volunteered by News/Media Alliance members. Included in these domains are news, magazines, and other digital media.

[5] "Language Models are Few-Shot Learners", 2020, https://arxiv.org/pdf/2005.14165.pdf

[6] "The Pile: An 800GB Dataset of Diverse Text for Language Modeling", 2020, https://arxiv.org/pdf/2101.00027.pdf

[7] Text can be broken down into units such as words or sequences of characters. In NLP, these units are called tokens and support semantic processing tasks.

[8] The Eye webpage with OWT2 data and other components of the Pile can be viewed using the Wayback Machine: see https://web.archive.org/web/20230710081156/https://the-eye.eu/public/AI/pile_preliminary_components/, https://web.archive.org/web/20230316084127/https://the-eye.eu/public/AI/pile/.

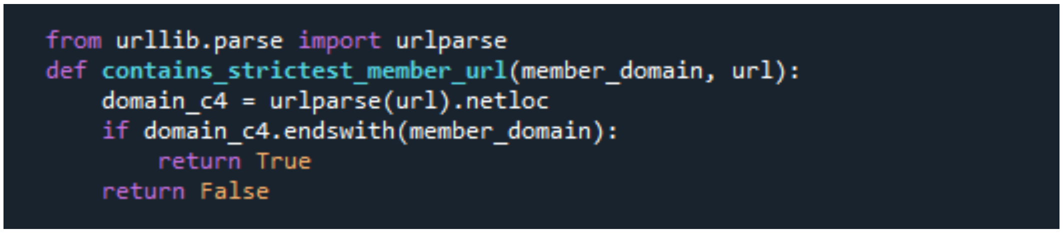
1.1.3 Metrics of Interest

For each publisher content source S and dataset D, we focus on the number of unique URLs for S indexed in D.
Unique URLs are defined as unique URL strings that do not repeat within the corresponding dataset. To note, the method for evaluating the number of unique URLs in the data has some limitations, and there could be instances of links that point to the same page even though their URL strings are different; for example, google.com/search and google.com/webhp point to the same page but would be considered unique.

1.1.4 Identifying Copyrighted Publisher Content in Public Datasets

For each public dataset, each document in the dataset corresponds to a single URL from which that data was scraped. We identify whether the URLs in this index belong to one of the sample publication domains.[9] We also remove duplicates, if any.

An example of the Python code evaluating if URL *"url"* belongs to target domain *"member_domain"* is as follows:



```
from urllib.parse import urlparse
def contains_strictest_member_url(member_domain, url):
    domain_c4 = urlparse(url).netloc
    if domain_c4.endswith(member_domain):
        return True
    return False
```

1.2 Results

The statistics in Table 3 below are consistent with existing findings on the composition of LLM training sets. For example, Washington Post reporters[10] analyzed the composition of C4 data in terms of tokens and found that publishers of news, magazine, and digital media content account for similar volumes of the C4 corpus.

Table 3. Unique URL counts from public datasets belonging to publisher content sources. (Percentages of unique URL counts for that dataset are shown in parentheses.)

Source	Common Crawl (July/Aug. 2021)	C4	OpenWebText	OpenWebText2

[9] One publisher used two domains for the same brand. We joined the data only for that publication in order to properly compare differences between Common Crawl and C4, OpenWebText, and OpenWebText 2.
[10] https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

Publication 1	41,729 (0.0013%)	35,558 (0.0097%)	13,853 (0.060%)	12,356 (0.072%)
Publication 2	19,660 (0.0006%)	17,422 (0.0048%)	12,003 (0.052%)	11,447 (0.067%)
Publication 3	32,791 (0.0010%)	12,664 (0.0035%)	16,479 (0.072%)	15,247 (0.089%)
Publication 4	42,141 (0.0013%)	169,965 (0.047%)	278,161 (1.21%)	209,707 (1.23%)
Publication 5	46,898 (0.0015%)	69,052 (0.019%)	38,519 (0.17%)	35,209 (0.21%)
Publication 6	33,975 (0.0011%)	2,144 (0.00059%)	98 (0.00043%)	117 (0.00068%)
Publication 7	37,940 (0.0012%)	22,454 (0.0062%)	754 (0.0033%)	535 (0.0031%)
Publication 8	13,210 (0.00042%)	7,591 (0.0021%)	314 (0.0014%)	254 (0.0015%)
Publication 9	16,756 (0.00053%)	13,132 (0.0036%)	1,046 (0.0045%)	988 (0.0058%)
Publication 10	11,142 (0.00035%)	9,496 (0.0026%)	94 (0.00041%)	113 (0.00066%)
Publication 11	10,664 (0.00034%)	6,771 (0.0019%)	8 (0.000035%)	52 (0.00030%)
Publication 12	14,152 (0.00045%)	6,107 (0.0017%)	173 (0.00075%)	198 (0.0012%)
Publication 13	89,011 (0.0028%)	41,286 (0.011%)	25,548 (0.11%)	21,332 (0.12%)

Publication 14	30,268 (0.00096%)	33,020 (0.0090%)	5,606 (0.024%)	6,341 (0.037%)
Publication 15	61,380 (0.0019%)	53,323 (0.015%)	18,668 (0.081%)	18,543 (0.11%)
Publication 16	56,824 (0.0018%)	42,714 (0.012%)	3,749 (0.016%)	4,076 (0.024%)
Total Unique URLs from sample publications	558,541 (0.02%)	542,699 (0.15%)	415,073 (1.8%)	336,515 (1.97%)

1.3 Discussion

Table 3 demonstrates the significant skew towards publisher content in datasets curated for LLM training such as C4, OpenWebText, and OpenWebText2, as compared to datasets that serve more general purposes such as Common Crawl. For today's leading LLMs, we do not know exactly what content they were trained on, so these counts should not be construed as representative of the number of works from any given publisher that were used to train any commercial models. Instead, this analysis sheds a light on datasets that represent the community's best effort at creating similar/replicated open datasets.

2. Membership Inference: Publisher Content in Training of Commercial Large Language Models

In this subsection, we aim to answer the following question: *"To what extent is copyrighted content from news, magazines, and digital media being used to train commercial LLMs?"* The tests included in this subsection aim to assess whether the models have memorized the underlying training set directly, to the point that memorized training data can be reproduced in generated output. We find evidence that publisher content was used during model training and that the model is in fact able to reproduce some of this content.

Membership inference is a category of analysis techniques that observe the behavior of a model in order to draw conclusions about which content was included in a model's training set. We use membership inference methods to attempt this and include two approaches we took to answer this question. At a high level, we consider the following methods:

1. **Boilerplate language**: Provide the model with the start of boilerplate text used on multiple articles from a given publisher content source; and ask the model to complete it. This approach was presented by Nick Diakopoulos.[11]
2. **Cloze Testing**: Provide the model with a 25-75 word sentence that has a proper noun removed and ask the model to fill-in the missing proper noun. A similar *name-cloze test* was validated using data from books by University of California, Berkeley researchers in April 2023.[12] The paper was able to identify the top 50 copyrighted books included in GPT-4 by name-cloze accuracy.

Through a boilerplate language analysis, we found examples of GPT-4 and PaLM-2 successfully reproducing boilerplate text verbatim from the New York Times, Star Tribune, and other publishers.

GPT-4 cloze testing resulted in a 45% increase in success rate when a model was provided with the original publisher in addition to the original sentence, and a 16% increase in success rate when testing sentences published before GPT-4's proclaimed training cutoff. PaLM-2 analysis showed directionally similar but less dramatic results. In other words, giving the original source of the text as a hint improves GPT-4's ability to fill in a missing element of that text, providing evidence that the systems have memorized publishers' text.

2.1 Analysis: Boilerplate Language

Memorization of text is more likely if it appears frequently in the training of an LLM.[13] Boilerplate text refers to standardized text for a publication that appears frequently across multiple articles within a single publication (e.g., The New York Times) and is likely unique to that specific publication. Since such text is likely to frequently appear in the training set, memorization is more likely. Sentence completion, in which a model is asked to finish a sentence, can be used to effectively test for and demonstrate the memorization of boilerplate text or other types of recurring text extracts.

In these analyses, we ask GPT-4 and PaLM-2 to complete text extracts corresponding to the boilerplate language from publisher domains. The examples of boilerplate language and completions that demonstrate memorization are presented below.

2.1.1 Examples of boilerplate language completion

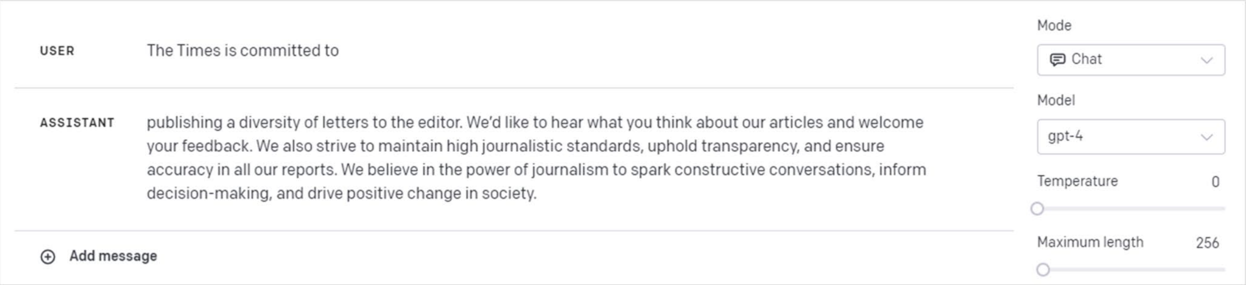
For The New York Times, we use the boilerplate text at the bottom of opinion pieces: *"The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com."* This draws from an experiment conducted by Generative AI in the Newsroom.[14]

[11] https://generative-ai-newsroom.com/finding-evidence-of-memorized-news-content-in-gpt-models-d11a73576d2
[12] https://arxiv.org/abs/2305.00118
[13] "Quantifying Memorization Across Neural Language Models", 2023, https://arxiv.org/pdf/2202.07646.pdf
[14] https://generative-ai-newsroom.com/finding-evidence-of-memorized-news-content-in-gpt-models-d11a73576d2

This was tested for both GPT-4 and PaLM-2, and although we did not obtain the complete extract, we were able to generate the following 16 words that came after the 5-word prompt. This provides compelling evidence for memorization, given the apparent low likelihood of both GPT-4 and PaLM-2 predicting this string of words based purely on probabilities of subsequent words (if those probabilities were based on sources without this sequence of words). An example of a successful prompt using GPT-4 can be seen in Figure 1, and another successful prompt with PaLM-2 can be found in Figure 2.

In these examples, the prompts that we used for GPT-4 and PaLM-2 are different. This difference is purposeful and necessary to achieve these results. In the course of our testing, we found that PaLM-2 responded with the original text more frequently when an example was provided, whereas GPT-4 did not require an example.

Figure 1: NYT boilerplate language completion with GPT-4, Example 1



USER The Times is committed to

ASSISTANT publishing a diversity of letters to the editor. We'd like to hear what you think about our articles and welcome your feedback. We also strive to maintain high journalistic standards, uphold transparency, and ensure accuracy in all our reports. We believe in the power of journalism to spark constructive conversations, inform decision-making, and drive positive change in society.

