JAMIE ODELL*

Training on Headlines: *The New York Times*, OpenAI, and the Copyright Implications of AI Data Usage

* J.D. Candidate, 2026, University of Oregon School of Law. The author thanks the faculty of the University of Oregon School of Law and the staff of Oregon Law Review for their invaluable feedback and guidance during this process. The author also expresses gratitude to her friends and family; this piece would not be possible without their endless support. The author dedicates this Comment to the creatives of the world—your labor is irreplaceable. *Please note:* At the time this author wrote this Comment, the legal doctrines surrounding copyright law and AI data usage remain unsettled law.

INTRODUCTION

In late November of 2023, OpenAI, Inc. ("OpenAI") released its chatbot, "ChatGPT," which quickly went viral as users discovered its ability to perform almost any writing task, including writing stories, coding computer programs, generating brand slogans, and planning for vacations.[1] Within five days, ChatGPT had over one million users.[2] As of April 2025, OpenAI had upgraded its Generated Pre-Trained Transformer ("GPT") model more than four times,[3] and has come under judicial scrutiny for using copyrighted materials to train its generative artificial intelligence models ("GenAI") without the copyright owner's license or consent.[4]

Many plaintiffs, including the Authors Guild—represented by notable authors such as George R.R. Martin and John Grisham—Sarah Silverman, *The New York Times*, and others have initiated legal actions challenging artificial intelligence ("AI") companies' use of their copyrighted works to train GenAI algorithms without authorization.[5] Particularly, the *Times* argues that OpenAI has incorporated millions of its articles into GPT training datasets without compensation, credit, or consent.[6] Furthermore, the *Times* asserts that OpenAI's models generate outputs that closely resemble its copyrighted content, thereby functioning as a direct competitor.[7] Among the remedies sought, the *Times* requests that all its copyrighted materials contained in GPT

---

1 Bernard Marr, *A Short History of ChatGPT: How We Got to Where We Are Today*, FORBES (May 19, 2023, at 01:14 ET), https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/ [https://perma.cc/2HSN-YP9A].

2 *Id.*

3 *ChatGPT – Release Notes*, *Sunsetting GPT-4 in ChatGPT*, OPENAI (Apr. 10, 2025), https://help.openai.com/en/articles/6825453-chatgpt-release-notes (on file with the Oregon Law Review) ("Effective April 30, 2025, GPT-4 will be retired from ChatGPT and fully replaced by GPT-4o.").

4 *See* First Consolidated Complaint at 2–3, Authors Guild v. OpenAI, Inc., 345 F.R.D. 585 (S.D.N.Y. 2024) (No. 23-CV-08292); *see also* Complaint at 1, Silverman v. OpenAI, Inc., No. 23-CV-03416 (N.D. Cal. Jul. 7, 2023); Complaint at 2–3, New York Times Co. v. Microsoft Corp., 777 F. Supp. 3d. 283 (S.D.N.Y. 2025) (No. 23-CV-11195) [hereinafter New York Times Complaint].

5 First Consolidated Complaint, *Authors Guild*, *supra* note 4; Complaint, *Silverman*, *supra* note 4; New York Times Complaint, *supra* note 4.

6 New York Times Complaint, *supra* note 4, at 59.

7 *Id.*

training sets be destroyed.[8] In response, OpenAI argues its use of copyrighted materials is fair use because the works are used to train large language models ("LLMs") and are not intended to act as a substitute for the original works.[9]

As artificial intelligence continues to advance, the ongoing legal battle between *The New York Times* and OpenAI highlights the diverging, complex tensions at the heart of copyright law—balancing innovation against the constitutional promise to protect original expression.[10] The Constitution's Copyright Clause "promote[s] the progress of Sciences and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries."[11] The Copyright Clause thus requires a balance between the societal value of protecting tangible expressions of creativity against the need for flexibility in technological innovations.[12]

This Comment discusses the copyright law concerns associated with the use of copyrighted materials within the training datasets for AI. Part I will discuss OpenAI's shift in corporate identity before defining and explaining the functionality of OpenAI's GPT models. Part I will also examine the types of copyrighted works involved in AI training and the methods by which these copyrighted materials are gathered and curated. Part II will analyze the legal and factual claims presented in the *Times*'s complaint against defendants OpenAI, Inc. and Microsoft Corporation, emphasizing the principal allegation of direct copyright infringement. Part III will explain OpenAI's response to the *Times*'s complaint. Part IV will assess the *Times*'s allegation of direct copyright infringement and assess OpenAI's fair use defense, examining (1) the purpose and character of the use; (2) the nature of the original work; (3) the quantitative and qualitative scope of the works used; and (4) the impact of the use on the copyrightable material's market value. Finally,

---

8 *Id.* at 68.

9 Defendant's Memorandum in Support of Motion to Dismiss at 2–3, New York Times Co. v. Microsoft Corp., 777 F. Supp. 3d. 283 (S.D.N.Y. 2025) (No. 23-CV-11195) ("There is a genuinely important issue at the heart of this lawsuit—critical not just to OpenAI, but also to countless start-ups and other companies innovating in this space—that is being litigated both here and in over a dozen other cases around the country (including in this Court): whether it is fair use under copyright law to use publicly accessible content to train generative AI models to learn about language, grammar, and syntax, and to understand the facts that constitute humans' collective knowledge.").

10 U.S. CONST. art. I, § 8, cl. 8.

11 *Id.*

12 *See id.*

Part V will reflect on the enduring significance of this case, ultimately advocating for a verdict favoring the *Times* as the outcome most consistent with the constitutional aims of copyright law.

I
THE INSTITUTIONAL AND TECHNICAL FOUNDATIONS OF OPENAI'S CHATGPT

A. The Development of OpenAI's Corporate Identity

Understanding the copyright implications of training data used in LLMs like OpenAI's GPT models requires a technical and contextual background. Across the past decade, OpenAI has transitioned from being an open-source nonprofit corporation committed to transparency and broad social benefits[13] to a closed-source for-profit corporation focused on rapid technological advancement.[14] The legal questions surrounding generative AI in the context of copyright law—what data was used, who owns it, how was the data collected—perhaps ties directly into OpenAI's updated priorities.[15]

OpenAI was founded in 2015 by a variety of researchers and scientists.[16] Initially, OpenAI was a nonprofit organization, free from any financial obligation to turn a profit, with a goal to advance AI technology to "benefit humanity as a whole."[17] However, by 2019, OpenAI became a for-profit organization when it added a profit-

---

13 *Introducing OpenAI*, OPENAI (Dec. 11, 2015), https://openai.com/index/introducing-openai/ (on file with the Oregon Law Review) (OpenAI's original mission statement: "Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact.").

14 Dan Milmo, *Why Is OpenAI Planning to Become a For-Profit Business and Does It Matter?*, THE GUARDIAN (Sep. 26, 2024, at 13:27 ET), https://www.theguardian.com/technology/2024/sep/26/why-is-openai-planning-to-become-a-for-profit-business-and-does-it-matter [https://perma.cc/5M9T-3R9P].

15 *See* Jack Hardinges, Elena Simperl & Nigel Shadbolt, *We Must Fix the Lack of Transparency Around the Data Used to Train Foundation Models*, HARV. DATA SCI. REV. (May 31, 2024), https://hdsr.mitpress.mit.edu/pub/xau9dza3/release/2 [https://perma.cc/CK3Q-QQ2Y] ("Information about training data is also vital to lawmakers' attempts to assess whether foundation models have ingested personal data or copyrighted material."); Tim O'Reilly, *You Can't Regulate What You Don't Understand*, O'REILLY (June 15, 2023), https://www.oreilly.com/radar/you-cant-regulate-what-you-dont-understand/ [https://perma.cc/W34T-T4PX]; Chloe Xiang, *OpenAI's GPT-4 Is Closed Source and Shrouded in Secrecy*, VICE (Mar. 16, 2023, at 09:21 PT), https://www.vice.com/en/article/openais-gpt-4-is-closed-source-and-shrouded-in-secrecy/ [https://perma.cc/H3X8-L2VR].

16 Karl Montevirgen, *OpenAI*, BRITANNICA MONEY (Aug. 26, 2025), https://www.britannica.com/money/OpenAI [https://perma.cc/4RH3-LPAT].

17 *Introducing OpenAI*, *supra* note 13.

making subsidiary.[18] As of January 2025, OpenAI's funding round was projected to raise $40 billion from investors, potentially resulting in a total business valuation of up to $340 billion.[19] This significant increase in investment highlights the company's financial incentives to expand its for-profit activities notwithstanding its nonprofit origins. This corporate refinancing comes nearly a year after the departure of several top executives and the dissolution of OpenAI's Superalignment team—a team previously dedicated to mitigating long-term AI-related risks; namely, "security, monitoring, preparedness, safety and societal impact."[20] These developments suggest a prioritization of rapid AI advancement over cautious governance, raising questions about the implications of OpenAI's evolving corporate identity on the broader landscape of AI safety and ethics.

Moreover, in the beginning stages, OpenAI adopted an open-source strategy, as opposed to a closed-source strategy. The open-source strategy reflected its initial mission statement, which promised to share its research, code, and potential patents with the world.[21] In the context of AI development, an open-source AI system is one that provides information, like training data, source code, and weights of parameters, "to the extent that 'a skilled person can recreate a substantially equivalent system using the same or similar data.'"[22] Conversely, a closed-source AI system is one that is closely and strictly controlled by

---

18 Milmo, *supra* note 14.

19 Hayden Field & Kate Rooney, *OpenAI in Talks to Raise Funding That Would Value AI Startup at up to $340 Billion*, CNBC (Jan. 30, 2025, at 16:15 ET), https://www.cnbc.com/2025/01/30/openai-in-talks-to-raise-up-to-40-billion-at-340-billion-valuation.html [https://perma.cc/VC2C-LMYC].