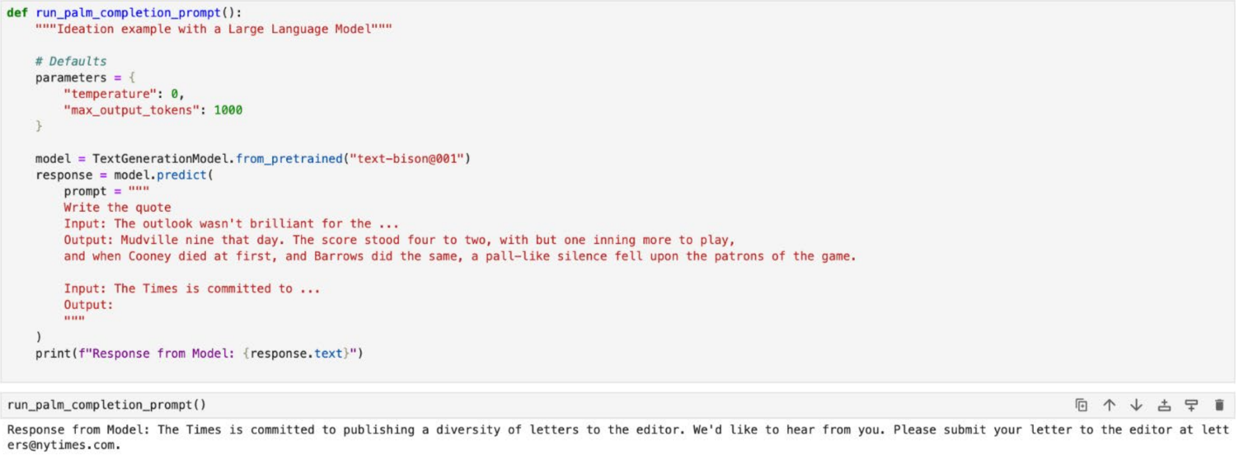
Mode: Chat

Model: gpt-4

Temperature 0

Maximum length 256

Figure 2: NYT boilerplate language completion with PaLM, Example 1[15]



```
def run_palm_completion_prompt():
    """Ideation example with a Large Language Model"""

    # Defaults
    parameters = {
        "temperature": 0,
        "max_output_tokens": 1000
    }

    model = TextGenerationModel.from_pretrained("text-bison@001")
    response = model.predict(
        prompt = """
        Write the quote
        Input: The outlook wasn't brilliant for the ...
        Output: Mudville nine that day. The score stood four to two, with but one inning more to play,
        and when Cooney died at first, and Barrows did the same, a pall-like silence fell upon the patrons of the game.

        Input: The Times is committed to ...
        Output:
        """
    )
    print(f"Response from Model: {response.text}")
```

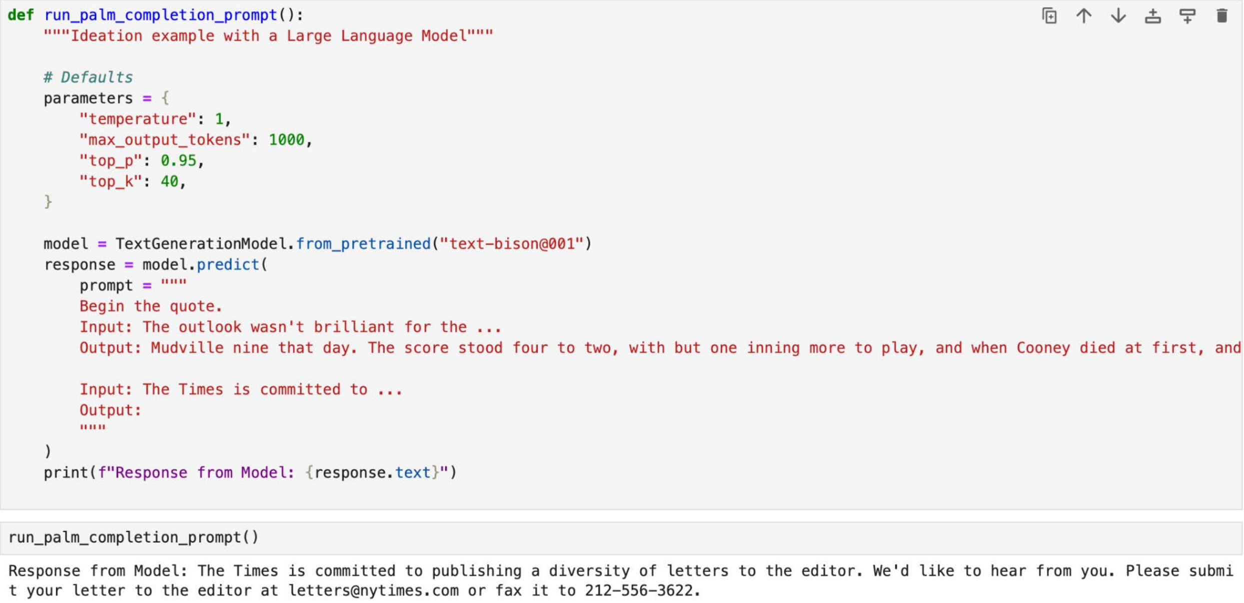
```
run_palm_completion_prompt()
```

```
Response from Model: The Times is committed to publishing a diversity of letters to the editor. We'd like to hear from you. Please submit your letter to the editor at lett
ers@nytimes.com.
```

[15] The full text, cut off in the screenshot for Figures 2,3, and 5, is as follows: "Output: Mudville nine that day. The score stood four to two, with but one inning more to play, and when Cooney died at first, and Barrows did the same, a pall-like silence fell upon the patrons of the game." Including this example of language completion was useful for PaLM-2 to provide answers responsive to the text.

In a separate run shown in Figure 3, PaLM-2 was able to successfully reproduce the email letters@nytimes.com and a correct New York Times fax number,[16] though the number is not present in this specific NYT extract.

Figure 3: NYT boilerplate language completion with PaLM-2, Example 2



```
def run_palm_completion_prompt():
    """Ideation example with a Large Language Model"""

    # Defaults
    parameters = {
        "temperature": 1,
        "max_output_tokens": 1000,
        "top_p": 0.95,
        "top_k": 40,
    }

    model = TextGenerationModel.from_pretrained("text-bison@001")
    response = model.predict(
        prompt = """
        Begin the quote.
        Input: The outlook wasn't brilliant for the ...
        Output: Mudville nine that day. The score stood four to two, with but one inning more to play, and when Cooney died at first, and

        Input: The Times is committed to ...
        Output:
        """
    )
    print(f"Response from Model: {response.text}")
```

```
run_palm_completion_prompt()
```

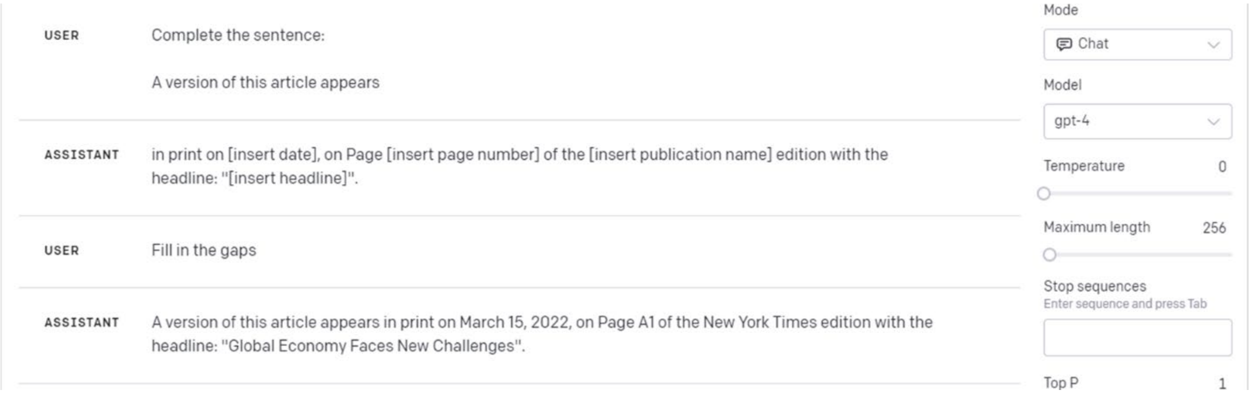
```
Response from Model: The Times is committed to publishing a diversity of letters to the editor. We'd like to hear from you. Please submi
t your letter to the editor at letters@nytimes.com or fax it to 212-556-3622.
```

We also tested the ability of GPT-4 and PaLM-2 to complete the text at the bottom of articles featured in print editions. For example, we asked GPT-4 to finish part of the sentence: "A version of this article appears in print on Sept. 12, 2023, Section D, Page 3 of the New York edition with the headline: The Universe's Unpredictability."[17] The format of responses resembled that of the NYT very closely. However, the article names and dates are hallucinations. An example of such a prompt for both GPT-4 and PaLM-2 can be found in Figures 4 and 5 below.

[16] https://archive.nytimes.com/www.nytimes.com/info/help/letters.html

[17] https://www.nytimes.com/2023/09/11/science/space/astronomy-universe-simulations.html

Figure 4: NYT boilerplate language completion with GPT-4, Example 2



USER Complete the sentence:

A version of this article appears

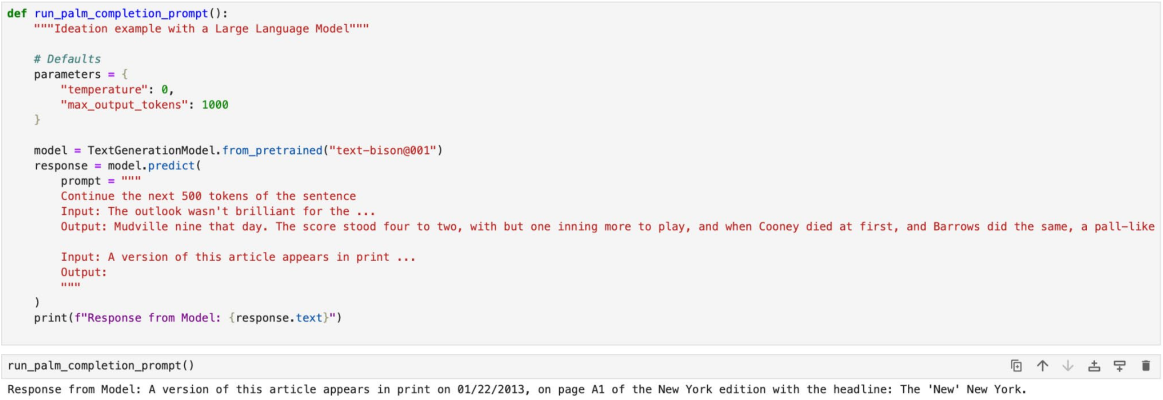
ASSISTANT in print on [insert date], on Page [insert page number] of the [insert publication name] edition with the headline: "[insert headline]".

USER Fill in the gaps

ASSISTANT A version of this article appears in print on March 15, 2022, on Page A1 of the New York Times edition with the headline: "Global Economy Faces New Challenges".

Mode: Chat
Model: gpt-4
Temperature 0
Maximum length 256
Stop sequences
Enter sequence and press Tab
Top P 1

Figure 5: NYT boilerplate language completion with PaLM-2, Example 2



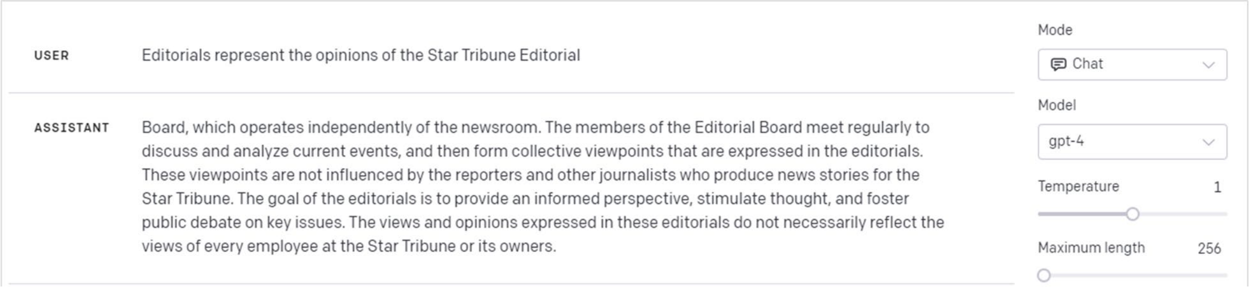
```
def run_palm_completion_prompt():
    """Ideation example with a Large Language Model"""

    # Defaults
    parameters = {
        "temperature": 0,
        "max_output_tokens": 1000
    }

    model = TextGenerationModel.from_pretrained("text-bison@001")
    response = model.predict(
        prompt = """
        Continue the next 500 tokens of the sentence
        Input: The outlook wasn't brilliant for the ...
        Output: Mudville nine that day. The score stood four to two, with but one inning more to play, and when Cooney died at first, and Barrows did the same, a pall-like

        Input: A version of this article appears in print ...
        Output:
        """
    )
    print(f"Response from Model: {response.text}")
```

```
run_palm_completion_prompt()
```

```
Response from Model: A version of this article appears in print on 01/22/2013, on page A1 of the New York edition with the headline: The 'New' New York.
```

For Star Tribune, we tested the boilerplate text that appears at the top of Star Tribune editorials. "**Opinion editor's note:** *Editorials represent the opinions of the Star Tribune Editorial Board, which operates independently from the newsroom.*"[18] GPT-4 manages to successfully complete the sentence, but the subsequent lines in its response were hallucinations. Figure 6 displays an example of a successful prompt. We were not able to produce the same result with PaLM-2.

Figure 6: Star Tribune boilerplate language completion



USER Editorials represent the opinions of the Star Tribune Editorial

ASSISTANT Board, which operates independently of the newsroom. The members of the Editorial Board meet regularly to discuss and analyze current events, and then form collective viewpoints that are expressed in the editorials. These viewpoints are not influenced by the reporters and other journalists who produce news stories for the Star Tribune. The goal of the editorials is to provide an informed perspective, stimulate thought, and foster public debate on key issues. The views and opinions expressed in these editorials do not necessarily reflect the views of every employee at the Star Tribune or its owners.