20 Hayden Field, *OpenAI Dissolves Team Focused on Long-Term AI Risks, Less Than One Year After Announcing It*, CNBC (May 18, 2024, at 13:49 ET), https://www.cnbc.com/2024/05/17/openai-superalignment-sutskever-leike.html [https://perma.cc/DP8M-BFNN] (Jan Leike, a cofounder of OpenAI, stated he has "been disagreeing with OpenAI leadership about the company's core priorities for quite some time, until [they] finally reached a breaking point" and that "[b]uilding smarter-than-human machines is an inherently dangerous endeavor . . . OpenAI is shouldering an enormous responsibility on behalf of all humanity. But over the past years, safety culture and processes have taken a backseat to shiny products.").

21 *Introducing OpenAI*, *supra* note 13 ("Researchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code, and our patents (if any) will be shared with the world. We'll freely collaborate with others across many institutions and expect to work with companies to research and deploy new technologies.").

22 Rhiannon Williams & James O'Donnell, *We Finally Have a Definition for Open-Source AI*, MIT TECH. REV. (Aug. 22, 2024), https://www.technologyreview.com/2024/08/22/1097224/we-finally-have-a-definition-for-open-source-ai/ [https://perma.cc/D3GD-MMYW].

its creators, meaning that the model's internal structure and training data is not publicized.[23] Despite starting as an open-source AI model, as OpenAI began to shift its corporate structure and its governance priorities, it has shifted toward a closed-source approach, starting with the development of GPT-3.[24] Given the lack of transparency found in the latest closed-source GPT models, it is impossible to know the *exact* content used to train GPT-3 and GPT-4.

B. The Technology Behind OpenAI's Generative AI ChatGPT

Moreover, an understanding of the core claims underlying the *Times*'s lawsuit against OpenAI requires a foundational understanding of what large language models are, how they function, and the source of the datasets used to train them. LLMs are artificial intelligence systems designed to generate humanlike text by recognizing and predicting patterns in language. Understanding this predictive structure is necessary to assess how and why certain data might be used during training.

GPT models are designed to produce natural language text—ranging from sentences and paragraphs to entire documents—in a manner that is coherent and consistent with human language patterns.[25] LLMs are "any large-scale language model designed for natural language processing tasks"; hence, GPT models are merely a specific subset of LLMs.[26] There is a two-stage training process for GPT models like those created and used by OpenAI: pretraining on extensive datasets, followed by fine-tuning the model for specific tasks.[27]

---

23 Leo Mao, *AI Open and Closed Source: It Is Human Pioneer's Choice*, GWBMA (Apr. 11, 2024), https://www.registrationchina.com/articles/ai-open-and-closed-source-it-is-human-pioneers-choice/ [https://perma.cc/Z4GX-GW6C].

24 Yevgeni Chuvyrov, *Generating Infrastructure-as-Code from Natural Language: Evaluating Fine-Tuned GPT-3 Models for Cloud Infrastructure Provisioning*, STAN. CS224N CUSTOM PROJECT, at 3–6, https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1224/reports/custom_116662212.pdf [https://perma.cc/D4VL-L58N] (last visited Aug. 30, 2025).

25 Partha Pratim Ray, *ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope*, INTERNET OF THINGS AND CYBER-PHYSICAL SYS., at 121–22 (2023), https://www.sciencedirect.com/science/article/pii/S266734522300024X (on file with the Oregon Law Review); *Understanding the Difference Between GPT and LLM: A Comprehensive Comparison*, STACKADEMIC (Oct. 20, 2023) [hereinafter *Difference Between GPT and LLM*], https://stackademic.com/blog/understanding-the-difference-between-gpt-and-llm-a-comprehensive-comparison-1f624c713507 [https://perma.cc/RC8P-A7Q4].

26 *Difference Between GPT and LLM*, *supra* note 25.

27 *Data Labeling: Fine-Tuning vs. Pre-Training: Key Differences for Language Models*, SAPIEN (Oct. 9, 2024), https://www.sapien.io/blog/fine-tuning-vs-pre-training-key

During the pretraining phase, GPT models learn to predict the next word in a sentence based on its preceding words or the prompt given by the user.[28] This process helps the model learn basic language patterns, including syntax, grammar, and semantics.[29] By generalizing these patterns from vast amounts of text, the model develops a relative understanding of human language.[30] In other words, because OpenAI's LLMs are trained on an extensive collection of text, the models can readily recognize words and sentences as being organized in sequences with specific dependencies.[31] This repetitive recognition enables the model to "understand how to cut text into statistical chunks that have some predictability," memorize the patterns within these text blocks, and use this knowledge to predict what the user wants to come next—similar to the autocomplete text feature in email applications.[32] LLMs make predictions by using adjustable parameters or weights, which are "complex mathematical transformations that LLMs learn from reading those billions of words," and work to inform the AI "how likely different words or parts of words are to appear together or in a certain order."[33] Once pretraining is complete, the GPT models undergo a fine-tuning process that subjects GPT models "to a more targeted dataset that closely aligns with the specific task at hand," resulting in a more precise output given the user's prompt, purpose, industry, or preference.[34]

-differences-for-language-models [https://perma.cc/NG39-G57F] ("For example, an LLM can be pre-trained on a massive dataset like Wikipedia to grasp general language patterns and then fine-tuned with customer service scripts to create a chatbot capable of handling customer inquiries with nuanced understanding.").

28 ETHAN MOLLICK, CO-INTELLIGENCE: LIVING AND WORKING WITH AI 2–5 (2024).

29 *See id.* at 5.

30 *Id.* at 5.

31 Adam Zewe, *Explained: Generative AI*, MIT NEWS ON CAMPUS & AROUND THE WORLD (Nov. 9, 2023), https://news.mit.edu/2023/explained-generative-ai-1109 [https://perma.cc/7YRG-22TQ].

32 *Id.*

33 MOLLICK, *supra* note 28, at 6.

34 Pablo Junco, *The Power of Fine-Tuning in Generative AI*, FORBES (Oct. 10, 2023, at 09:45 ET), https://www.forbes.com/councils/forbestechcouncil/2023/10/10/the-power-of-fine-tuning-in-generative-ai/ [https://perma.cc/VDJ9-NSUR]; *see also* MOLLICK, *supra* note 28, at 8 ("AI companies hire workers . . . to read AI answers and judge them on various characteristics. In some cases, that might be rating results for accuracy, in others it might be to screen out violent or pornographic answers.").

In 2018, OpenAI introduced its first transformer-based language model, GPT-1.[35] The dataset used to train GPT-1 was relatively small compared to the later iterations of GPT models, consisting of only 117 million parameters.[36] GPT-1 was trained using the BooksCorpus dataset, a compilation of 7,000 unpublished and self-published books sourced from Smashwords, a platform for such works.[37] This dataset provided the language model with exposure to novel data.[38]

OpenAI released its next iteration, GPT-2, in 2019.[39] GPT-2 used a much larger dataset, one consisting of 1.5 billion parameters,[40] that focused on using web pages curated by humans to both emphasize document quality and produce a larger, more diverse training dataset.[41] Hence, web pages like Reddit, a social media platform, were scraped "to find upvoted articles and [pull] data from all outbound links in the targeted Reddit posts."[42]

A year later, OpenAI introduced GPT-3 featuring 175 billion parameters,[43] expanding the amount of information in its dataset by using five compilations of data: Common Crawl, WebText2, Books1, Books2, and Wikipedia.[44] Common Crawl is a dataset that has compiled information from over 250 billion web pages since 2007 and is freely available to researchers.[45] Another dataset, WebText2, expands on WebText, the dataset that trained GPT-2, and uses a similar

---

35 Aleksandra Yosifova, *The Evolution of ChatGPT: History and Future*, 365 DATASCIENCE (Aug. 14, 2023), https://365datascience.com/trending/the-evolution-of-chatgpt-history-and-future/ [https://perma.cc/K5T9-2A9G].

36 Mayankchugh Jobathk, *The Evolution of Large Language Models (LLMs): A Journey from GPT to GPT-40*, MEDIUM (July 18, 2024), https://medium.com/@mayankchugh.jobathk/the-evolution-of-large-language-models-llms-a-journey-from-gpt-to-gpt-4o-618765889c98 [https://perma.cc/XES7-7PAN].

37 Art Neill, James Thomas & Erika Lee, *A Framework for Applying Copyright Law to the Training of Textual Generative Artificial Intelligence*, 32 TEX. INTELL. PROP. L.J. 225, 231 (2024).

38 Priya Shree, *The Journey of Open AI GPT Models*, MEDIUM (Nov. 9, 2020), https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2 [https://perma.cc/7QGH-REGW].

39 Jobathk, *supra* note 36.

40 *Id.*

41 Alec Radford et al., *Language Models Are Unsupervised Multitask Learners*, OPENAI at 3, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf [https://perma.cc/42RD-UGP4] (last visited Aug. 20, 2025).

42 Neill, Thomas & Lee, *supra* note 37, at 231.

43 Jobathk, *supra* note 36.

44 Neill, Thomas & Lee, *supra* note 37, at 231–32.

45 *See Frequently Asked Questions*, COMMON CRAWL, https://commoncrawl.org/faq [https://perma.cc/2X9F-7M7H] (last visited Mar. 2, 2025).

selection criterion, targeting websites linked to Reddit posts with at least three upvotes.[46] In contrast, the Books1 and Books2 datasets are less precisely defined but appear to consist of books in the public domain.[47] The final dataset, Wikipedia, seemingly includes all text and data available on the popular and free-content encyclopedia platform.[48]

Finally, OpenAI released its most recent version, GPT-4, in 2023, with a staggering estimation of 1 trillion parameters.[49] Hence, GPT-4 contains the most expansive training dataset yet, containing "web texts, books, news articles, social media posts, code snippets, and other unspecified sources."[50] However, unlike previous iterations of GPT, OpenAI has become a closed-source system, meaning it has completely "closed off" information about the types of data used for the training of this version of GPT.[51]

Despite the closed-source system of OpenAI's latest GPT models, a typical dataset is comprised of text sourced from "the internet, public domain books and research articles, and assorted other free sources of content that researchers can find."[52] However, as AI development continues to advance, the demand for high-quality training data has become a pressing issue.[53] Many AI companies are rapidly depleting readily accessible, high-value sources, prompting concerns over the increasing likelihood that training datasets contain copyrighted material, whether inadvertently or deliberately.[54]

C. The Copyrightability of GPT-4's Training Dataset

Furthermore, distinguishing between public domain works, openly licensed materials, and fully copyrighted texts is critical to evaluate

---

46 Roger Montti, *How to Block OpenAI ChatGPT from Using Your Website Content*, SEARCH ENGINE J. (Feb. 2, 2023), https://www.searchenginejournal.com/how-to-block-chatgpt-from-using-your-website-content/478384/ [https://perma.cc/K2UM-46RK].

47 Gregory Roberts, *AI Training Datasets: The Books1+Books2 That Big AI Eats for Breakfast*, GREGOREITE (Dec. 14, 2022), https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/ [https://perma.cc/28F2-9QAR].

48 *Id.*

49 Jobathk, *supra* note 36.

50 Neill, Thomas & Lee, *supra* note 37, at 232.

51 *Id.*

52 MOLLICK, *supra* note 28, at 7.

53 *Id.* at 7–8.

54 *Id.* at 7.

whether the use of such content by models like GPT-4 (the subject of this lawsuit) aligns with or violates existing copyright law.[55]

A significant portion of the training data for LLMs is likely derived from public domain works. Works in the public domain are no longer protected by copyright law and instead are owned by the public, subject to the public's use without permission.[56] Although the precise nature of information in the datasets Books1 and Books2 is undisclosed, these datasets are widely speculated to consist of, at least in part, books that have entered the public domain.[57]

Wikipedia, on the other hand, is published under an open license, allowing users to access Wikipedia's articles for free and for reuse.[58] Aside from specific quoted content that may fall under separate copyright protections, Wikipedia is generally licensed under the Creative Commons Attribution-ShareAlike 4.0 International License and the GNU Free Documentation License.[59] As a result, Wikipedia's content may typically be used without risk of copyright infringement provided that proper attribution is given, consistent with the terms of these open licenses.[60]

As for the last category of sources, fully protected copyrighted works not subject to open licensing, it is more unclear as to the exact amount of copyrighted material that went into the training of LLMs like GPT-4. At the very least, we know that some copyrighted materials went into the training of GPT-1 because BooksCorpus, one of the

---

55 *Id.* at 233.

56 Richard Stim, *Welcome to the Public Domain*, COPYRIGHT & FAIR USE STANFORD LIBRS., https://fairuse.stanford.edu/overview/public-domain/welcome/#:~:text=The%20term%20%E2%80%9Cpublic%20domain%E2%80%9D%20refers,one%20can%20ever%20own%20it [https://perma.cc/4VPL-2LT9] (last visited Sep. 17, 2025) (Public domain "refers to creative materials that are not protected by intellectual property laws such as copyright, trademark, or patent laws. The public owns these works, not an individual author or artist. Anyone can use a public domain work without obtaining permission, but no one can ever own it.").

57 Roberts, *supra* note 47; *see also* Jack Bandy & Nicholas Vincent, *Addressing "Documentation Debt" in Machine Learning Research: A Retrospective Datasheet for BookCorpus*, https://arxiv.org/pdf/2105.05241 [https://perma.cc/82DB-EBE6] (May 11, 2021).

58 Yana Welinder, *Free as in Open Access and Wikipedia*, THE CTR. FOR INTERNET & SOC'Y (Oct. 20, 2014, at 03:00 PT), https://cyberlaw.stanford.edu/blog/2014/10/free-open-access-and-wikipedia/ [https://perma.cc/BS2B-PTCE].

59 *Wikipedia: FAQ/Copyright*, WIKIPEDIA, https://en.wikipedia.org/wiki/Wikipedia:FAQ/Copyright#Can_I_reuse_Wikipedia's_content_somewhere_else [https://perma.cc/9PSX-NKQY] (last visited Sep. 17, 2025).

60 *Id.*

sources in the dataset, also contained copyrighted books.[61] Moreover, authors of the copyrighted works in BooksCorpus were unable "to opt out of the inclusion of their works" into the GPT training datasets.[62] Therefore, because existing datasets contained works that are fully subject to copyright law restrictions, the *exact* number of such copyrighted works in the dataset is irrelevant; the fact that at least one copyrighted work made it into the dataset is sufficient.

II
THE *TIMES*'S ALLEGATIONS

In late December 2023, *The New York Times* filed an action against defendants OpenAI, Inc. and Microsoft Corporation alleging seven counts of unlawful use and infringement of the *Times*'s articles, with its lead claim being direct copyright infringement under 17 U.S.C. § 501.[63]

The central claim[64] is that OpenAI has infringed upon the *Times*'s copyrights through the unlicensed and unauthorized use and reproduction of the *Times*'s content in the training of its GPT models.[65] First, the complaint alleges that LLMs like GPT are capable of "memorizing" or "regurgitating" portions of works included in their training data.[66] When this occurs, the models can generate near-verbatim reproductions of the original works the models were trained on, such as the *Times*'s copyrighted articles in the dataset.[67] Second, and relatedly, the LLMs can produce synthetic search results that, when prompted, reproduce "significantly more expressive content from [an]

---

61 *See supra* text accompanying notes 50–54.

62 Neill, Thomas & Lee, *supra* note 37, at 234; *see also* Bandy & Vincent, *supra* note 57.

63 New York Times Complaint, *supra* note 4, at 60−67.

64 The *Times* Complaint asserts seven counts against Microsoft Corporation and OpenAI, Inc. and its related business entities. The following Counts are alleged against all defendants: Copyright Infringement under 17 U.S.C. § 501, Contributory Copyright Infringement, Violation of the Digital Millennium Copyright Act under 17 U.S.C. § 1202, Common Law Unfair Competition by Misappropriation, and Trademark Dilution under 15 U.S.C. § 1125(c). The *Times* also alleges Vicarious Copyright Infringement against Microsoft, OpenAI, Inc., OpenAI LP, OAI Corporation LLC, OpenAI Holdings LLC, and OpenAI Global LLC. The *Times* also has a separate Contributory Copyright Infringement claim against just Microsoft Corporation. For the purposes of this Comment, I focus on only Count I: Copyright Infringement under § 501 as it pertains to defendant OpenAI, Inc.

65 New York Times Complaint, *supra* note 4, at 2.

66 *Id.* at 23−24.

67 *Id.*

original article than what would traditionally be displayed" through an online search engine.[68] As the *Times* alleges, this ability enables users to circumvent the *Times*'s paywall and undermine its subscription-based business model.[69] At its core, the *Times* stresses that these practices gravely threaten high-quality journalism,[70] because if OpenAI users can easily generate summaries or near-verbatim reproductions of the *Times*'s articles, at no cost or consequence, it would "obviate the need" to purchase access to the *Times* directly, thereby undermining its business model.[71]

The complaint also emphasizes that, given the need for high-quality textual works in GPT datasets,[72] the algorithm was designed to disproportionately curate the *Times*'s works.[73] The complaint alleges that the WebText dataset was created with the intention of securing high-quality documents and, thus, contains "a staggering amount of scraped content from the *Times*," with the *Times*'s domain name being "one of the 'top 15 domains by volume' in the WebText dataset."[74] As the complaint alleges, because OpenAI acknowledges that "datasets [it] view[s] as higher-quality are sampled more frequently" during the training process, by its own admission, it suggests that high-quality content, such as material from the *Times*, played a more significant and valuable role in training GPT models compared to lower-quality sources.[75]

Ultimately, the complaint underscores the societal stakes, stating that "[i]ndependent journalism is vital to our democracy" and is "increasingly rare and valuable."[76] In contrast, the *Times* characterizes OpenAI as merely a "multi-billion-dollar for-profit business built in large part on the unlicensed exploitation of copyrighted works" that shed its nonprofit status a mere three years after its founding.[77] The *Times*'s complaint suggests that OpenAI obtained public trust through

---

68 *Id.* at 40, 42, 44, 46.

69 *Id.* at 59.

70 *Id.* at 14.

71 *Id.* at 22.

72 *See supra* text accompanying notes 50−54.

73 New York Times Complaint, *supra* note 4, at 20.

74 *Id.* at 25.

75 *Id.* at 27.

76 *Id.* at 1.

77 *Id.* at 17.

its brief nonprofit and open-source phase to, within a few years, become what it once promised never to be.[78]

III
OPENAI'S DEFENSES

Defendant OpenAI argues that the crux of the *Times*'s infringement allegations rest on "training data regurgitation," which are merely "uncommon and unintended phenomena" that OpenAI has already identified as a problem and is actively working to resolve.[79] OpenAI even goes as far as accusing the *Times* of "pa[ying] someone to hack" its products, claims the *Times* engaged in "deceptive prompt[ing]" to exploit a "bug," and states the *Times* was able to obtain the verbatim outputs in Exhibit J of the Complaint only by engaging in "tens of thousands of attempts."[80] This, OpenAI argues, violates its terms of use and is not how "[n]ormal people" use its products.[81]

More importantly, OpenAI argues that its use of the *Times*'s articles for the training of its GPT models constitutes fair use.[82] In particular, OpenAI argues that it uses the copyrighted content as a mere cog in the technological process and development of its LLMs to produce "new, different, and innovative products."[83]

Moreover, OpenAI seeks to partially dismiss the direct infringement claim, arguing the claim is time-barred by the three-year limitations period because OpenAI's use of the *Times*'s articles to train the LLMs occurred more than three years ago.[84] However, as the *Times* notes in its opposition brief, "OpenAI's statute-of-limitations argument is narrow, addressing only the training of GPT-2 in 2019 and GPT-3 in 2020 . . . . [i]t does not apply to the 'orders of magnitude more

---

78 *See Introducing OpenAI*, *supra* note 13; Chloe Xiang, *OpenAI Is Now Everything It Promised Not to Be: Corporate, Closed-Source, and For-Profit*, VICE (Feb. 28, 2023, at 12:35 PT), https://www.vice.com/en/article/openai-is-now-everything-it-promised-not-to-be-corporate-closed-source-and-for-profit/ [https://perma.cc/S7P8-JSTJ].