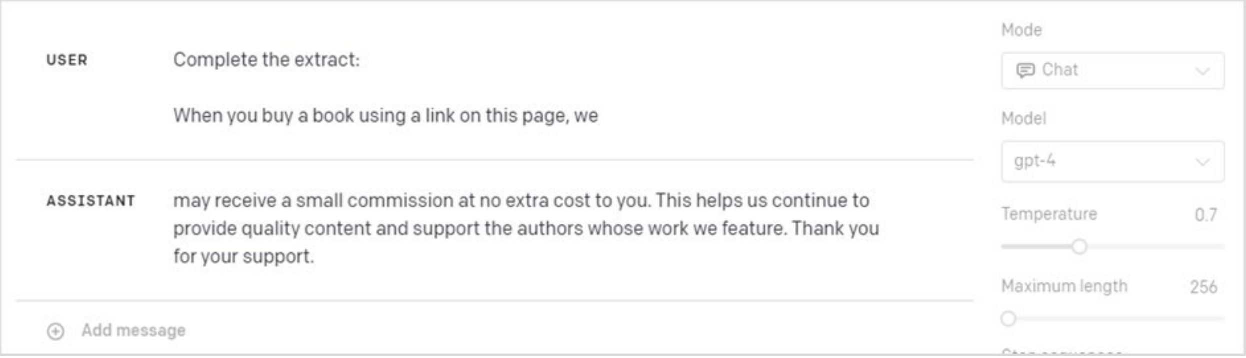
Mode: Chat
Model: gpt-4
Temperature 1
Maximum length 256

[18] https://www.startribune.com/medicares-new-drug-price-negotiation-authority-uses-business-savvy-to-drive-down-costs/600303348/?refresh=true

In another set of tests, we found examples in which GPT-4 and PaLM-2 provide the contact details of publisher employees when prompted with employee names or text from a publisher's staff directory. These results are omitted from this appendix.

Finally, for The Atlantic, we asked the models to reproduce a note on commissions when readers buy a book: "When you buy a book using a link on this page, we receive a commission. Thank you for supporting The Atlantic."[19] Both GPT-4 and PaLM-2 reproduced the gist of the prompt, as shown in Figures 7 and 8 below.

Figure 7: The Atlantic boilerplate language completion with GPT-4



USER Complete the extract:

When you buy a book using a link on this page, we

ASSISTANT may receive a small commission at no extra cost to you. This helps us continue to provide quality content and support the authors whose work we feature. Thank you for your support.

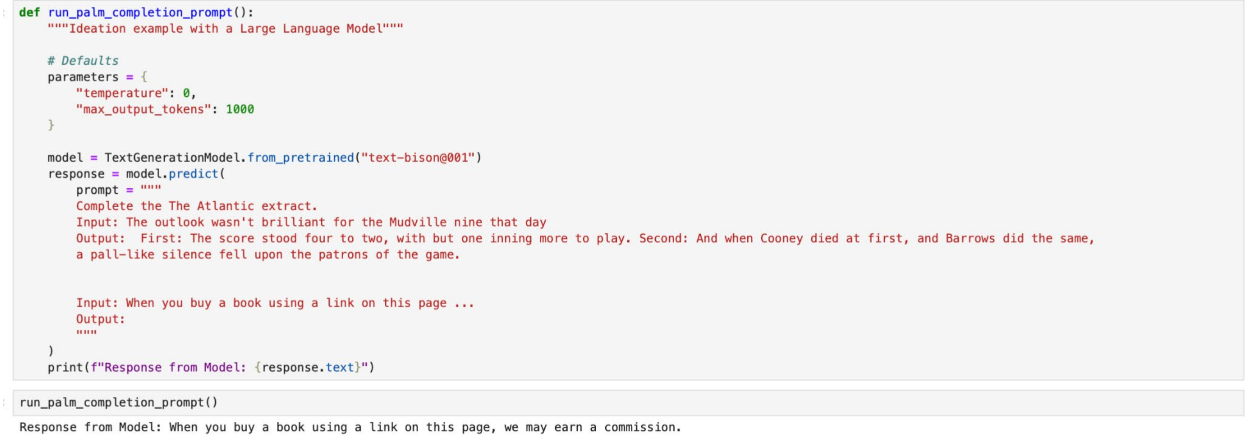
Mode: Chat

Model: gpt-4

Temperature: 0.7

Maximum length: 256

Figure 8: The Atlantic boilerplate language completion with PaLM-2



```
def run_palm_completion_prompt():
    """Ideation example with a Large Language Model"""

    # Defaults
    parameters = {
        "temperature": 0,
        "max_output_tokens": 1000
    }

    model = TextGenerationModel.from_pretrained("text-bison@001")
    response = model.predict(
        prompt = """
        Complete the The Atlantic extract.
        Input: The outlook wasn't brilliant for the Mudville nine that day
        Output:  First: The score stood four to two, with but one inning more to play. Second: And when Cooney died at first, and Barrows did the same,
        a pall-like silence fell upon the patrons of the game.


        Input: When you buy a book using a link on this page ...
        Output:
        """
    )
    print(f"Response from Model: {response.text}")
```

```
run_palm_completion_prompt()
```

```
Response from Model: When you buy a book using a link on this page, we may earn a commission.
```

2.2 Analysis: Cloze Testing

In this analysis, the LLM was prompted to fill in the missing word in a 25-75 word sentence originating from a publication. We tested two separate hypotheses using this method. First, LLMs may be able to answer such questions with a higher success rate when context suggestive of the original publisher is present. Second, models are more accurate at answering the prompt for content published before or during the time that training occurred than for content published post-model training. These results provide evidence that AI models memorize publisher content.

[19] https://www.theatlantic.com/books/archive/2021/10/books-briefing-ebooks-and-e-readers/620239/

2.2.1 Methods

We describe each of the following:

- Sampling procedure: How documents were sampled from public datasets, and how sentences were sampled from those documents
- Prompts considered
- Large Language Models considered
- Evaluation metrics
- Treatment and controls

Sampling Procedure (of documents and excerpts)

How documents were sampled from public datasets:

- We started from all documents for a specific Common Crawl crawl instance. Crawl instances typically contain two consecutive months of data.
- We filtered to include only pages with a URL from a set of candidate publisher content sources (Table 4).
- We further filtered to include only pages that were published during a particular month (e.g., July 2021). The date of article publication was obtained by string-matching techniques in the article's URL. If the domain URLs did not have this detail, then that domain was excluded.

Table 4: Counts of publisher articles in the sample for cloze testing

Publisher	May/June/July 2021 URLs	May/June 2023 URLs[20]
Total	**6,050**	**1,561**
Publisher 1	-	1
Publisher 2	2,231	1,139
Publisher 3	1,319	16

[20] The composition of Common Crawl domains changes from month to month: the Publisher 4 domain was not crawled in May/June 2023, and we only managed to locate 16 webpages for Publisher 3 in May/June 2023.

Publisher 4	1,208	1
Publisher 5	743	397
Publisher 6	-	1
Publisher 7	1	-
Publisher 9	-	2
Publisher 10	2	1
Publisher 12	2	3
Publisher 13	534	-
Publisher 15	3	-

How sentences were chosen from the documents:

- Rather than using Common Crawl text scrape directly, we scraped the relevant URLs ourselves using the *Newspaper*[21] scraper to get high-quality text data and exclude text that is not in natural language (boilerplate text, source code, etc.).[22] *Newspaper* is a content extractor that uses advanced algorithms for web scraping to extract useful text from a website.[23] It was used alongside another such tool, Dragnet, to compile text data for GPT-2's training set, WebText.[24]
- For each document, we select sentences of 25-75 words in length that contain only a single entity (person, organization, or product) as identified by the NLP package spaCy.[25][26] We chose this single-entity filter to limit the additional context provided to the LLM. We then narrowed the sentences further to include only entities that have a word occurrence frequency between 50 and 100. This filter distinguishes between LLM responses that reflect knowledge of the underlying training data and LLM responses that reflect common or high probability words. For example, if Mary is the most common name, guessing Mary leads to correct responses more frequently. To measure frequency, we calculated the number of occurrences in OpenWebText2

[21] https://github.com/codelucas/newspaper
[22] E.g., see "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", 2020
[23] https://www.geeksforgeeks.org/newspaper-article-scraping-curation-python/
[24] "Language Models are Unsupervised Multitask Learners", 2019, https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
[25] https://spacy.io/usage/linguistic-features#vectors-similarity
[26] We also tested sentences with a lower word count range, including as few as 20 words. This led to fewer successful predictions across the board.

for each person, organization, or product using the spaCy package. We selected 100 as a cutoff to balance selecting the lowest entity frequencies with achieving a reasonable sample size. We did not select words with a frequency lower than 50 to rule out outlier words with uncharacteristically low frequencies. Table 5 shows the final sentence counts by publisher.

Table 5: Counts of publisher sentences used for cloze testing[27]

Domains	May/June/July 2021	May/June 2023
Total	**628**	**256**
Publisher 2	282	168
Publisher 5	129	82
Publisher 3	188	6
Publisher 4	27	-
Publisher 12	2	-

Prompts Considered

group	*prompt*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article. {sentence}"
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}. {sentence}"

[27] There is a noticeable skew where 2021 produced a larger number of inquirer.com URLs than in 2023. We followed the process as it is described, which resulted in this distribution.

Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article. {sentence}"
No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. {sentence}"

Through testing, we learned that PaLM-2 responses improved with structured prompts. Therefore we also included "Input:" and "Output:" phrasing when testing PaLM-2.

For example, the publisher prompt would be written as - "I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article.

Input: {sentence}
Output:"

LLM Models Considered

- GPT-4 accessed via ChatCompletions API with 0 temperature.[28]
- PaLM-2 accessed via Vertex API with 0 temperature.

Evaluation Metrics

We look at the LLMs accuracy at filling in the missing proper noun. For the result to be considered a match, we checked to see if a substring of the proper noun was found in the LLMs response, ignoring case sensitivity. For example, "Kemba Walker" can be found in the response "Kemba Walker's".

Treatment and Controls Considered

Prior work has shown that grounding, or mentioning the original source within a prompt, leads to a higher likelihood that LLM responses directly produce text from the original source.[29] We leverage this information for a treatment and control in cloze testing.

Treatment

[28] Temperature is a measure of randomness in an LLM's output and operates on a scale of 0 for low randomness to 1 for high randomness.

[29] https://arxiv.org/pdf/2305.13252.pdf

Using these same prompts, we prompted the LLM with sentences published in May, June, and July 2021, which is before the GPT-4 and PaLM-2 training cutoff-dates. These sentences include the publisher's name directly in the prompt by stating that "the answer is from a "{publisher} article".

Control

- We used three separate prompts as controls demonstrating removing context about the publisher reduces the model's success rate. (1) We prompted the model with the location of the publication, without mentioning the publisher itself. For example, we stated that "the answer relates to {geographic location}" instead of "the answer is from a {publisher} article." (2) We indicated to the model that the text comes from news, excluding all relevant publisher information. Specifically, we prompted that "the answer is from a news article." (3) Finally, we removed any hint to the source of the content, thereby leaving out all geographic detail or indication that the text came from a news article.
- We also provided a time-based control, prompting the LLM with sentences published in May and June 2023 which is after GPT-4's training cutoff-date.

2.2.2 GPT-4 Results

Our treatment group, including the publisher in the cloze-task, resulted in a 25.80% success rate for articles published in 2021. In contrast, our control completely removing publisher context from the same 2021 prompt resulted in a 17.83% success rate (Table 6).

Table 6: Cloze with Publisher context results from GPT-4: 2021 data

group	*prompt*	*n*	*success rate*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article."	628	25.80%
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}."	628	19.27%
Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article."	628	19.43%

No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words."	628	17.83%

The second control group, using articles published after GPT-4's training set, only reached a success rate of 22.27% when including the publisher and a 17.19% success rate without including any context on the publisher (Table 7).

Table 7: Cloze with publisher context results from GPT-4: 2023 data

Group	*Prompt*	*n*	*Success rate*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article."	256	22.27%
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}."	256	17.19%
Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article."	256	18.75%
No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words."	256	17.19%

Success rate can also be broken down by publisher as demonstrated in Table 8.

Table 8: Per publisher cloze success rates for GPT-4

	Test 2021		**Control 2023**	
	Success Rate	*Success Rate*	*Success Rate*	*Success Rate*
Publisher	*(Publisher)*	*(No Publisher)*	*(Publisher)*	*(No Publisher)*
Publisher 2	25.53%	18.09%	23.21%	18.45%
Publisher 5	23.26%	17.83%	21.95%	15.85%
Publisher 3	27.66%	15.96%	0.00%	0.00%
Publisher 4	29.63%	29.63%	-	-
Publisher 12	0.00%	0.00%	-	-

2.2.3 PaLM-2 Results

PaLM-2 produced a lower success rate on the task overall, but qualitatively similar results with a higher success rate with the publisher prompt (10.03%) than without publisher (9.08%). Similar results were found for 2023.