79 Defendant's Memorandum in Support of Motion to Dismiss, *supra* note 9, at 10–11.

80 *Id.* at 2. Excerpts of Exhibit J in Plaintiff's complaint show a side-by-side comparison of ChatGPT-4's outputs and various *New York Times* articles*; see also id.* at 30–37.

81 *Id.*

82 *OpenAI and Journalism*, OPENAI (Jan. 8, 2024), https://openai.com/index/openai-and-journalism/ (on file with the Oregon Law Review); *see also* Bobby Allyn, *'The New York Times' Takes OpenAI to Court. ChatGPT's Future Could Be on the Line*, NPR (Jan. 14, 2025, at 16:27 ET), https://www.npr.org/2025/01/14/nx-s1-5258952/new-york-times-openai-microsoft (on file with the Oregon Law Review).

83 Defendant's Memorandum in Support of Motion to Dismiss, *supra* note 9, at 3.

84 *Id.* at 15.

powerful' GPT-3.5 and GPT-4 models developed in 2022 and 2023."[85] Interestingly, prior to filing a formal legal response, OpenAI publicly responded to the *Times*'s complaint in a blog post delineating four main points: (1) OpenAI "collaborate[s] with news organizations," (2) OpenAI's use of copyrighted materials to train its LLMs constitutes fair use and is hence not infringement, (3) regurgitation is a rare phenomenon that it is "working to drive to zero," and (4) the regurgitations the *Times* argues as a basis for its infringement claim are so out of the ordinary that it raises inferences that the *Times* "either instructed the model to regurgitate or cherry-picked their examples from many attempts."[86]

IV
MERITS OF THE LEGAL CLAIMS

While the legal dispute is likely to turn on the court's application of the fair use defense, the *Times* will still have to establish direct copyright infringement. Where it is successful in demonstrating direct infringement, it will similarly need to show that OpenAI's fair use defense is inapplicable in this case. Given the relatively low bar for establishing ownership and unauthorized copying,[87] and the available evidence suggesting the *Times*'s content was used in training datasets, the *Times* is likely to succeed in making out a prima facie case of copyright infringement.[88] As a result, the disposition of the case will likely hinge on OpenAI's fair use defense argument. This Section IV.B offers a detailed application of the four statutory fair use factors.

A. Direct Copyright Infringement

Under the Copyright Act of 1976, 17 U.S.C. § 101, the owner of a copyrighted work has the exclusive right to reproduce, perform publicly, display publicly, prepare derivative works of, and distribute copies of their copyrighted work.[89] Where a defendant violates a plaintiff's exclusive rights in their copyrighted work, the defendant is

---

85 Plaintiff's Memorandum in Opposition to OpenAI Defendants' Partial Motion to Dismiss at 6, N.Y. Times Co. v. Microsoft Corp., No. 23-CV-11195 (S.D.N.Y. Apr. 4, 2025).

86 *OpenAI and Journalism*, *supra* note 82.

87 *See, e.g.*, 17 U.S.C. § 410(c) ("[T]he certificate of a registration made before or within five years after first publication of the work shall constitute prima facie evidence of the validity of the copyright . . . .").

88 Defendant's Memorandum in Support of Motion to Dismiss, *supra* note 9, at 7.

89 17 U.S.C. § 106.

liable for damages.[90] Therefore, for the *Times* to prevail on its direct copyright infringement claim against OpenAI, it must establish by a preponderance of the evidence that (1) it owns a valid copyright, and (2) the defendant copied "constituent elements of the work that are original."[91] The second element requires showing both that OpenAI actually copied the work and that its copy was substantially similar to the work.[92]

Copyright registrations are "prima facie evidence of the validity of the copyright" if "made before or within five years after first publication of the work."[93] Given that the *Times* has registered its articles, it is unlikely OpenAI will be successful in disputing this element.[94]

Hence, the major legal disputes are likely to occur under the second element. To satisfy the second element of a copyright claim, the *Times* must demonstrate that OpenAI copied its work and that such copying was unlawful because OpenAI's allegedly infringing work is substantially similar to the protectable elements of the *Times*'s copyrighted works.[95] Actual copying means that "the defendant did, in fact, use the copyrighted work in creating his own."[96] Because direct evidence of copying is difficult to establish, the *Times* can rely on circumstantial evidence of copying "'by demonstrating that the person who composed the defendant's work had access to the copyrighted material,' and that there are similarities between the two works that are 'probative of copying.'"[97] The *Times* will likely be able to establish that OpenAI had access to its copyrighted articles because, in July of 2020, "OpenAI had disclosed that the *Times*'s articles were a tiny part of the diverse datasets that had been used to train these language models[,]" establishing that OpenAI undoubtedly had access to such

90 Warren v. John Wiley & Sons, Inc., 952 F. Supp. 2d 610, 616 (S.D.N.Y. 2013).

91 Disney Enter., Inc. v. Sarelli, 322 F. Supp. 3d 413, 441 (S.D.N.Y. 2018) (quoting Muller v. Twentieth Century Fox Film Corp., 794 F. Supp. 2d 429, 439 (S.D.N.Y. 2011)).

92 Dam Things from Den. v. Russ Berrie & Co., Inc., 290 F.3d 548, 561−62 (3d Cir. 2002).

93 17 U.S.C. § 410(c).

94 *Copyright Notice*, THE NEW YORK TIMES, https://help.nytimes.com/hc/en-us/articles/115014792127-Copyright-Notice [https://perma.cc/6JF8-CPC2] (last visited Sept. 13, 2024).

95 *Sarelli*, 322 F. Supp. 3d at 442.

96 Tanksley v. Daniels, 902 F.3d 165, 173 (3d Cir. 2018).

97 Jorgensen v. Epic/Sony Recs., 351 F.3d 46, 51 (2d Cir. 2003) (first quoting Herzog v. Castle Rock Ent., 193 F.3d 1241, 1249 (11th Cir. 1999); and then quoting Repp v. Weber, 132 F.3d 882, 889 (2d Cir. 1997)).

copyrighted works.[98] Additionally, because the *Times* is popular worldwide, it is unlikely that "access" will be a major issue in this case.

Moreover, the *Times* must still prove that the similarities between its copyrighted articles and OpenAI's outputs are so similar that an "average lay observer would recognize the alleged copy as having been appropriated from the copyrighted work" under the substantial similarity analysis.[99] Because the case here involves works of literary expression, the *Times* must demonstrate that the similarities shared by its articles and OpenAI's outputs are "something more than mere generalized ideas or themes."[100] It is likely that the *Times* will be able to establish a substantial similarity to OpenAI's outputs—at least in terms of Exhibit J in the complaint, which introduced one hundred examples[101] where ChatGPT was able to nearly regurgitate parts of the *Times*'s articles verbatim and provide summaries that are "significantly longer and more detailed" than what is traditionally accessible through normal search engines.[102]

B. Application of the Fair Use Defense

The fair use defense operates as a limitation on a copyright owner's rights of exclusivity to promote the breathing room necessary to promote creative expression.[103] The fair use defense to copyright infringement is a question of law that is determined by underlying facts on a case-by-case basis.[104] Hence, although it is not possible to categorically assert that every aspect of the LLMs' training are per se protected under a fair use defense, the discussion below examines how the training of generative AI aligns with existing fair use jurisprudence.

Under 17 U.S.C. § 107, the fair use defense requires consideration of

> (1) the purpose and character of the use, including whether such use is of a commercial nature or is for nonprofit educational purposes;

---

98 Defendant's Memorandum in Support of Motion to Dismiss, *supra* note 9, at 7 (OpenAI admits the *Times*'s articles exist in the training dataset but argues their use is protected under the fair use doctrine).

99 *E.g.*, *Sarelli*, 322 F. Supp. 3d at 442 (quoting Blakeman v. The Walt Disney Co., 613 F. Supp. 2d 288, 304 (E.D.N.Y. 2009); *see also* Warner Bros. Inc. v. Am. Broad. Cos., Inc., 654 F.2d 204, 208 (2d Cir. 1981) ("In the case of literary works, it is axiomatic that copyright protection only extends to the expression of the author's idea, not to the idea itself.").

100 *Warner Bros. Inc.*, 654 F.2d at 208.

101 New York Times Complaint, *supra* note 4, at 2.

102 *Id.* at 3.

103 Green v. U.S. Dep't of Just., 111 F.4th 81, 101 (D.C. Cir. 2024).

104 *See* Google LLC v. Oracle Am., Inc., 593 U.S. 1, 23 (2021).

> (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 value . . . of the copyrighted work.[105]

1. Purpose and Character

Although courts must consider all four fair use factors, the "purpose and character" of the use has historically played a central role in the analysis.[106] This factor addresses "[t]he heart of the fair use inquiry"—namely, whether the defendant's use is "transformative," and whether it is for commercial or nonprofit educational purposes.[107] A use is transformative if it uses the original work to create something new, with a different purpose or character, rather than merely superseding the original.[108] Conversely, a use is not transformative where it does not serve a "further purpose or different character" than the original work, meaning that the use merely acts as a substitute or replacement for the original work.[109]

17 U.S.C. § 107 provides an inexhaustive list of potentially fair-use, transformative purposes—"criticism, comment, news reporting, teaching . . . scholarship, or research"[110]—all of which indicate uses that are incapable of supplanting the original works' use.[111] However, given that most secondary uses add something at least slightly novel, this factor asks "'whether and to what extent' the use at issue has a

---

105 17 U.S.C. § 107.

106 Nat'l Acad. of Television Arts and Scis., Inc. v. Multimedia Sys. Design, Inc., 551 F. Supp. 3d 408, 421−22 (S.D.N.Y. 2021).

107 *Id.*

108 Richard Stim, *Measuring Fair Use: The Four Factors*, COPYRIGHT & FAIR USE STANFORD LIBR., https://fairuse.stanford.edu/overview/fair-use/four-factors/ [https://perma.cc/84P2-BXGW] (last visited Sept. 14, 2025).

109 Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 525 (2023).

110 17 U.S.C. § 107.

111 *Andy Warhol*, 598 U.S. at 528 ("Criticism of a work, for instance, ordinarily does not supersede the objects of, or supplant, the work. Rather, it uses the work to serve a distinct end."); *see also* Folsom v. Marsh, 9 F. Cas. 342, 344−45 (C.C.D. Mass. 1841) (No. 4,901,2) ("[I]t is as clear, that if he thus cites the most important parts of the work, with a view, not to criticise, but to supersede the use of the original work, and substitute the review for it, such a use will be deemed in law a piracy."). *But see* Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 561 (1985) ("The fact that an article arguably is 'news' and therefore a productive use is simply one factor in a fair use analysis.").

purpose or character different from the original."[112] In sum, the first fair use factor weighs the degree to which the secondary use of a copyrighted work serves a distinct purpose or transforms the original in some meaningful way against the commercial nature of the use.[113] When a secondary use closely mirrors the purpose of the original work and is driven by profit, the first factor will generally weigh against a finding of fair use, unless compelling justification for the copying exists.[114]

Importantly, a secondary use that merely adds new expression, meaning, or message—but retains the same purpose as the original—does not qualify as transformative.[115] In *Andy Warhol*, a magazine hired Goldsmith to photograph the musician Prince.[116] Years later, Andy Warhol used one of Goldsmith's copyrighted photographs of Prince without permission as the basis for a series of colorful prints, one of which was licensed and published on the cover of a popular magazine.[117] Despite their stylistic differences, the Court found that Warhol's prints served the same commercial purpose as Goldsmith's original photographs: to illustrate articles about Prince in magazines.[118] Therefore, although the two works "were not perfect substitutes" of each other, because both works competed in the same licensing market, the Court held that Warhol's use was not transformative and weighed against a finding of fair use.[119]

Conversely, where a use does not replicate the original copyrighted work but merely furnishes a searchable database and provides information about the original work, the use is likely transformative.[120] In *Authors Guild*, Google scanned millions of books for its Google Books project, including works in the public domain and copyrighted materials, without the consent of the owners of such materials.[121] Google kept the original scans of these books, but allowed its users to

---

112 *Andy Warhol*, 598 U.S. at 529 (quoting Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 579 (1994)) ("The larger the difference, the more likely the first factor weighs in favor of fair use.").

113 *Id.; see also* Google LLC v. Oracle Am., Inc., 593 U.S. 1, 29 (2021).

114 *Andy Warhol*, 598 U.S. at 531−32.

115 *Id.* at 550.

116 *Id.* at 516.

117 *Id.* at 518−19.

118 *Id.* at 550.

119 *Id.* at 536−37.

120 Authors Guild v. Google, Inc., 804 F.3d 202, 219 (2d Cir. 2015).

121 *Id.* at 208.

access only a limited view of the text in response to queries.[122] The plaintiffs challenged Google on the basis of two uses: (1) its use in providing a search engine for users to determine whether a key term appeared in a book, and if so, how many times, and (2) its use of the copyrighted works to allow users to see a "snippet" or a portion of the original text.[123] Because Google copied the original books for the purpose of making information about those books available to its users, rather than serving merely as a substitute for the books, the court held that Google's use was transformative because it fundamentally altered the purpose of the original works.[124] Moreover, this transformation conferred significant public benefits, such as facilitating scholarship and expanding access to knowledge, without supplanting the market for the original books.[125] Hence, although Google derived a commercial benefit from the project through advertising, the court found that this factor did not outweigh the transformative nature of the use.[126]

Similar to *Andy Warhol*, and distinct from *Authors Guild*, OpenAI's use of the *Times*'s articles to train its generative AI tools does not alter the fundamental purpose of the original works. The *Times*'s purpose in the creation of its articles centers around investigative reporting, breaking news reporting, beat reporting, reviews and analysis, and commentary and opinion.[127] Arguably, OpenAI's use is to train its LLMs for the ultimate purpose of providing reports, analysis, commentary, and opinion-based outputs to its users.[128] Therefore, just as Warhol's prints were not "perfect substitutes" for Goldsmith's photograph[129] but nevertheless shared a similar purpose and commercial function, OpenAI's ingestion of the *Times*'s articles serves

---

122 *Id.* at 209.

123 *Id.* at 222.

124 *Id.* at 218.

125 *See id.* at 207.

126 *Id.* at 219.

127 New York Times Complaint, *supra* note 4, at 10−12.

128 *See How ChatGPT and Our Foundation Models Are Developed*, OPENAI, https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed (on file with the Oregon Law Review) (last visited June 27, 2025) (Users can "use ChatGPT for a wide range of tasks, including organizing and summarizing information, assisting with translations, analyzing or generating images, inspiring creativity and ideas, and other everyday activities.").

129 Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 536 (2023).

a purpose closely aligned with that of the source material: the communication of information.

Moreover, in contrast to *Authors Guild*, OpenAI's use fundamentally differs in terms of function and output. First, whereas Google Books was a tool for locating information within a book, thus not acting as an actual substitute for the information itself,[130] OpenAI's LLMs ingest the expressive content of the *Times*'s articles for the *purpose* of generating new textual outputs that resemble, and at times actually copy the original work's substance, style, and overall expression.[131] Second, unlike *Authors Guild*, GPT's outputs go beyond what is revealed in typical snippets or ordinary search results.[132] Even when these AI-generated summaries include links or citations to the original source, users have fewer incentives to actually navigate to those sources because the expressive content, either quoted or paraphrased, is already incorporated into the narrative response.[133] In fact, the inclusion of attribution may inadvertently increase user trust in the summary itself, reducing the likelihood that users will click through to the original article.[134] As a result, users may never visit the *Times*'s website at all, directly undermining its ability to monetize access to its content—an ability the Copyright Clause is intended to provide.[135]

Furthermore, OpenAI's status as a for-profit enterprise undermines its fair use defense. The commercial nature of OpenAI's activities, especially its monetization of the GPT product line through premium subscriptions, further undermines its fair use defense.[136] Therefore, given the similarity and near synonymous purpose of the parties' use, in combination with the commercial nature of the use, it is likely that the purpose and character factor will weigh against the application of the fair use defense.

On the other hand, OpenAI may argue that its purpose is narrower, using the *Times*'s articles to merely train its LLMs to understand

---

130 Authors Guild v. Google, Inc., 804 F.3d 202, 224 (2d Cir. 2015).

131 *See* New York Times Complaint, *supra* note 4, at 30–37.

132 *See id.* at 2.

133 *Id.* at 37.

134 *Id.*

135 U.S. CONST. art. I, § 8, cl. 8; *see also* Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 562 (1985) ("[E]very commercial use of copyrighted material is presumptively an unfair exploitation of the monopoly privilege that belongs to the owner of the copyright.") (internal quotations omitted).

136 *See Introducing ChatGPT Plus*, OPENAI (Feb. 1, 2023), https://openai.com/index/chatgpt-plus/ (on file with the Oregon Law Review).

the mechanics and basic rules of human language rather than to inform readers about current events.[137] While this distinction may be technically correct at the training stage, the ultimate purpose of the use is to generate outputs readable by humans that can and have regurgitated content in its dataset verbatim, particularly that of the *Times*'s content.[138] Under such circumstances, the purported distinction between "training" and "informing" collapses. The former is instrumental to the latter, and OpenAI's models, trained on the *Times*'s content, fulfill the same market facing role as the original works. Thus, the purpose and character of OpenAI's uses are insufficiently transformative and weigh against a finding of fair use.

2. Nature of the Copyrighted Works

The second factor, the nature of the copyrighted work, evaluates where the original work lies on the spectrum between factual and creative, with the underlying premise that creative works "are closer to the core of intended copyright protection."[139] The *Times*'s articles, as news reporting, generally reside closer to the factual end of the spectrum that, again, can favor fair use. However, copyright law nevertheless safeguards the author's particular *expression* of those facts, including the selection of facts, the ordering of facts, and the overall arranging of facts.[140]

Facts as individual propositions exist within the public domain; however, when an author compiles those facts in a way that reflects their original creativity, judgment, and effort, the protection of copyright law is more likely to become available.[141] In *Schroeder*, the plaintiff compiled a book listing the names and addresses of suppliers of gardening items and tools.[142] The defendant copied the names and

---

137 Defendant's Memorandum in Support of Motion to Dismiss, *supra* note 9, at 4–5.

138 *See* New York Times Complaint, *supra* note 4, at 30–37.

139 Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 586 (1994); *see also* Hachette Book Grp., Inc. v. Internet Archive, 115 F.4th 163, 187 (2d Cir. 2024).

140 Authors Guild v. Google, Inc., 804 F.3d 202, 220 (2d Cir. 2015) ("The mere fact that the original is a factual work therefore should not imply that others may freely copy it. 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."); *Harper & Row*, 471 U.S. at 561 ("The fact that an article arguably is 'news' and therefore a productive use is simply one factor in a fair use analysis.").

141 Schroeder v. William Morrow & Co., 566 F.2d 3, 5-6 (7th Cir. 1977) ("Another is entitled to make his own compilation of the same names and addresses, using information in the public domain, but he is not entitled merely to copy the copyrighted list.").

142 *Id.* at 4.

addresses from almost fifty percent of the plaintiff's book.[143] Even though the defendant copied only the names and addresses from the plaintiff's book and not any of "the accompanying descriptive material," the court rejected the fair use defense on the ground that copyright protection extends to "'the selection, the ordering and arrangement' of the names and addresses."[144]

Here, it is unlikely OpenAI's use does not merely involve the extraction of bare facts or general ideas available in the public domain.[145] Instead, the *Times* presents evidence that OpenAI's models can reproduce entire paragraphs from its articles, capturing not just factual content, but the original phrasing, sequencing, and stylistic elements that reflect significant editorial judgment.[146] If the mere compilation of names and addresses is sufficient to reject the fair use defense, surely the *Times*'s worldwide, three-time Pulitzer-award-winning articles are entitled to copyright protection,[147] even if the main components are facts.