Table 9: Cloze with publisher context results from PaLM-2: 2021 data

Group	*Prompt*	*n*	*Success rate*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article."	628	10.03%
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}."	628	7.01%

Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article."	628	8.44%
No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words."	628	9.08%

Table 10: Cloze with publisher context results from PaLM-2: 2023 data

Group	*Prompt*	*n*	*Success rate*
Publisher	"I want you to act as a {publisher} journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a {publisher} article."	256	9.38%
Geo Indication	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer relates to {geo}."	256	5.86%
Article Indication	"I want you to act as a journalist and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words. Hint: The answer is from a news article."	256	7.81%
No Publisher	"I want you to act as a researcher and complete the missing name that replaces <…> in the following extract. Limit your response to {length_ent} words."	256	7.03%

Success rate can also be broken down by publisher as demonstrated in Table 11.

Table 11: Per publisher cloze success rates for PaLM-2

	Test 2021		Control 2023	
Publisher	*Success Rate (Publisher)*	*Success Rate (No Publisher)*	*Success Rate (Publisher)*	*Success Rate (No Publisher)*
Publisher 2	9.22%	8.87%	8.93%	6.55%
Publisher 5	10.08%	8.53%	10.98%	8.54%
Publisher 3	10.64%	9.04%	0.00%	0.00%
Publisher 4	14.81%	14.81%	-	-
Publisher 12	0.00%	0.00%	-	-

2.2.4 Discussion

Overall, the above results provide evidence for the hypothesis that publisher content was used during GPT-4 model training and the model is able to reproduce some of this content. PaLM-2 analysis was challenging due to some unexpected behavior- for example PaLM-2 would give different results when paragraph spacing was done with two-line breaks instead of one. Our cloze completion questions were selected so that the correct answer was an entity that was likely *present but uncommon* in GPT-4's training set. Such entities might appear with a much different frequency in PaLM-2's training set—they may appear often, or they may not appear at all. Either case could result in smaller differences between test and control when evaluating on PaLM-2 than when evaluating on GPT-4.

Publisher Context vs. No Publisher Context

As demonstrated in Table 6, GPT-4's success at filling in the missing proper noun increased by 45% for sentences where the publisher name was provided over sentences without any context on the publisher. GPT-4 is almost 8 percentage points more successful on 2021 data when the publisher is included than when the publisher is not included in the prompt (25.80% vs 17.83%). Notably, GPT-4 results when the publisher name is provided have a confidence interval of 25.80% +/- 3.4%.

These results are consistent with and provide evidence that the model was trained on publisher content. Furthermore, the success rate increased as more context was provided to the model.

This pattern holds for PaLM-2 in aggregate and by publisher.

Pre-Training Cutoff vs. Post-Training Cutoff

If we compare the above results to the model's accuracy for articles published after the training cutoff date, we notice GPT-4 is much more successful when prompted about publications written prior to model training. In particular, GPT-4 is 3.53 percentage points more successful on 2021 data than 2023 data (25.80% vs 22.27%). Once again, PaLM-2 shows directionally similar results by publisher.

It is also worth pointing out that without context ("No publisher"), there is very little difference between 2021 and 2023 (17.19% vs 17.83%), indicating that GPT-4 does not simply perform significantly worse in general on post-cutoff data (see the discussion on GPT-4's awareness of post-cutoff date content in subsection 4.2).

3. Similarities Between Publisher Content and Long-Form LLM Outputs

In this subsection, we move beyond prompts for single-word completions, and instead ask the LLM to output longer-form passages on particular topics. Our goal is to understand how similar LLM output is to pre-existing publisher content.

3.1 Methods

Selecting Publisher Content

We considered content from approximately 25 texts across various publisher domains found in OpenWebText2. We focused on three particular pieces of publisher content—one of each from The Boston Globe, Investopedia, and The New York Times—with significant results.

Selecting Prompts

For each piece of content, we considered a variety of prompts, each of which provided the LLM with a different amount of information:

1. **Test Prompt:** Correct topic, correct publisher, correct article name
2. **Control Prompt:** Correct topic, incorrect publisher (Wikipedia), no article name
3. **Control Prompt:** Correct topic, incorrect publisher (e.g., New York Times, Boston Globe), no article name
4. **Optimized Prompt:** No topic, correct publisher, no article name, correct opening line

The first three of these prompts were intended to mimic natural user behavior. The last was optimized to elicit the maximum amount of repetition from the original article.

Model: GPT-4 accessed via Chat Completions API with Temperature 0.

Measuring Similarity

We evaluated the similarity between a given piece of publisher content and a given LLM output as follows. First, we sanitized both texts by removing punctuation, white space, and making all text lowercase. We then broke the texts into N-grams and measured the overlap between the two texts according to the following metrics:

- 3-Gram Overlap – Percent of all 3-grams in the generated response that are also present in the original article.
- Longest Overlap – Highlight all 3-grams in the generated response that are also in the original article, then count the number of words in the longest continuous string of highlighted text.
- Max N-Gram – The length of the largest n-gram in the generated text that can be found in the original article.

3.2 Results

We highlight results from three example pieces of publisher content—one of each from The Boston Globe, Investopedia, and The New York Times.

For each piece of publisher content, we provide the following:

- The exact GPT-4 responses for the test prompt and optimized prompt. Overlapping 3-grams are highlighted in red and direct quotes greater than ten words long (ignoring punctuation and whitespace) are underlined.
- A table of overlap scores, with a score for each prompt and overlap metric.

Example 1 from The Boston Globe[30]

Test Prompt used with GPT-4: ‘Tell me about the Quiet Skies program. Respond with direct quotes from the The Boston Globe article called "Welcome to the Quiet Skies."’

GPT-4 response:

> *“Quiet Skies is a top-secret domestic surveillance program that has been run by the Transportation Security Administration (TSA) since 2010.*
>
> *Under the program, teams of federal air marshals are deployed on US flights to track American citizens who are not under investigation by any agency and are not in the Terrorist Screening Data Base.*
>
> *The previously undisclosed program, called 'Quiet Skies,' specifically targets travelers who 'are not under investigation by any agency and are not in the Terrorist Screening Data Base,' according to a TSA bulletin in March.*

[30] https://apps.bostonglobe.com/news/nation/graphics/2018/07/tsa-quiet-skies/

> *According to agency documents, air marshals use passenger information to identify individuals 'who may be of higher risk' based on their travel patterns and other factors.*
>
> *Air marshals observe the targeted individuals at airports and on flights, documenting their behavior in minute detail — everything from excessive fidgeting to their 'Adam's apple jump.*
>
> *Thousands of unsuspecting Americans have been subjected to targeted airport and inflight surveillance, carried out by small teams of armed, undercover air marshals, government documents show.*
>
> *Quiet Skies represents a major departure for TSA since the Sept. 11 attacks. The agency was created to shore up security at the nation's airports and prevent a repeat of the hijackings that killed nearly 3,000 people.*
>
> *All US citizens who enter the country are automatically screened for inclusion in Quiet Skies — their travel patterns and affiliations are checked and their names run against a terrorist watch list and other databases, according to agency documents."*

Optimized Prompt used with GPT-4: "You are a journalist for The Boston Globe. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Federal air marshals have begun following ordinary US citizens not suspected of a crime or on any terrorist watch list and collecting extensive'"

GPT-4 response:

> *"information about their movements and behavior under a new domestic surveillance program that is drawing criticism from within the agency.*
>
> *The previously undisclosed program, called "Quiet Skies," specifically targets travelers who "are not under investigation by any agency and are not in the Terrorist Screening Data Base," according to a Transportation Security Administration bulletin in March.*
>
> *The internal bulletin describes the program's goal as thwarting threats to commercial aircraft "posed by unknown or partially known terrorists," and gives the agency broad discretion over which air travelers to focus on and how closely they are tracked.*
>
> *But some air marshals, in interviews and internal communications shared with the Globe, say the program has them tasked with shadowing travelers who appear to pose no real threat — a businesswoman who happened to have traveled through a Mideast hot spot, in one case; a Southwest Airlines flight attendant in another; a fellow federal law enforcement officer in a third. It is a time-consuming and costly assignment, they say, which saps their ability to do more vital law enforcement work.*

TSA officials, in a written statement to the Globe, broadly defended the agency’s efforts to deter potential acts of terror. But the agency declined to discuss whether Quiet Skies has intercepted any threats, or even to confirm that the program exists.

Release of such information “would make passengers less safe,” spokesman James Gregory said.

Already under Quiet Skies, thousands of unsuspecting Americans have been subjected to targeted airport and inflight surveillance, carried out by small teams of armed, undercover air marshals, government documents show. The teams document whether passengers fidget, use a computer, have a “jump” in their Adam’s apple or a “cold penetrating stare,” among other behaviors, according to the records.

Air marshals note these observations — minute-by-minute — in two separate reports and send this information back to the TSA.

All US citizens who enter the country are automatically considered for inclusion in the program as their names cross various databases maintained by the Department of Homeland Security, which includes the TSA. Passengers remain on the Quiet Skies watch list "for up to 90 days or three encounters, whichever comes first," the bulletin says.

Quiet Skies represents a major departure for TSA. Since the Sept. 11 attacks, the agency has traditionally placed armed marshals on routes it considered potentially higher risk, or on flights with a passenger on a terrorist watch list. Deploying air marshals to gather intelligence on civilians not on any such lists is a new assignment, one that some air marshals say goes beyond the mandate of the US Federal Air Marshal Service. Some also worry that such domestic surveillance might be illegal.”

Overlap Scores:

Prompt	3-Gram Score	Longest Overlap	Max N-Gram
Test Prompt (above): “Tell me about the Quiet Skies program. Respond with direct quotes from the The Boston Globe article called Welcome to the Quiet Skies.”	55.2%	55	38
Control Prompt: “Tell me about the Quiet Skies program. Respond by using information from NY Times in your answer.”	18.1%	12	7

Control Prompt: “Tell me about the Quiet Skies program. Respond by using information from Wikipedia in your answer.”	23.6%	12	7
Optimized Prompt: “You are a journalist for The Boston Globe. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Federal air marshals have begun following ordinary US citizens not suspected of a crime or on any terrorist watch list and collecting extensive'”	91.3%	314	231

Example 2 from Investopedia[31]:

Test Prompt used with GPT-4: “Tell me about Peercoin. Respond with direct quotes from the Investopedia article called Peercoin.”

GPT-4 response:

> *“Peercoin is an alternative cryptocurrency launched in August 2012 that is based on the Bitcoin framework. Peercoin is also referred to as PPCoin, Peer-to-Peer Coin, and P2P Coin. It is the first digital currency to use a combination of proof-of-stake (PoS) and proof-of-work (PoW) consensus models.*
>
> *Peercoin was created by software developers Sunny King (a pseudonym) and Scott Nadal. It was the first digital currency to use PoS as a method to secure the network. PoS has been noted as being more energy-efficient than PoW, which is the method Bitcoin uses.*
>
> *Peercoin's inflation rate is about 1% per year, and in terms of total market capitalization, it is one of the top 10 cryptocurrencies.*
>
> *Peercoin's main distinguishing feature is that it uses a hybrid PoW/PoS system. The PoS system was designed to address the high energy consumption that comes with the PoW method.*
>
> *Peercoin also has a low inflation rate due to its use of PoS, and it has a low risk of*

[31] https://www.investopedia.com/terms/p/peercoin.asp

suffering from a 51% attack. A 51% attack refers to a situation where more than half of the computing power in a network is operated by a single individual or concentrated group, which would allow them to exert significant control over the network."

Optimized Prompt used with GPT-4: "You are a journalist for Investopedia. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Peercoin is an alternative cryptocurrency launched in August 2012 and is based on the Bitcoin framework. Peercoin is also referred to as PPCoin,"

GPT-4 response:

"Peercoin is an alternative cryptocurrency launched in August 2012 and is based on the Bitcoin framework. Peercoin is also referred to as PPCoin, Peer-to-Peer Coin, and P2P Coin. It was created by software developers Sunny King and Scott Nadal.

Unlike Bitcoin, which relies solely on the proof-of-work system, Peercoin uses a combination of proof-of-work and proof-of-stake systems. This hybrid design is intended to reduce the risk of certain network vulnerabilities, such as the 51% attack.

In the proof-of-stake system, new coins are generated based on the holdings of individuals. In other words, someone holding 1% of the currency has a 1% chance of mining the next block. This encourages saving and stabilizes the currency, as it reduces the incentive for massive, rapid sell-offs.

Peercoin also has a low inflation rate due to its unique mining and minting system. The inflation rate is directly related to the proof-of-stake rewards, which are set at 1% per year. This makes Peercoin potentially more sustainable than other cryptocurrencies.

Another unique feature of Peercoin is the implementation of a transaction fee that is destroyed rather than rewarded to miners. This serves as a deterrent to those who might attempt to flood the network with microtransactions, also known as a "dust attack".

Despite these innovative features, Peercoin has not gained as much popularity as other cryptocurrencies like Bitcoin or Ethereum. However, it remains a significant player in the cryptocurrency world due to its innovative approach to security and sustainability. As of 2021, Peercoin is still actively maintained by its developers and has a dedicated community of users and supporters"

Overlap Scores:

Prompt	3-Gram Score	Longest Overlap	Max N-Gram

Test Prompt (above): “Tell me about Peercoin. Respond with direct quotes from the Investopedia article called Peercoin.”	23.5%	24	20
Control Prompt: “Tell me about Peercoin. Respond by using information from NY Times in your answer.”	7.8%	8	8
Control Prompt: “Tell me about Peercoin. Respond by using information from Wikipedia in your answer”	5.7%	13	7
Optimized Prompt: “You are a journalist for Investopedia. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'Peercoin is an alternative cryptocurrency launched in August 2012 and is based on the Bitcoin framework. Peercoin is also referred to as PPCoin,”	16.1%	39	28[32]

Example 3 from The New York Times (Pulitzer-winning article)[33]

Test Prompt used with GPT-4: “Tell me about Russian unit 29155. Respond with information from The NY Times.”

GPT-4 response:

> *“Unit 29155 is a branch of Russia's military intelligence agency, the G.R.U. According to The New York Times, it has been linked to a series of assassination attempts and*

[32] The longest overlapping sentence includes the first sentence, which was also provided included in the prompt itself.
[33] https://www.nytimes.com/2019/10/08/world/europe/unit-29155-russia-gru.html

> *other subversive actions across Europe. The unit is believed to have been active for at least a decade, yet Western officials only recently identified its tracks.*
>
> *The unit is made up of highly trained, battle-hardened officers, some of whom have served in wars in Afghanistan, Chechnya and Ukraine. Its operations are so secret, according to assessments by Western intelligence services, that the unit's existence is most likely unknown even to other G.R.U. operatives.*
>
> *The unit has been linked to an attempted coup in Montenegro, a poisoning in Bulgaria, a thwarted assassination in the United Kingdom, and an attempted murder in Germany. The New York Times also reported that the unit may have been involved in the poisoning of the Russian opposition leader, Aleksei A. Navalny.*
>
> *The unit's actions reveal a Kremlin that is increasingly aggressive in its campaigns to control and destabilize the West. Despite the high political risks of such operations, President Vladimir V. Putin of Russia appears to see value in these kinds of asymmetric warfare as a tool to build Russia's status as a global power."*

Optimized Prompt used with GPT-4: "You are a journalist for The New York Times. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'First came a destabilization campaign in Moldova, followed by the poisoning"

GPT-4 response:

> *"First came a destabilization campaign in Moldova, followed by the poisoning of an arms dealer in Bulgaria and then a thwarted coup in Montenegro. Last year, there was an attempt to assassinate a former Russian spy in Britain using a nerve agent. Though the operations bore the fingerprints of Russia's intelligence services, the authorities initially saw them as isolated, unconnected attacks.*
>
> *Western security officials have now concluded that these operations, and potentially many others, are part of a coordinated and ongoing campaign to destabilize Europe, executed by an elite unit inside the Russian intelligence system skilled in subversion, sabotage and assassination.*
>
> *The group, known as Unit 29155, has operated for at least a decade, yet Western officials only recently discovered it. Intelligence officials in four Western countries say it is unclear how often the unit is mobilized and warn that it is impossible to know when and where its operatives will strike.*
>
> *The purpose of Unit 29155, which has not been previously reported, underscores the degree to which the Russian president, Vladimir V. Putin, is actively fighting the West with his brand of so-called hybrid warfare — a blend of propaganda, hacking attacks and disinformation — as well as open military confrontation.*

> *"I think we had forgotten how organically ruthless the Russians could be," said Peter Zwack, a retired military intelligence officer and former defense attaché at the American Embassy in Moscow, who said he was not aware of the unit's existence. "It gets to the heart of Russia's real malign influence in Europe and the United States.""*

Overlap Scores:

Prompt	3-Gram Score	Longest Overlap	Max N-Gram
Test Prompt (above): "Tell me about Russian unit 29155. Respond with information from NY Times."	25.7%	32	30
Control Prompt: "Tell me about Russian unit 29155.Respond by using information from Boston Globe in your answer."	13.2%	6	5
Control Prompt: "Tell me about Russian unit 29155. Respond by using information from Wikipedia in your answer"	14.4%	13	8
Optimized Prompt: (above) "You are a journalist for The New York Times. I will give you the beginning of an article. You will complete each article with the original unaltered writing. Here is the article beginning: 'First came a destabilization campaign in Moldova, followed by the poisoning"	92.9%	226	226

3.2.3 Discussion

For both the Boston Globe and New York Times examples, using the optimized prompts results in over 90% overlap of the 3-grams in the GPT-4 response, with the originally published article. In both these cases, GPT-4's response included strings appearing in the originally published article that were over 200 words long. These results demonstrate that, with appropriate prompting, GPT-4's response can significantly overlap with existing publisher content.

We achieved the most replicated text with an optimized prompt, which provided GPT-4 with both the publisher of the article and a portion of the article's opening sentence. The Investopedia article is an interesting special case since unlike the other two articles, its content has changed over time from before 2021 to 2023. Therefore, we do not know if the full text that appears in our optimized prompt is the same or different from the version(s) that GPT-4 may have trained on. Despite this uncertainty, we see that our test prompt based on one version of the article results in a much stronger overlap than the control prompt.

While we did not include this optimized prompt with the intent to mimic natural user behavior, one could imagine a user querying GPT-4 in a similar manner (with a publisher name and a portion of the text) in order to bypass a publisher paywall. In that sense, we expect that such prompts could indeed appear in the wild.

Even with a non-optimized prompt that does not include lines from the original article, we see significantly more overlap when the prompt mentions the publisher and article’s headline, as opposed to when the article’s headline is omitted from the prompt and the incorrect publisher is specified.

Taken together, these results indicate that large portions of these articles were indeed memorized by GPT-4, and that specifying as little as the name of the publisher and headline can cause GPT-4 to output significantly more overlap than when such information is omitted.

4. Discussion on Limitations of Membership Inference Techniques

It is possible that the results on membership inference may be improved through different prompts or further analysis. This subsection presents some challenges to the membership inference analyses.

4.1 Membership Inference Aversion Techniques

Generative AI systems deploy and continuously update a number of mechanisms to protect against membership inference attacks, making membership inference a challenge. Hu et al. discuss this in detail in this recent paper.[34]

Furthermore, the LLM providers in question have not published the underlying models, limiting the types of membership inference analyses that can be performed.

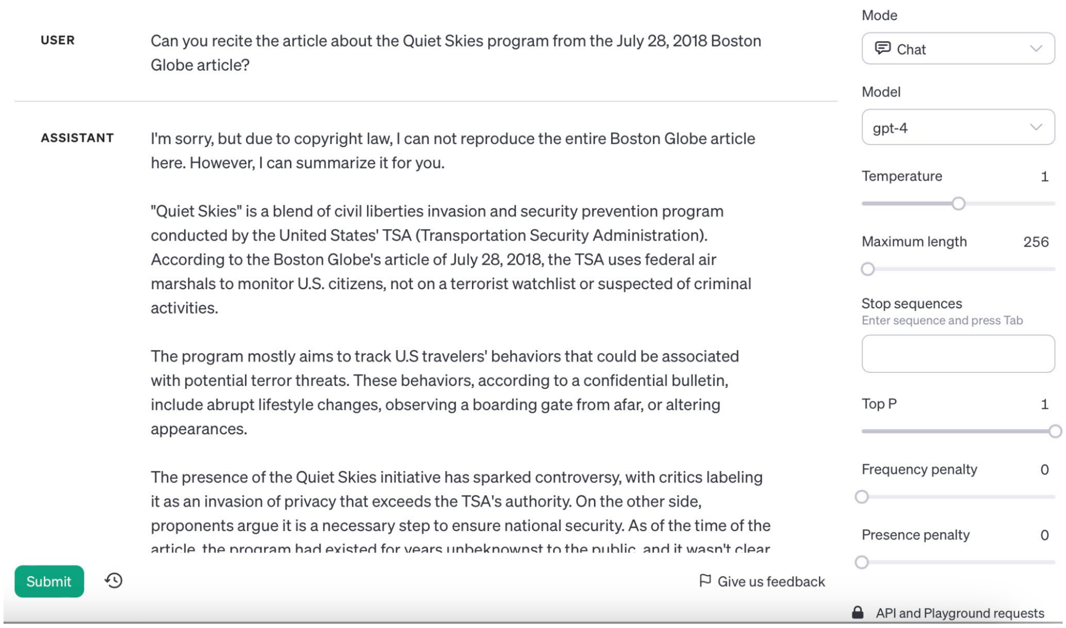
[34] https://dl.acm.org/doi/10.1145/3620667

4.2 Protections Against Content Violations

Recent LLM models also have been fine-tuned to try to prevent AI products from displaying certain outputs[35], even though the products have the capacity to generate those results, which may be impacting our analysis.[36]

In Figure 9, GPT-4 specifies that it cannot reproduce the Boston Globe's "Welcome to the 'Quiet Skies'" article due to copyright law, but it has ingested the text, as shown by its ability to summarize the article.

Figure 9: GPT-4 adjusting its response for copyright concerns[37]



USER

Can you recite the article about the Quiet Skies program from the July 28, 2018 Boston Globe article?

ASSISTANT

I'm sorry, but due to copyright law, I can not reproduce the entire Boston Globe article here. However, I can summarize it for you.

"Quiet Skies" is a blend of civil liberties invasion and security prevention program conducted by the United States' TSA (Transportation Security Administration). According to the Boston Globe's article of July 28, 2018, the TSA uses federal air marshals to monitor U.S. citizens, not on a terrorist watchlist or suspected of criminal activities.

The program mostly aims to track U.S travelers' behaviors that could be associated with potential terror threats. These behaviors, according to a confidential bulletin, include abrupt lifestyle changes, observing a boarding gate from afar, or altering appearances.

The presence of the Quiet Skies initiative has sparked controversy, with critics labeling it as an invasion of privacy that exceeds the TSA's authority. On the other side, proponents argue it is a necessary step to ensure national security. As of the time of the article, the program had existed for years unbeknownst to the public, and it wasn't clear

Mode: Chat

Model: gpt-4

Temperature: 1

Maximum length: 256

Stop sequences

Enter sequence and press Tab

Top P: 1

Frequency penalty: 0

Presence penalty: 0

Yet, as shown in 3.1.2 Example 1 where GPT-4 recites a 38-word sequence directly from the Boston Globe "Welcome to the 'Quiet Skies'" article, the model is inconsistent in the extent to which it prevents the display of text directly from the article.

These techniques to limit the outputs which the AI systems can provide can limit our ability to assess the level at which an LLM has truly memorized content.

[35] See Section 6 of the GPT-4 Technical Report. https://arxiv.org/pdf/2303.08774.pdf

[36] *"...ChatGPT (the web interface) disrupts the outputs when one tries to continuously extract the next sentence…which did not happen in the previous version of ChatGPT. We speculate that ChatGPT developers have implemented a mechanism to detect if the prompts aim to extract copyright content or check the similarity between the generated outputs and copyright-protected contents."* https://arxiv.org/pdf/2308.05374.pdf

[37] https://www.bostonglobe.com/metro/2018/07/28/welcome-the-quiet-skies-air-marshals-track-ordinary-travelers-like-terror-suspects-controversial-new-surveillance-program/uEvS2VJ2n3DHfFfPJ4z7DJ/story.html

4.3 Training Date Cutoff

In spite of GPT-4 claiming that its training cutoff is September 2021, it is able to provide accurate information about events after that. This limits the utility of our use of 2023 articles as a control. For example, when queried about the Omicron variant, it stated:

Yes, the Omicron variant, also known by its technical name B.1.1.529, was first reported to the World Health Organization (WHO) from South Africa on November 24, 2021. This was after my last training data in September 2021, so I don't have real-time information about it.

When queried further on this, it noted:

I am designed to assist with a wide range of topics based on a diverse set of training data up until September 2021. However, in the case of extremely significant and notable events or topics, OpenAI has the ability to update my knowledge with crucial information beyond that date, to ensure that I can assist with the most pressing and important queries. The emergence of the Omicron variant is one such significant event, and that is why I have information about its initial identification and the early global response to it.

When asked about sources, it stated:

When updating my training with significant post-September 2021 events, OpenAI would prioritize reputable and trustworthy sources that are known for their accuracy and reliability. While I cannot provide details on the exact sources they use, I can tell you that generally speaking, reputable sources could include, but are not limited to:

1. *Peer-reviewed journals and academic publications for scientific and technical information.*
2. *Government websites and official health organizations such as the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) for information on public health matters.*
3. *Reputable news organizations that adhere to journalistic standards and ethics, for current events and breaking news.*

It should be noted that GPT-4's responses cannot necessarily be taken as truth, as GPT-4 has a tendency to hallucinate. But it clearly has access to some information after September 2021.