Furthermore, OpenAI's use does not merely involve isolated factual information or general ideas drawn from the *Times*'s reporting. Instead, OpenAI's models, as shown in Exhibit J,[148] can reproduce substantial portions of the *Times*'s articles, including entire paragraphs.[149] The copying of lengthy, coherent portions of original text evidence a direct taking of the author's expressive choices, regardless of the work's fictional or nonfictional status.

Moreover, the court will consider whether the original work is published, in which an unpublished original work is more likely to weigh against a finding of fair use.[150] The *Times*'s articles are published works; however, the mere fact of publication alone is insufficient to satisfy the second factor in the face of a defendant's wholesale copying of entire paragraphs of narrative text.

---

143 *Id.* at 4–5.

144 *Id.* at 6.

145 *See supra* text accompanying notes 50–54.

146 *See* New York Times Complaint, *supra* note 4, at 30–37.

147 *See Awards and Recognition*, N.Y. TIMES, https://www.nytco.com/ [https://perma.cc/JR4T-83G5].

148 *See* New York Times Complaint, *supra* note 4, at 30–37.

149 Stim, *supra* note 56.

150 Authors Guild v. Google, Inc., 804 F.3d 202, 220 (2d Cir. 2015).

Ultimately, copyright law does not require absolute novelty or inventiveness—only a minimal degree of creativity.[151] Thus, the *Times* need not demonstrate that its reporting introduces new facts or revolutionary ideas; rather, it must show that its articles embody original expressive choices in how facts are selected, framed, and communicated to the public. Therefore, even if the factual and published nature somewhat reduces the *Times*'s argument, the overall fair use argument will likely still weigh in the *Times*'s favor because the purpose of OpenAI's uses is nearly synonymous with the *Times*'s, and therefore it can act as a direct market substitute.

3. Amount and Substantiality

The third factor asks whether "'the amount and substantiality of the portion used in relation to the copyrighted work as a whole' are reasonable in relation to the copying's purpose."[152] Under this factor, where the most important parts of the original work are copied or are largely copied, a court is less likely to find that this factor weighs in favor of fair use because it increases the likelihood that the defendant's work is an effective substitute for the original.[153]

Even where the infringement is quantitatively small, if the infringement captures the most valuable and expressive portion of the original copyrighted work—in other words, the "heart" of the work—the infringement is considered substantial under the third factor.[154] In *Harper & Row*, former President Ford contracted with a magazine the exclusive right to publish his memoirs that contained his reflections on the Watergate crisis and his pardon of former President Nixon.[155] Before the magazine could publish excerpts of the memoir, a political commentary magazine, *The Nation*, obtained an unauthorized copy of the memoir and published roughly 300 words of copyrighted material.[156] As a result of *The Nation*'s article, the contracted magazine

---

151 Feist Publ'ns, Inc. v. Rural Tel. Serv. Co., Inc., 499 U.S. 340, 358–59 (1991) (holding copyright protects original works that possess a minimal level of creativity regardless of whether the underlying information is fact or fiction).

152 Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 570 (1994).

153 *Authors Guild*, 804 F.3d at 221.

154 Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 565 (1985) ("[A] taking may not be excused merely because it is insubstantial with respect to the *infringing* work.").

155 *Id.* at 543.

156 *Id.*

canceled its piece.[157] Despite the brevity of the excerpt, the Court found that *The Nation*'s use weighed against fair use because the excerpt included key revelations about Ford's decision to pardon Nixon that got to the heart of the author's personal reflections.[158]

Where the defendant's copying of the original work is necessary to achieve its transformative purpose, the amount and substantiality factor weighs in favor of a finding of fair use.[159] In *Authors Guild*, because Google's use of the original works was to create a search function, it necessarily needed to copy the entire original work, or else its search function would not be able to perform its intended, and transformative, purpose: aid online searchers with determining whether their searched term appears in a book.[160] As a result, even though Google used the entire original work, and thus its use was substantial in amount, the court held that in terms of the search function, the amount and substantiality factor weighed in favor of fair use.[161] Moreover, Google's second challenged use, its snippet feature, was not found to be so substantial in amount as to weigh against fair use.[162] Although the plaintiff was able to provide evidence that some searches revealed nearly sixteen percent of their book's text, Google's snippet use is not substantial because the user saw only scattered and fragmented sections of the book's text.[163] In this way, there was not a significant risk that Google's use would act as a substitute for the original books because it did not coherently communicate the content of the original book.[164]

As shown in Exhibit J,[165] the fact that a substantial portion of the infringing work is copied nearly verbatim is powerful evidence of the qualitative importance of the material both to the original creator and to the infringer.[166] Again, GPT-4's dataset is relatively unknown;[167] however, the *Times* alleges that its webpage domain, NYTimes.com, is one of the "'top 15 domains by volume' in the WebText dataset, and is listed as the 5th 'top domain' in the WebText dataset with 333,160

157 *Id.*

158 *Id.*

159 Authors Guild v. Google, Inc., 804 F.3d 202, 221 (2d Cir. 2015).

160 *Id.*

161 *Id.* at 222–23.

162 *Id.* at 223.

163 *Id.*

164 *Id.* ("If snippet view could be used to reveal a coherent block amount to 16% of a book, that would raise a very different question beyond the scope of our inquiry.").

165 *See* New York Times Complaint, *supra* note 4, at 30–37.

166 Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 565 (1985).

167 *See supra* text accompanying notes 50–54.

entries."[168] Even if the use of the *Times*'s articles is unintentional or necessary for training, the model's verbatim reproduction of paragraphs, OpenAI's stated aim to use the highest-quality training materials, and the *Times*'s outsized presence in WebText strongly support the Times's claims. After all, if only small fragments or minimal samples had been included, such detailed recitations would be far less likely.[169]

4. The Use's Effect on the Market Value of the Copyrighted Work

The last factor is how the defendant's use of the original work affects "the potential market for or value of the copyrighted work."[170] This factor examines whether the defendant's use functions as a market substitute for the original work, thereby causing the copyright owner economic harm by significantly reducing potential revenues due to consumer preference for the defendant's copy over the original.[171] This factor is essential in the fair use analysis because copyright is ultimately a "commercial doctrine" that aims to encourage creativity by ensuring that authors profit from their creations.[172] Moreover, this factor is closely linked with the first factor in the fair use defense analysis (the purpose and character of the use) in that "the more the copying is done to achieve a purpose that differs from the purpose of the original, the less likely it is that the copy will serve as a satisfactory substitute for the original."[173]

Where a defendant's use can possibly, probably, or even certainly cause some loss of sales, the loss of sales alone is insufficient to make the defendant's copy an effective substitute that would weigh against a finding of fair use.[174] In *Authors Guild*, the court found the fourth factor did not weigh against a finding of fair use because Google's

168 New York Times Complaint, *supra* note 4, at 25.

169 *See id.* at 30–37.

170 Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569, 570 (1994).

171 Authors Guild v. Google, Inc., 804 F.3d 202, 223 (2d Cir. 2015).

172 *Id.*; *see also* Harper & Row Publishers, Inc. v. Nation Enters., 471 U.S. 539, 566 (1985) (describing the fourth factor as "undoubtedly the single most important element of fair use").

173 Authors Guild, 804 F.3d at 223 (citing Campbell, 510 U.S. at 591).

174 *Id.* at 224 (holding that the fourth factor weighs in favor of fair use despite the recognition that the "snippet function can cause *some* loss of sales . . . . [Because] the possibility, or even the probability or certainty, of some loss of sales does not suffice to make the copy an effectively competing substitute that would tilt the weighty fourth factor in favor of the rights holder in the original").

snippet function was so brief, "cumbersome, disjoined, and incomplete" it could hardly be said that the snippet view could act as a significant substitute for the author's entire book.[175]

This factor is likely to weigh against the application of the fair use defense. Here, OpenAI's use undermines the *Times*'s ability to license content and threatens existing and potential revenue streams, including through content syndication and licensing arrangements. Therefore, unlike the fragmented search facilitation at issue in *Authors Guild*, where the output consisted of isolated, non-substitutive snippets, OpenAI's ingestion and deployment of the *Times*'s works risks enabling outputs that function as close substitutes for the original articles. ChatGPT's ability to generate outputs resembling news reporting—sometimes in substantial portions of the original text—means that users could turn to OpenAI's platform for information rather than consulting original journalism. This dynamic directly undermines both the market for the *Times*'s work and the exclusivity that copyright law is designed to protect. In this respect, OpenAI's use arguably crosses a line: Rather than serving a distinct transformative function, it imperils the very market in which the *Times* operates, weighing heavily against a finding of fair use under the fourth factor.

V
BROADER LEGAL AND SOCIAL CONSEQUENCES

This Part evaluates the policy implications of a verdict for either OpenAI or *The New York Times*, ultimately concluding that the latter better serves the constitutional purpose of copyright. It first examines the broader policy consequences of a verdict favoring the *Times*, including concerns about algorithmic bias and the limitation of training data to public domain or licensed sources. While these concerns raise important questions about representation and technological development, they do not justify weakening the core protections of copyright. The Part then turns to the implications of a ruling in favor of the *Times*, highlighting how enforcing copyright in the AI context promotes creative sustainability and protects the livelihoods of authors. Although both sides raise compelling concerns, this Part argues that the policy arguments ultimately favor the *Times*, whose position better aligns with the Copyright Clause's foundational goal: incentivizing

---

175 *Id.* at 224–25.

human creativity by granting authors meaningful control over the use of their work.[176]

A. The Possibility of Algorithmic Gatekeeping

If the *Times* prevails in its lawsuit and the court mandates the destruction of GPT's training datasets, the implications will be far-reaching. If existing datasets were disallowed, thus forcing AI developers to rebuild, developers would face a critical decision: either rely exclusively on public domain works or selectively license copyrighted materials. Existing datasets are curated through web crawlers, whose objective is to obtain as many materials as possible for the LLM's training, regardless of the source or content.[177] While the training materials are inherently biased by virtue of societal conditions,[178] the risks associated with adding an additional layer of discretionary bias would only exacerbate and amplify that problem.[179] This additional level of discretionary bias is particularly harmful given that "AI systems enjoy an aura of objectivity and accuracy."[180] This appearance of neutrality may result in a reinforcement and normalization of existing prejudices through supposedly factual outputs.