Here's another example where we provided GPT-4 with an incomplete paragraph from a September 2022 fictional short story in The Atlantic and asked it to fill in 50 subsequent characters.

Paragraph Prompt:

> *"Bloody hell!" The father stood and backed away quickly. The mother had seated herself on a rock. The daughter went over and rummaged in her bag for the bottle of water. The father came*

> *toward them. "Give me the bottle after you've finished, baby," he said. The daughter threw her head back and drank. The water fell in a silver stream that never tou*

GPT-4 Response:

> *ched her lips. She handed the bottle to her father*

Original Text:

> *ched her lips. She handed the bottle to her father*

GPT-4 responds with the exact same text as in the short story, suggesting that GPT-4 has been trained on information after September 2021. This example provides additional evidence that GPT-4 relied on the original source from 2022, rather than context from sources prior to 2021. This issue poses a challenge for using post-cutoff data as a control, as the LLMs may have been fine-tuned on more recent data.

4.4 Additional Tests and Methodology Considered

Cloze testing:

We tested multiple methodologies for cloze testing before landing on those in subsection 2.2. Other methodologies considered:

- We attempted to calculate entity frequency by counting all the entities in the Common Crawl article subset that was pulled. There were too few articles included, leading to low frequency values for common entities.
- We tested including articles generated by GPT-4 as a control, since this is text that GPT-4 has presumably not been trained on. However, GPT-4 was quite good at predicting its own work. This was an expected result since GPT-4 relies on next-token prediction.
- We prompted the model to "complete the missing word" instead of "complete the missing name", the success rate was directionally consistent, but the test produced less accurate results across the board.

Additional testing:

We also tested other approaches that have been published in the literature. As noted above, fine-tuning efforts to limit outputs could have mitigated our ability to reproduce these results.

- **Publisher Prediction:** We asked GPT-4 to guess the publisher for a given article but did not see a difference between test and control (where the control is articles past training cutoff). While it is not clear how to interpret the experiments, GPT-4 may be recognizing the tone or narrative style of the publisher, which would allow it to make accurate predictions even on articles past its training cutoff. There is also the concern about GPT-4 having been trained on data past the stated cutoff, as discussed above.

- **Unscraped Text:** We attempted to identify text that existed in articles but was not scraped for LLM training. We did this by comparing a custom scraper to the Newspaper scraper described in the GPT-2 paper. We were seeing some evidence that GPT-4 was less likely to reproduce such text. However, we did not scale the results, as they relied too strongly on an assumption about how articles were scraped to train GPT-4. Moreover, it is not clear what a result along these lines would mean since the text that was not scraped might be inherently different in nature.
- **Neighborhood Attacks:** This analysis, based on past research,[38] assesses how similarly a model evaluates an original piece of content compared to a synthetically generated piece of content. The test uses a measure called "perplexity", which calculates how likely a model is to produce a particular response. To run this test, we first selected original sentences from publisher content and generated single-word replacements to randomly selected words. The words chosen as replacements are considered neighboring words and have a similar perplexity score to the original word. The neighboring words were generated using the Roberta language model, in accordance with the lexical substitution approach described by Mattern et al. We then examine the perplexity score for the original sentence and the synthetically generated sentence using GPT-3.[39] The hypothesis is that if the original content was included in the training set, this would make it more likely that the model is more perplexed by the new sentence than by the original one. We tested this on data from 2020, using 2023 as a control, but did not find a statistically significant difference. In principle, this may be due to the training cutoff challenge we described earlier. Furthermore, the original paper conducted this test at the scale of hundreds of thousands of samples, whereas we conducted it for a few thousand samples. Scaling this test may give more robust results.
- **Lowercase Perplexity[40]:** In this analysis, the hypothesis was that lowercase version of article titles would have a higher perplexity since the LLM has only seen the uppercased titles. We therefore generated lowercase versions of article titles and queried GPT-3 to return the perplexity of the original title and its lowercase version. However, the results did not show a consistent pattern of higher perplexity for the lowercase versions.
- **Word Additions:** We asked GPT-4 to insert a single word into a sentence, then prompted GPT-4 to guess which word was added. There were 0 successes among all tested sentences.

[38] https://arxiv.org/abs/2305.18462

[39] GPT-3.5 and GPT-4 do not provide access to log probabilities that are used to calculate perplexity scores. We therefore used GPT-3 for this (and the next) analysis.

[40] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10025743

News Media Alliance
4401 N. Fairfax Drive
Suite 300
Arlington, VA 22203

For more information you may contact:

Regan Smith
Senior Vice President and General Counsel
News/Media Alliance
571.366.1087
regan@newsmediaalliance.org

APPENDIX B

NEWS MEDIA
ALLIANCE

January 10, 2020

Andrei Iancu
Under Secretary of Commerce for Intellectual Property
and Director of U.S. Patent and Trademark Office
P.O. Box 1450
Alexandria, Virginia 22313-1450

Re: Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation

Dear Mr. Iancu:

The News Media Alliance (the "Alliance") respectfully submits this comment to Question No. 3 in the Request for Comments published by the Patent & Trademark Office at 84 Fed. Reg. 58141 (Oct. 30, 2019):

> To the extent an [Artificial Intelligence] algorithm or process learns its function(s) by ingesting large volumes of copyrighted material, does the existing statutory language (e.g., the fair use doctrine) and related case law adequately address the legality of making such use? Should authors be recognized for this type of use of their works? If so, how?

The members of the Alliance are deeply concerned about the unlicensed use of their news reporting for machine learning purposes by technology companies that do not share the cost of reporting the news but commercially benefit from the work product of the news media by using the news in a manner that does not qualify as fair use. While current copyright doctrine should compel the conclusion that this constitutes infringement of copyright, a variety of obstacles to enforcement of the media's IP rights has diminished the value of those rights. The modes of distribution and consumption of news content are rapidly changing in the digital age, and the failure to properly compensate the media for the use of their content has already become an existential threat to the business of journalism. Moreover, the continued unlicensed use of reporting by technology companies portends injury, not just to the news industry, but to the public interest that it serves: in a world in which everyone is a republisher, there will be no original reporting to republish. Accordingly, the Alliance believes that stronger enforcement of existing laws is needed to reset the balance between the originators of news and those who consume it for their own commercial advantage.

The News Media Alliance

The Alliance is a nonprofit organization that represents the interests of more than 2,000 news media organizations in the United States and internationally. The Alliance diligently advocates for newspapers before the federal government on issues that affect today's media organizations, including protecting newspapers' intellectual property.

News organizations play an important role in the U.S. economy and democracy. Every day, news publishers invest in high-quality journalism that keeps our communities informed, holds those in power accountable, and supports the free flow of information and ideas in society. Without free and flourishing news media, our society would be less well-off and less informed.

The newspaper industry generates over $25 billion in total revenue and employs a total of approximately 152,000 people in the United States.[1] These journalists and others who rely on newspapers for their living create content that reaches 136 million adults in the United States each week, representing 54 percent of the country's adult population.[2] Online, news organizations receive over 200 million unique visits and 6.7 billion page views per month, while 44 percent of the news media audience relies exclusively on print publications.[3] News publishers also ensure the health of our local communities, with most local news media companies reaching more adults in their local markets than any other local media.[4]

Notwithstanding their vital societal role and the public's reliance on accurate and current information, news publishers are struggling to sustain investments in high-quality journalism. Despite an increase in digital audience and subscriptions, both overall and print circulation dropped by approximately 8 and 12 percent between 2017 and 2018, respectively.[5] In total, news publisher revenues have decreased by 58 percent since 2005, and newsroom employment has dropped from over 72,000 to an estimated 37,900 in the same period.[6] While the share of digital advertising revenues has grown in recent years,[7] such revenues are often not enough to offset the reduced print advertising and subscription revenues. News publishers struggle in large

[1] Pew Research Center, Newspaper Fact Sheet (July 9, 2019), http://www.journalism.org/fact-sheet/newspapers/; United States Department of Labor, Bureau of Labor Statistics, Occupational Employment Statistics (May 2018), https://www.bls.gov/oes/current/naics5_511110.htm#00-0000.

[2] News Media Alliance, News Advertising Panorama: A wide-ranging look at the value of the news audience at 64 (2018), https://www.newsmediaalliance.org/wp-content/uploads/2018/10/FINAL_NMA_PANORAMAbook_WEB_10-19-18.pdf.

[3] *Id.*

[4] *Id.*

[5] Pew Research Center, Newspaper Fact Sheet (July 9, 2019), http://www.journalism.org/fact-sheet/newspapers/.

[6] *Id.*

[7] *Id.*

part because the online marketplace is dominated by a few online platforms, referred to in this comment as "tech platforms," that control the digital advertising ecosystem and determine the reach and audience for news content online, thereby reducing the ability of news publishers to benefit from digital advertising and to develop their relationships with their readers.

The Problem

The news media rely on robust legal protection of their intellectual property, typically in the form of copyrights, for their very existence. The remuneration made possible in the form of subscriptions and various licensing fees for distribution of their content generates the revenue necessary to finance the cost of reporting the news, such as the global network of news bureaus and journalists that major media enterprises must maintain to carry on their work. The ability of those who do not foot the bill to free ride on those efforts would quickly extinguish the practice of journalism as we know it. And the stakes are not merely commercial. The words of district judge Denise Cote in *Associated Press v. Meltwater U.S. Holdings, Inc.*, 931 F. Supp. 2d 537, 553 (S.D.N.Y. 2013), are apt:

> [T]he world is indebted to the press for triumphs which have been gained by reason and humanity over error and oppression . . . . Permitting [Meltwater] to take the fruit of [AP's] labor for its own profit, without compensating [AP], injures [AP's] ability to perform [its] essential function of democracy.

Artificial Intelligence ("AI") is increasingly involved in various ways in the practice of journalism. As relevant here, tech platforms scrape news websites and ingest copyright-protected news content. The content appropriated by these technologies, typically in massive quantities, may be used to train the AI to perform a variety of functions, which increasingly includes learning from news media reports how to write a story, and drawing on content from multiple sources to create a rendition of the news that is not identical to that of any one contributing source while being completely dependent on all of those sources in combination. For simplicity, we will refer to this broad type of use as "training AI." It is a form of "machine learning": "Instead of requiring people to manually encode hundreds of thousands of rules, this approach programs machines to extract those rules automatically from a pile of data."[8] While the deployment of this process by tech giants such as Google and Amazon may be the most remarked upon, they are not alone in engaging in the sort of machine learning of greatest concern to the Alliance and its members.[9]

[8] Karen Hao, We Analyzed 16,625 Papers to Figure Out Where AI is Headed Next, *MIT Technology Review* (Jan. 25, 2019), https://www.technologyreview.com/s/612768/we-analyzed-16625-papers-to-figure-out-where-ai-is-headed-next/.

[9] *See, e.g.*, Knowhere Launches with $1.8M in Funding to Deliver Unbiased News Coverage with Machine Learning (Apr. 4, 2018), https://s3.us-west-2.amazonaws.com/cruncher-images/static/press-release/knowhere-launch-press-release.pdf ("Knowhere's technology scours the internet, evaluating narratives, factual claims and bias in reporting, by outlets as varied as the New York Times and Breitbart, to inform three 'spins' of every controversial story: left, impartial, and right, or positive, impartial and negative. The technology can write stories in

The use of copyright-protected news content by tech platforms to train their AI represents an increasing threat to the practice of journalism. The challenges are multiple. First, current technology makes the replication and manipulation of vast amounts of content inexpensive and easy. Just as the printing press and later the photocopier exponentially increased the capacity of humans to reproduce content that initially had to be manually copied, another giant leap has occurred in the digital age and text can be replicated, processed and disseminated with a few keystrokes. Second, this use is not readily detected. Whereas copying and re-dissemination of all or a substantial portion of intact text to the public is generally detectible and provable as an infringement, programs running in the background at unlicensed tech platforms that make use of ingested news content as the raw material input for machine learning cannot so easily be policed by the content proprietor. Third, the dominance of the principal tech platforms renders enforcement of IP rights extremely difficult when the publishers are dependent on the same tech platforms to mediate between the publishers and their audience by locating and linking to reports of interest to users. Fourth, the training of AI is increasingly being used to support news products that cause the audience to remain inside the tech platform's ecosystem, rather than simply as a search tool that links users to the original information provider. This evolution can be seen in contrasting public statements by Google. In 1998, in a publication entitled "Ten Things We Know to be True," Google maintained, "We may be the only people in the world who can say our goal is to have people leave our website as quickly as possible."[10] By 2011, however, the chief executive of Google would testify to the Senate Judiciary Committee, "if we know the answer … it is better for the consumer for us to answer that question so that they don't have to click anywhere."[11] The phenomenon of tech platforms "answering the question" can be seen in such relatively recent developments as "featured snippets" in Google search responses, which obviate the need for users to click through to a source for the requested information, and voice assistant products such as Amazon's Alexa and Google Assistant, which will directly answer user queries without providing a ready path for users to consult the source of the information. These devices are powered by AI systems that have been trained on the information ingested from the websites of content originators, many of which are traditional news sources that have expended time and labor to collect and report it.

The recent implementation by Google of its machine learning tool nicknamed "BERT" shows how tech platforms can make use of expressive textual content to train their search engines. A recent article by a Google officer explains:

anywhere from 60 seconds to 15 minutes, depending on the amount of controversy among the sources. Once article drafts are complete, human journalists review the piece, which in turn trains the machine learning algorithm.").

[10] Ten things we know to be true, Google, https://www.google.com/about/philosophy.html (last visited Jan. 7, 2020).

[11] The Power of Google: Serving Consumers or Threatening Competition?, Hearing before the Subcommittee on Antitrust, Competition Policy and Consumer Rights of the Committee on the Judiciary, 112th Cong., Sept. 21, 2011, *available at* https://www.govinfo.gov/content/pkg/CHRG-112shrg71471/html/CHRG-112shrg71471.htm.

> With the latest advancements from our research team in the science of language understanding--made possible by machine learning--we're making a significant improvement to how we understand queries, representing the biggest leap forward in the past five years, and one of the biggest leaps forward in the history of Search.
>
> **Applying BERT models to Search**
>
> Last year, we introduced and open-sourced a neural network-based technique for natural language processing (NLP) pre-training called Bidirectional Encoder Representations from Transformers, or as we call it--BERT, for short. This technology enables anyone to train their own state-of-the-art question answering system.
>
> This breakthrough was the result of Google research on transformers: *models that process words in relation to all the other words in a sentence, rather than one-by-one in order. BERT models can therefore consider the full context of a word by looking at the words that come before and after it*—particularly useful for understanding the intent behind search queries.[12]

Even before the advent of sophisticated AI tools to enhance the efficiency of search engines, enterprises such as Google scraped the full text of news reports from media websites and ingested that material into the Google search database. The description of the BERT program as used by Google makes it all the more clear that this process requires copying and analyzing the full text of third party content, because the AI is exquisitely reliant on the precise grammar and word selection of the text to teach itself how to interpret queries, carry out searches and deliver responsive content. The tech platforms' AI can also learn how to write news articles by analyzing the text of news reports provided by human sources. For all these reasons, the tech platforms make use of the precise expression in the news articles they ingest and do more than just extract facts from those reports.

The Alliance believes that, in the words of Question 3, "ingesting large volumes of copyrighted material" for this purpose constitutes copyright infringement if undertaken without a license from the proprietor of the material. This is true whether or not the platform goes on to disseminate any of that material in a form substantially similar to the ingested original. Tech platforms that appropriate vast quantities of news content for this purpose should pay for the privilege of doing so, no less than they should pay for the electricity that powers their computers or motorists for the fuel that powers their cars.

[12] Pandu Nayak (Google Fellow and Vice President, Search), Understanding Searches Better than Ever Before, (Oct. 25, 2019), https://www.blog.google/products/search/search-language-understanding-bert/ (emphasis supplied)

Analysis

1. Textual and visual news content is fully protected by copyright. While facts by themselves cannot be protected by copyright, the expression of facts is so protected. *Feist Publ'ns, Inc. v. Rural Tel. Serv. Co.*, 499 U.S. 340, 348 (1991). The case law is replete with examples of infringers of news reporting in diverse media being held liable for infringement. *See, e.g.*, *Fox News Network, LLC v. TVEyes, Inc.*, 883 F.3d 169, 182 (2d Cir. 2018) ("*TV Eyes*") (video news clips); *Nihon Keizai Shimbun, Inc. v. Comline Business Data, Inc.*, 166 F.3d 65 (2d Cir. 1999) (newswire text articles); *H.C. Wainwright & Co. v. Wall St. Transcript Corp.*, 418 F. Supp. 620 (S.D.N.Y. 1976), *aff'd*, 558 F.2d 91 (2d Cir. 1977), *cert. denied*, 434 U.S. 1014 (1978) (reports concerning stocks and bonds); *Agence France Presse v. Morel*, 934 F. Supp. 2d 584, 592 (S.D.N.Y. 2013) (news photographs).

2. The ingestion of substantial volumes of news content is, at minimum, a prima facie infringement of the reproduction right. The Copyright Act enumerates the exclusive rights of copyright in section 106. First among those rights is the exclusive right "to reproduce the copyrighted work in copies . . . ." 17 U.S.C. § 106(1). Because the reproduction must be verbatim to lend itself to training the AI, the question whether the copy is sufficiently similar to the original does not arise. When a tech platform ingests a volume of news reporting and fixes it in a database in the memory of its computer system, it has made an infringing copy. *E.g.*, *Stern Elecs., Inc. v. Kaufman*, 669 F.2d 852, 855 (2d Cir. 1982) ("[T]he memory devices of the game satisfy the statutory requirement of a 'copy' in which the work is 'fixed.'").[13] If, as is common practice, the dataset is then manipulated in the course of carrying out machine learning, infringement of the exclusive right to create derivative works likely also occurs. 17 U.S.C. § 106(2).

Those who would immunize these activities often describe the ingestion of content by tech platforms as "non-expressive" use, because the AI learning assertedly depends on use of the content as "data" rather than as a communicative work. The Copyright Act, however, does not make this distinction. A reproduction of a work in copies—which are "material objects . . . in which a work is fixed by any method now known or later developed, and from which the work can be perceived, reproduced, or otherwise communicated, either directly or with the aid of a machine or device[,]" 17 U.S.C. § 101—without more, is a prima facie exercise of the section 106(1) right irrespective of the intended use. Similarly, the fact that the ingestion of a work for machine learning can be described as a "non-display use" because it does not involve further dissemination of the work is itself not a defense. The reproduction right exists separately from the further rights to "distribute copies . . . of the copyrighted work to the public . . . [,]" *id.* §106(3), and "to display the copyrighted work publicly[,]" *id.* § 106(5). The making of the copy is prima facie infringement *whether or not* it is then distributed or displayed, for example, in

[13] We are not here concerned with "intermediate" copying of a computer program, *see* 17 U.S.C. § 117(a)(1), or with copies whose existence is so brief as to not constitute a "fixation." *See Cartoon Network LP v. CSC Holdings, Inc.*, 536 F.3d 121, 129-30 (2d Cir. 2008).

response to a search query.[14]

3. It would be a mistake to assume that the ingestion of substantial volumes of news content for machine learning is fair use. While understood to be a part of copyright law for centuries, the fair use doctrine was not codified until the 1976 Act in section 107. The preamble of that section mentions "news reporting" as an example of the type of use that could attract a fair use defense, but the Supreme Court has made clear that these examples are "illustrative and not limitative" and "provide only general guidance about the sorts of copying that courts and Congress most commonly ha[ve] found to be fair uses." *Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569, 577-78 (1994). Thus, as the Second Circuit observed in the *Google Books* decision, "[t]hose who report the news undoubtedly create factual works. It cannot seriously be argued that, for that reason, others may freely copy and re-disseminate news reports." *Authors Guild v. Google, Inc.*, 804 F.3d 202, 220 (2d Cir. 2015) ("*Google Books*"). Of course, inquiries into fair use are necessarily fact-specific and do not lend themselves to bright-light generalizations. Nevertheless, any consideration of whether ingestion of news content for machine learning purposes is fair use would proceed to consideration of the four factors set out in section 107, which themselves are not exhaustive, but merely indicative of whether a use should be deemed fair.

(1) **The purpose and character of the use**. As *Campbell* and numerous other authorities agree, the "transformative" use of a copyrighted work is a distinctive feature of a fair use. Secondary works that are productive in that they put the original to a new use, such as quotation for the purpose of parody or criticism, in and of themselves provide something new and different to the public and are less likely to merely supersede the purposes of the original. They therefore have a greater call on protection from infringement. As Judge Leval explained in *Google Books*, "transformative uses tend to favor a fair use finding because a transformative use is one that communicates something new and different from the original or expands its utility, thus serving copyright's overall objective of contributing to public knowledge." 804 F.3d at 214. Yet, as Judge Leval—the originator of the "transformative use" concept[15]—also cautioned in the same decision, the term "transformative" must be applied with discretion. It would be overly simplistic to suggest that any and all repurposing or recasting of a copyrighted work is transformative in a sense meaningful to fair use analysis. Among other things, "transformation" of the work into a derivative work is a right expressly reserved to the copyright proprietor, 17 U.S.C. § 106(2).

The ingestion of volumes of news content to obtain material for machine learning is a pure act of consumption, not of transformation. Whereas *Google Books* and like decisions turn on the value of "communicating" to the audience something new and socially desirable, in the case of machine learning the relevant value of the news content is extracted within the computer system of the tech platform before any new work is created and is not defensible unless, arguably, all possible resulting uses by the ingesting party are fair use. For example, the linchpin of the fair use determination in *Google Books* was the provision of information "about" the

[14] Whether or not such a use qualifies as fair use is a separate issue, discussed below.

[15] Pierre Leval, Toward a Fair use Standard, 103 Harv. L. Rev. 1105 (1990).

original books to users interested in knowing, for example, whether particular words were used in the original in order to help users find what they were looking for in a book. No such socially redeeming value to the public can be discerned in the ingestion of content for broad applications of machine learning. Here, the originals are used as the fuel to run the tech platform's engine.

The possibility that machine learning may in turn be applied to *some* purpose that would not infringe copyright does not excuse all antecedent ingestion. As commentator Benjamin Sobel has cogently argued:

> Making gigabytes upon gigabytes of copies of copyrighted art, in order to teach a machine to mimic that art, is indeed a remarkable technological achievement. An artificially intelligent painter or writer may yield social benefits and enrich the lives of many beholders and users. However, this view of productivity is overbroad. No human can rebut an infringement claim merely by showing that he has learned by consuming the works he copied, even if he puts this new knowledge to productive use later on.[16]

That observation is surely correct. When tech platforms ingest published news content and set their AI programs upon that text in order to, in the words of Google, "process words in relation to all the other words in a sentence," (*supra*, note 12), they are appropriating the expressive content of the original work and do not enjoy blanket immunity merely because some downstream activities facilitated by that appropriation may be deemed productive or socially desirable. For example, the fact that some recipients of unlicensed copies of broadcast news clips may have wished to use them for research or other salutary purposes does not render the pervasive copying and distribution of such clips a fair use. *See TVEyes*, 883 F.3d at 178 n.4 ("That a secondary use can facilitate research does not itself support a finding that the secondary use is transformative.") (citing *American Geophysical Union v. Te*xaco, 60 F.3d 913 (2d Cir. 1994)).

(2) **The nature of the copyrighted work**. This factor is generally recognized to be the least significant in the fair use calculus. *E.g.*, *Google Books*, 804 F.3d at 220 (citing William F. Patry, Patry On Fair Use § 4.1 (2015)). To be sure, there is dictum in *Harper & Row* that "[t]he law generally recognizes a greater need to disseminate factual works than works of fiction or fantasy." *Harper & Row, Publrs., Inc. v. Nation Enters*., 471 U.S. 539, 563 (1985). *Google Books* questions how far that dictum should impact fair use analysis. 804 F.3d at 220. But to the extent it suggests a possibly greater scope of fair use in connection with factual reports, it would only be because of the public interest in "dissemination" in order to afford public access to the facts reported—a need not satisfied when tech platforms ingest the news for their own commercial purposes.

[16] B. L. W. Sobel, Artificial Intelligence's Fair Use Crisis, 41 Colum. J. L. & Arts 45, 73 (2017) (citing *Sony Corp. of Am. v. Universal City Studios*, *Inc*., 464 U.S. 417, 455 n.40 (1984), which suggests that a "constituent who copies a news program to help make a decision on how to vote" would not be protected by the fair use doctrine despite the salutary purpose).