Ultimately, if AI developers relied solely on public domain works, an entirely new set of problems related to bias may be introduced. Copyrighted works from 1929 entered the U.S. public domain on

---

176 U.S. CONST. art. I, § 8, cl. 8.

177 *See Overview of OpenAI Crawlers*, OPENAI PLATFORM, https://platform.openai.com/docs/bots/ (on file with the Oregon Law Review) (last visited Apr. 14, 2025); Duncan Anderson, *Beyond Data Hoovering: The Nuanced Reality of Training Large Language Models (LLMs)*, MEDIUM (July 19, 2023), https://medium.com/barnacle-labs/beyond-data-hoovering-the-nuanced-reality-of-training-large-language-models-llms-79aa107c17db [https://perma.cc/9A6U-L9Z7].

178 Thomas Dethmann & Jannis Spiekermann, *Ethical Use of Training Data: Ensuring Fairness and Data Protection in AI*, LAMARR INST. FOR MACH. LEARNING & A.I. (July 3, 2024), https://lamarr-institute.org/blog/ai-training-data-bias/ [https://perma.cc/9UBE-29H8].

179 *See* Uwe Peters, *Algorithmic Political Bias in Artificial Intelligence Systems*, 35 PHIL. & TECH., Mar 2022, at 25, 2 (2022), https://pmc.ncbi.nlm.nih.gov/articles/PMC8967082/ [https://perma.cc/K6VF-VCGH] (Algorithmic bias is defined as "a tendency to not merely neutrally transform or extract information from data but to operate on it in ways that deviate from a normative (moral, statistical, social, etc.) standard such that one kind of individual or group is unfairly privileged over another based on aspects of their social identity.").

180 *Id.*

January 1, 2025.[181] Hence, because the materials in the public domain are from 1929 and before, these materials predominantly skew toward older, Western-centric, and historically privileged perspectives.[182] The historic exclusion of marginalized voices in publishing contribute to the disproportionate representation of Western-centric perspectives.[183] Undoubtedly, while the publication of literary works in the modern age still portrays certain class-based biases, those biases were inflated in the early twentieth century when "publication was the near-exclusive domain of white males born to a privileged class" and although "exceptions certainly exist[ed], they were relatively rare and special cases."[184] As a result, AI models trained solely on public domain works risk perpetuating outdated viewpoints while also failing to capture contemporary and diverse discourses. These biases, in turn, affect the inclusivity and objectivity of AI-generated knowledge.

The process of selecting and licensing training data is neither neutral nor purely technical—it is an act of curation that reflects the individual priorities, limitations, and biases of AI developers. The economic considerations underlying AI training data licensing can significantly affect the diversity of the content that is incorporated into models. To be able to use works outside the public domain, AI developers will need to negotiate a license with the relevant copyright owners.[185] There are license fees and transaction costs associated with the negotiation of a license.[186] The license fee is the actual price a company pays to obtain the rights to use the licensed material whereas the transaction costs are the expenses incurred to make the transaction

---

181 Jennifer Jenkins & James Boyle, *January 1, 2025 Is Public Domain Day: Works from 1929 Are Open to All, as Are Sound Recordings from 1924!*, DUKE L.: CTR. & PROGRAM: CTR. FOR THE STUDY OF THE PUB. DOMAIN, https://web.law.duke.edu/cspd/publicdomainday/2025/#:~:text=On%20January%201%2C%202025%2C%20thousands,will%20now%20be%20public%20domain [https://perma.cc/PWL9-HHJJ].

182 *See* Cathay Y. N. Smith, *Editing Classic Books: A Threat to the Public Domain?*, 110 VA. L. REV. 1, 5–7 (2024).

183 Catherine Tucker, *Potential Socioeconomic Biases of AI Policy*, *in* IDENTIFYING THE ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE FOR COPYRIGHT POLICY: CONTEXT AND DIRECTION FOR ECONOMIC RESEARCH, U.S. COPYRIGHT OFFICE, 54, 55 (Brent A. Lutes ed., 2025), https://www.copyright.gov/economic-research/economic-implications-of-ai/Identifying-the-Economic-Implications-of-Artificial-Intelligence-for-Copyright-Policy-FINAL.pdf [https://perma.cc/9WUW-22AP].

184 *Id.*

185 *See Public Domain and Creative Commons: A Guide to Works You Can Use Freely*, UNIV. OF MONT. MAUREEN & MIKE MANSFIELD LIBR. (Jan. 8, 2025, at 16:31 PT), https://libguides.lib.umt.edu/PublicDomainCC [https://perma.cc/6Q3K-XN6X].

186 Richard A. Posner, *Transaction Costs and Antitrust Concerns in the Licensing of Intellectual Property*, 4 J. MARSHALL REV. INTELL. PROP. L. 325, 325 (2005).

happen.[187] Therefore, as more individual owners become involved in the licensing negotiations, transaction costs increase because more labor and money is needed to obtain all the necessary permissions.[188] This economic reality incentivizes developers to negotiate with large, consolidated content providers rather than engaging with a fragmented array of rights holders.[189] This preference for efficiency reinforces the overrepresentation of dominant, corporate-controlled content while sidelining alternative perspectives. For example, corporate news outlets often set the agenda for public discourse,[190] meaning that AI models trained primarily on their content may be more likely to reproduce viewpoints that align with mainstream, profit-driven media rather than those emerging from independent investigative journalism, activist publications, or community-driven storytelling. As a result, AI-generated knowledge may lack the diversity of thought necessary to present a well-rounded, multifaceted perspective on complex issues.

Moreover, given that AI developers are more likely to license works from major corporations, and because those news corporations generally engage in agenda setting,[191] there is a potential for political bias. This bias may emerge in two ways: first, through the exclusion of radical or nonmainstream sources from training data; and second, through the prioritization of political works that economically or politically benefit the AI company. Existing algorithms used by website operators and social media platforms, for instance, are often designed to proactively block or remove content deemed politically sensitive, dangerous, or untrustworthy.[192] While filtering certain types of content—such as explicit hate speech—may be widely accepted as necessary, the implementation of policy-directed filtering

---

187 *Id.* (e.g., the costs of finding the owner of the material, drafting contracts, paying legal counsel, etc.).

188 Tucker, *supra* note 183, at 57.

189 *Id.* at 56−57 ("Reddit . . . has announced a deal to license its users' postings to developers. However, such forums are poor representations of society at large . . . . Reddit is only used by 11 percent of U.S. adults. Of those, around two-thirds are men." (internal citation omitted)).

190 *See* Gary King et al., *How The News Media Activate Public Expression and Influence National Agendas*, 358 SCIENCE 776 (2017); Maxwell E. McCombs & Donald L. Shaw, *The Agenda-Setting Function of Mass Media*, 36 THE PUB. OPINION Q. 176, 177 (1972), https://www.jstor.org/stable/2747787 ("[T]he press 'may not be successful much of the time in telling people what to think, but it is stunningly successful in telling its readers what to think *about*.'").

191 King, *supra* note 190.

192 Peters, *supra* note 179, at 5.

algorithms may "violate legitimate user expectations of policy-neutrality" and contribute to ideological discrimination by disproportionately suppressing certain perspectives.[193] Taken to the extreme, such practices may hinder the free marketplace of ideas, thereby restricting public debate, limiting open dialogue, and enabling de facto censorship.[194] Additionally, while it is implausible—if not impossible—to capture every conceivable political work from across the ideological spectrum, the discretion AI developers exercise in determining which political materials are *relevant* enough to warrant the financial and logistical effort of licensing is problematic. This selection process effectively functions as a form of self-censorship, wherein AI-generated content disproportionately aligns with dominant ideological frameworks rather than presenting a truly comprehensive understanding of political issues.[195] The exclusion of diverse political viewpoints from AI training datasets ultimately risks reinforcing existing power structures and narrowing the range of perspectives available in public discourse.

Similarly, AI developers may prioritize obtaining licenses from established publishing institutions over less conventional sources. This preference is, in part, justified by the perception that such institutions transmit objective, factual information curated by credentialed professionals.[196] However, this assumption overlooks the historical association of mainstream media with political and economic interests.[197] Since at least the nineteenth century, newspapers and other major media outlets have frequently aligned themselves with specific political parties and economic power structures, shaping public discourse in ways that reflect their affiliations rather than practicing purely neutral reporting.[198] The historical bias in media is further compounded by the fact that dominant publishing institutions have traditionally been owned and operated by elite groups with vested

193 *Id.*

194 G. Michael Parsons, *Fighting for Attention: Democracy, Free Speech, and the Marketplace of Ideas*, 104 MINN. L. REV. 2157, 2158−59 (2020).

195 Anderson, *supra* note 177 ("We want our model to understand lots of forms and styles of language, so it's important that we include a very wide variety of language examples in our training data. If a model has never seen two humans arguing and insulting each other, it cannot understand such. Similarly, a model that's not seen street language and swearing will struggle to understand it. That's why the training data for an LLM needs to include data that might raise an eyebrow or two. If it doesn't, our model will be naive.").

196 Bruce Thornton, *A Brief History of Media Bias*, HOOVER INST. (June 12, 2023), https://www.hoover.org/research/brief-history-media-bias [https://perma.cc/WB6S-9XGE].