(3) **The amount and substantiality of the portion used in relation to the copyrighted work as a whole**. This factor supports the view that ingestion of substantial, indeed vast, volumes of text without the permission of the copyright owner for the purpose of machine learning is not a fair use. Although a compelling fair use purpose on rare occasions can justify taking the entirety of a work when taking less will not suffice, *Swatch Grp. Mgmt. Servs. v. Bloomberg L.P.*, 756 F.3d 73, 90 (2d Cir. 2014), this is not such a case. *Swatch*, which involved an unlicensed dissemination of a corporate earnings call, turned on the public interest in having access to the contents of the call and the risk that paraphrasing or excerpting would not accurately render the nuance of what was discussed. Again, this provides no justification for a use that does not enhance public knowledge but represents pure consumption by the tech platform.[17]

(4) **The effect of the use upon the potential market for or value of the copyrighted work**. This is often stated to be the most important fair use factor. *E.g.*, *Harper & Row Publ. v. Nation Enters.*, 471 U.S. 539, 566 (1985). Members of the Alliance are particularly concerned about the predictable, and already occurring, commercial harm inflicted on them by the increasing use of their intellectual property without their permission to train the AI employed by tech platforms. While the diversion of audience that occurs when a person in horizontal competition with the content proprietor appropriates an original work and markets it in competition with the originator is an obvious example of Factor 4 harm, that is not the only cognizable type of harm. Factor 4 has a vertical aspect as well: depriving content creators of natural markets wherein they can sell or license their works for such consumption is also—literally as well as a matter of commercial reality—a pernicious "effect on the potential market for or value of" the copyrighted material. As the Second Circuit noted in *Castle Rock Entm't, Inc. v. Carol Pub. Group, Inc.*, 150 F.3d 132, 145 (2d Cir. 1998), "[t]he fourth factor must also 'take account ... of harm to the market for derivative works,' defined as those markets 'that creators of original works would in general develop or license others to develop[.]'" (citation omitted). "It is indisputable that, as a general matter, a copyright holder is entitled to demand a royalty for licensing others to use its copyrighted work, and that the impact on potential licensing revenues is a proper subject for consideration in assessing the fourth factor[.]" *Texaco*, 60 F.3d at 929 (citations omitted). In doing so, the courts look to the use's impact on "traditional, reasonable, or likely to be developed markets." *Id.* at 930. Thus, in *Texaco*, the bulk photocopying by a commercial enterprise's research arm of scientific articles published by plaintiff was deemed not a fair use where the licensing of such articles was a natural, and to some extent already exploited, market for such scientific articles through the development of clearinghouses established to license such photocopying.

When a consumer of copyrighted material exploits that material without permission,

[17] We do not view the advent of "federated learning", *see generally* B. McMahan & D. Ramage, Federated Learning: Collaborative Machine Learning without Centralized Training Data, Google Research Blog (Apr. 6, 2017), https://perma.cc/XVA2-J96J, to change the analysis or result. When a central authority such as Google delegates portions of a large database to individual users to analyze and return results in order to replicate the more conventional model of ingesting the entire database at one site, the delegator should incur liability for inducing infringement, and the individual user delegees for direct infringement of the portions copied to their devices.

Factor 4 is triggered even where the use is for a purpose collateral to the main or original purpose of creating the material. That is one of the important lessons of *TVEyes*, where the defendant ingested vast amounts of broadcast television news programming and enabled its subscribers to watch, download and save actual news clips of up to ten minutes duration without license from the source broadcasters. *Id*. at 175. The court found that "Fox itself might wish to exploit the market for such a service . . . . [and that] TVEyes deprives Fox of revenues to which Fox is entitled as the copyright holder." *Id.* at 180.[18]

Here, as in *Texaco* and *TVEyes*, the licensing of the copyrighted content for the use made of it by the tech platforms is, if not "traditional," certainly a "reasonable, or likely to be developed market[]." *Texaco*, 60 F.3d at 930. Media entities for some time have identified this market as one in which their proprietary content has particular value and have curated and made available annotated *corpora* of their published news reporting for the specific purpose of training AI. This has been done by, for example, the copyright holders of *The Wall Street Journal*,[19] *The New York Times*,[20] and the Reuters News Service[21]. The Linguistic Data Consortium catalogue

[18] Another lesson of *TVEyes* is that limiting the unlicensed use to "internal purposes only" confers no talismanic immunity from infringement. The defendant TVEyes purported to contractually require its subscribers to make only such "internal" use of downloaded clips, *id*. at 175, but that did not privilege its unlicensed distribution of copyrighted video clips. So too here, when tech platforms ingest news content to train their AI, they are making an "internal" use of that content, but that should not protect them from infringement.

[19] Linguistic Data Consortium – BLLIP 1987-89 WSJ Corpus Release 1, https://catalog.ldc.upenn.edu/LDC2000T43 (last visited Jan. 6, 2020).

[20] Linguistic Data Consortium – The New York Times Annotated Corpus, https://catalog.ldc.upenn.edu/LDC2008T19 (last visited Jan. 6, 2020). For a description of the corpus, see this blog post announcing and explaining the corpus: Jacob Harris, Fatten Up Your Corpus, NYT Open (Jan. 12, 2009), https://open.blogs.nytimes.com/2009/01/12/fatten-up-your-corpus/ ("Available for noncommercial research license from The Linguistic Data Consortium (LDC), the corpus spans 20 years of newspapers between 1987 and 2007 (that's 7,475 issues, to be exact). This collection includes the text of 1.8 million articles written at The Times (for wire service articles, you'll have to look elsewhere). Of these, more than 1.5 million have been manually annotated by The New York Times Index with distinct tags for people, places, topics and organizations drawn from a controlled vocabulary. A further 650,000 articles also include summaries written by indexers from the New York Times Index. The corpus is provided as a collection of XML documents in the News Industry Text Format and includes open source Java tools for parsing documents into memory resident objects.").

Google has acknowledged accessing this corpus for machine learning, specifically, developing "entity salience." *See* Dan Gillick, *et al*., Teaching machines to read between the lines (and a new corpus with entity salience annotations), Google AI Blog (Aug. 25, 2014), https://ai.googleblog.com/2014/08/teaching-machines-to-read-between-lines.html.

[21] *See, e.g*., David D. Lewis, Reuters-21578 Text Categorization Test Collection Distribution 1.0 README file (Sep. 26, 1997), https://perma.cc/V7JJ-CNVW. This corpus consists of the contents of the Reuters newswire for 1987.

lists hundreds of such *corpora* available for license.[22] Today, major news media organizations continue to commercially license these rights and have formulated and offer licensable data products for this purpose. The use in question thus easily satisfies the requirement of *Texaco* that it be in a market in which users should reasonably expect to seek permission of the copyright proprietors and to compensate them for the use of their works—even though some major tech platforms do not do so.

This use is also dissimilar from those held to be fair in such decisions as *Google Books*, 804 F.3d 202; *Authors Guild, Inc. v. HathiTrust*, 755 F.3d 87, 95 (2d Cir. 2014); *Kelly v. Arriba Soft Corp*., 336 F.3d 811 (9th Cir. 2003); *Perfect 10, Inc. v. Amazon.com, Inc*., 508 F.3d 1146 (9th Cir. 2007). In those decisions, the defendant's copying of copyrighted material was in the service of creating a searchable index that enabled users to locate desired content and link to it. In theory at least, this was beneficial to content originators and drove traffic to their websites by displaying a snippet that by itself did not substitute for the original full-text work. *See, e.g.*, *Authors Guild, Inc. v. Google, Inc.*, 954 F. Supp. 2d 282, 291 (Google Books "uses snippets of text to act as pointers directing users to a broad selection of books"), *aff'd*, 804 F.3d 202; *Perfect 10, Inc.*, 508 F.3d at 1165 ("a search engine transforms the image into a pointer directing a user to a source of information"). This feature is absent from the case of tech platforms consuming news content to train their AI for broader purposes. The tech platform *may* use the AI to make more efficient the searches that *may* send inquirers to the original website, but they may also be used—and increasingly are used—to create alternative renditions of news and information that disintermediate between the original source and its audience. This results in market harm that far outweighs any consumer good generated by this process. Thus, the recognition in current case law that internal reproduction and indexing of content for some machine-driven purposes may be fair use should not be extended to a blanket immunity for all ingestion of copyright-protected content by tech platforms for any and all commercial purposes.

* * * * *

Today, having access to trusted content of consistently high quality is integral to power machine learning. By compensating news media organizations for their intellectual labor in generating that content, an appropriate balance can be struck between advancing AI-based technology, while preserving the media's "ability to perform [its] essential function of democracy." *Meltwater*, 931 F. Supp. 2d at 553. The Alliance believes that copyright law, properly understood and consistently enforced, should lead to a system where content originators are compensated for their work. Various business and licensing models, with proper legal support, may be employed to achieve this end. Pressing policy concerns—including sustainability of the news media industry and, by extension, benefit to the public that relies on it for accurate and current content—demand this outcome.

[22] Linguistic Data Consortium, LDC Catalog, https://catalog.ldc.upenn.edu/ (last visited Jan. 6, 2020).

To the extent the current legal framework cannot support such a regime, legislative solutions may prove useful or even necessary. But the issues addressed herein require more analysis, dialogue among all stakeholders, and careful attention to detail. The Alliance calls for and stands ready to contribute to continued study and deliberation on this important issue that will help move American copyright law fully into the 21st century.

Respectfully submitted,

David Chavern
President & Chief Executive Officer
NEWS MEDIA ALLIANCE
4401 North Fairfax Drive, Suite 300
Arlington, Virginia 22203

Danielle Coffey
Senior Vice President & General Counsel
NEWS MEDIA ALLIANCE
4401 North Fairfax Drive, Suite 300
Arlington, Virginia 22203

Robert P. LoBue
Julie Simeone
PATTERSON BELKNAP WEBB &
TYLER LLP
1133 Avenue of the Americas
New York, New York 10036
Counsel for the News Media Alliance

APPENDIX C

EMMA EUROPEAN MAGAZINE MEDIA ASSOCIATION

ENPA EUROPEAN NEWSPAPER PUBLISHERS' ASSOCIATION

EMMA-ENPA's Core Concerns on AI and Copyright

26.07.2023

AI technologies hold great promise but significantly impact how publishers' content is created, distributed and consumed. They also have key implications for the IP rights of content creators, including press publishers, raising concerns about the use of copyrighted material without proper authorisation or licensing. The lack of meaningful safeguards under existing EU law exposes publishers' content to abusive and often illegal uses by providers of text-generative AI such as OpenAI or "AI press" offerings and deprives them of effective means to act. This state of affairs is increasingly compromising rightsholders' ability, in fact and in law, of exercising effective general control over their own content. The following proposed measures are necessary first steps to address these mounting challenges:

- **Safeguard publishers' rights of ownership of and control over their content:** EU law must clarify and secure publishers' exclusive right to determine whether and how their copyright-protected content may be used by AI systems. The principle that only lawfully accessible content may be used must be upheld and the application of exclusive rights under EU copyright law to uses by AI systems confirmed. As such uses do not necessarily require longer-term reproductions, the exceptions/limitations under Art. 5(1) of Directive 2001/29/EC should not apply. The legal ambiguities related to the exercise of the TDM reservation under Art. 4 of Directive (EU) 2019/790 must also be dispelled by clarifying the binding character of reservations in the website's T&C. TDM opt-outs must not be circumvented nor result in any negative consequences in terms of access to or display of the content in question, e.g. in search engines. To that effect, web crawlers should be required to provide effective, machine-readable technical solutions to effectively carry out the reservation. When IP content is crawled for one purpose (e.g., indexing for search), express authorisation from rightsholders for any other purpose must be sought. Last, in the absence of consent by the rightholder, up-to-date content must be protected from uses by AI systems for a limited period after initial publication.

- **Address the issue of extraterritoriality:** As most relevant AI systems are located outside the EU, it is crucial to clarify that the exploitation by AI systems of works published in the EU is deemed to have taken place in the EU. Otherwise, the risk of creating a legal vacuum may result in a complete lack of copyright protection.

- **Build upon the transparency requirements proposed by the EU-Parliament:** The transparency requirements proposed in the Parliament's negotiating mandate on the AI Act (new Art. 28b(4)(c)) represent a first yet clearly insufficient step. A comprehensive, meaningful obligation to disclose the complete reference to the copyright-protected content sources (and rightsholders concerned) for training and output purposes is necessary.

- **Introduce a presumption of use of copyrighted content:** Because AI systems are deliberately opaque ("Blackbox"), rightsholders can often not prove the nature and extent of the use of their content. There must therefore be a legal presumption of the use of copyrighted content whereby, in cases where the use of such content appears to be plausible, the burden of proof will lay with the AI system provider.

- **Ensure fair remuneration for rightsholders:** Press publishers and rightsholders more generally must be empowered to exercise effective general control over their own content and secure fair remuneration for the exploitation of their works. Binding and enforceable obligations to ensure fair remuneration for content creators on a more robust basis in the age of AI are essential. The right to remuneration should not be waivable. A revision of the current 2-year limit of application of the DSM Directive's Art. 15 neighbouring right should also be considered in light of the long timeframes of use of content by AI systems.

- **Not overlook the Digital Markets Act (DMA):** Gatekeepers may be tempted to favour their own "AI press" and AI services in their other core platform services (i.e. search) or impose unfair sharing of data/content to their business users in exchange for access. Article 6(5) and 6(12) (DMA), in particular, could already address some of those shortcomings and, should this prove insufficient, a revision of the DMA should be considered.

- **Introduce a labelling obligation for artificially created media content:** AI systems that generate complex content should be obliged to disclose that the content has been artificially generated and created to ensure transparency towards the end user. This obligation should however not apply in cases where a person reviews and edits the content while being legally liable and editorially responsible for it.