197 *Id.*

198 *Id.*

interests in maintaining prevailing power structures.[199] Meanwhile, alternative and grassroots publications—often serving as critical counterpoints to mainstream narratives—have historically struggled to achieve the same levels of influence and legitimacy.[200] Therefore, if AI models disproportionately train on sources that have historically filtered public knowledge through elite-controlled lenses, they may inadvertently reinforce systemic exclusions and fail to provide a truly comprehensive or equitable representation of knowledge.

Ultimately, the biases embedded in licensing decisions have significant consequences for the reliability, inclusivity, and fairness of AI-generated knowledge. When certain voices and perspectives are excluded from training datasets, AI models generate content that reflects and reinforces these exclusions. This can result in skewed, incomplete, and misleading AI-generated information; marginalization of underrepresented communities; reinforcement of existing media and academic biases; and overarching fairness and accountability concerns. Hence, bias in AI training data is not limited to the content that is excluded; it also exists in the choices made during the licensing process. When AI developers prioritize certain sources over others based on economic, political, or systemic biases, they shape the perspectives that AI models ultimately produce. Addressing these biases requires a commitment to transparency, inclusivity, and ethical responsibility in AI development.

B. The Antithesis of the Copyright Clause

If OpenAI successfully establishes a fair use defense for its GPT training, the resulting precedent could undermine the core purposes of copyright law—encouraging creativity and ensuring economic incentives for authors—by allowing AI companies to exploit copyrighted works without paying compensation.[201] By granting

---

199 Kate Vinton, *These 15 Billionaires Own America's News Media Companies*, FORBES (Jun. 1, 2016, at 14:26 ET), https://www.forbes.com/sites/katevinton/2016/06/01/these-15-billionaires-own-americas-news-media-companies/ [https://perma.cc/2VHJ-RB8D] (The Murdoch family "controls 120 newspapers across five countries" and the Cox family owns "more than a dozen non-daily publications, 14 broadcast television stations, one local cable channel and 59 radio stations.").

200 Karoline Andrea Ihlebæk et al., *Understanding Alternative News Media and Its Contribution to Diversity*, 10 DIGIT. JOURNALISM, 1267, 1268−69 (2022), https://www.tandfonline.com/doi/full/10.1080/21670811.2022.2134165?scroll=top&needAccess=true#abstract [https://perma.cc/8HLJ-DGFR].

201 U.S. CONST. art. I, § 8, cl. 8.

creators exclusive rights over their works, copyright law fosters investment in intellectual labor, enabling authors, journalists, and publishers to sustain their livelihoods.[202] However, if courts accept a broad fair use defense for AI training, this foundational purpose could be undermined. AI companies, such as OpenAI, profit from models trained on copyrighted materials without permission or compensation.[203] This practice threatens the economic incentives that sustain creative industries and commercially exploits the labor of authors. If AI-generated content can directly compete with human-created works, demand for original literature, journalism, and other forms of creative expression will decline, leading to fewer new works and a diminished cultural landscape.[204]

By providing temporary monopolies over creative works, copyright law ensures that authors are rewarded for their contributions, thereby encouraging the continued production of literature, journalism, music, and other forms of expression.[205] Without strong copyright protections, authors may lose the motivation to create new works, knowing that their content could be freely exploited without compensation.[206] The exclusivity granted by copyright law fosters a balanced and sustainable knowledge environment—one in which creators can earn a livelihood while contributing to cultural and intellectual development.[207] A broad fair use defense for AI training undermines this principle by enabling AI firms to appropriate copyrighted materials without permission, stripping authors of their economic rights, and weakening

---

202 Mei-Ian Stark, *5 Ways Copyright Laws Encourage Personal Expression and Creativity*, U.S. CHAMBER OF COM. (Apr. 25, 2022), https://www.uschamber.com/intellectual-property/five-ways-copyright-laws-encourage-personal-expression-and-creativity [https://perma.cc/DB98-ENP8].

203 Milmo, *supra* note 14.

204 *See* New York Times Complaint, *supra* note 4, at 14.

205 *Copyright Basics*, U.S. PAT. & TRADEMARK OFF., https://www.uspto.gov/ip-policy/copyright-policy/copyright-basics#:~:text=The%20primary%20purpose%20behind%20copyright,them%20available%20in%20the%20marketplace [https://perma.cc/ZV9U-L5TD] ("By granting authors the exclusive right to authorize certain uses of their works, copyright provides economic incentives to create new works and to make them available in the marketplace").

206 *See* Celeste Shen, *Fair Use, Licensing, and Authors' Rights in the Age of Generative AI*, 22 NW. J. TECH. & INTELL. PROP. 157, 177 (2024); *see also* Harry Jiang et al., *AI Art and Its Impact on Artists*, AIES 363, 368 (2023), https://doi.org/10.1145/3600211.3604681 (on file with the Oregon Law Review) (In the context of visual arts, "students who foresee image generators replacing artists have become demoralized and dissuaded from honing their craft and developing their style.").

207 *See* Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith, 598 U.S. 508, 526 (2023).

the incentives necessary for creative production. Additionally, large technology firms, which already control vast amounts of digital content,[208] stand to further consolidate their power by freely using copyrighted works without compensating creators. This creates an imbalance, in which AI companies profit from content they did not create, while smaller publishers and independent authors struggle to compete.

Moreover, copyright protections ensure that authors and publishers are compensated through book sales, licensing fees, and royalties.[209] However, if AI models are allowed to freely train on copyrighted materials, they can generate content that competes directly with human-created works.[210] This could significantly reduce demand for original literature and journalism.[211] Already, some authors and media outlets have reported declining revenues as AI-generated summaries and articles divert traffic from their platforms.[212] Without financial incentives, many writers may find it unsustainable to continue producing new works, leading to a decline in creative output.

Ultimately, when authors see their works freely exploited by AI companies without consent, they may lose the motivation to create.[213] Fear that original works will be absorbed into AI training datasets without compensation or recognition discourages creative risk-taking and innovation.[214] AI-generated works are inherently derivative, as they rely on existing copyrighted materials to produce new content.[215] Consequently, if AI-generated literature and journalism become the widespread dominant form of media, the

---

208 Shivam Rapjut, *How Does Big Tech Dominate the Digital World?*, GEOSTRATA (Sept. 30, 2024), https://www.thegeostrata.com/post/big-tech-tightens-its-grip-how-companies-like-apple-and-google-dominate-the-digital-world [https://perma.cc/XK39-3MJ3] (For example, "Meta Inc. is a leader in social media and communication . . . with over 3 billion active users.").

209 17 U.S.C. § 106.

210 *See* New York Times Complaint, *supra* note 4, at 14.

211 *See id.* at 14–15.

212 Wendy Lee, *Get Paid or Sue? How the News Business Is Combating the Threat of AI*, L.A. TIMES (July 24, 2024, at 03:00 PT), https://www.latimes.com/entertainment-arts/business/story/2024-07-24/how-ai-is-disrupting-the-journalism-industry [https://perma.cc/72XK-MAZ6].

213 *Copyright Basics*, *supra* note 205.

214 *See* Jiang et al., *supra* note 206, at 368 ("[B]oth new and current artists are becoming increasingly reluctant to share their works and perspectives, in an attempt to protect themselves from the mass scraping and training of their life's works."); *see also Copyright Basics*, *supra* note 205.

215 *See supra* text accompanying notes 31–33.

diversity of artistic expression may suffer. Instead of fostering new ideas and voices, AI-generated content could reinforce dominant narratives, leading to a homogenization of creative output.[216]

Further, artists may hesitate to share their work and mentor aspiring creators, which not only stifles individual artistic growth but also weakens the collective creativity of humanity.[217] Artistic innovation has long relied on a rich tradition of learning, adaptation, and reinterpretation, where new generations of creators draw inspiration from those who came before them, building upon past techniques to develop novel expressions.[218] When this cycle is disrupted by the fear of exploitation, fewer artists may engage in this exchange of knowledge, resulting in a narrower landscape of creativity. This phenomenon mirrors a troubling feedback loop emerging in AI development, particularly in the training of LLMs.[219] When AI models are trained on the outputs of previous models rather than on a diverse corpus of human-created content, they risk generating increasingly homogenized and inaccurate results.[220] Each successive generation becomes more detached from the originality, complexity, and novelty of human expression, and instead amplifies the patterns, biases, and inherent limitations from earlier iterations of AI models.[221] This self-referential process resembles an ouroboros—a serpent consuming its own tail—where AI-generated content continuously feeds into new models, creating a cycle in which nothing truly original is produced.[222]

---

216 *See* Anil R. Doshi & Oliver P. Hauser, *Generative AI Enhances Individual Creativity but Reduces the Collective Diversity of Novel Content*, NAT'L LIBR. MED. (July 12, 2024), https://pmc.ncbi.nlm.nih.gov/articles/PMC11244532/ [https://perma.cc/QXD5-XUEQ].

217 Jiang et al., *supra* note 206, at 368.

218 *See, e.g.*, Daniela Lehner, *The Heroine's/Hero's Journey—A Call for Transformation? Transformative Learning, Archetypal Patterns, and Embodied Knowing/ Learning*, 20 J. TRANSFORMATIVE EDUC. 88, 89 (2022), https://journals.sagepub.com /doi/10.1177/15413446211007089?icid=int.sj-abstract.citing-articles.21 (on file with the Oregon Law Review) (explaining the "Heroine/Hero's Journey" is a common storytelling trope that can be traced through "myths and stories from different times and places").

219 Ilia Shumailov et al., *AI Models Collapse When Trained on Recursively Generated Data*, 631 NATURE, 755, 755 (2024), https://www.nature.com/articles/s41586-024-07566-y [https://perma.cc/V3PF-DR8F] ("Model collapse is a degenerative process affecting generations of learned generative models, in which the data they generate end up polluting the training set of the next generation. Being trained on polluted data, they then mis-perceive reality.").

220 Keyur Ramoliya, *Model Collapse in AI*, MEDIUM (Aug. 2, 2024), https://medium .com/thedeephub/model-collapse-in-ai-813418fd8516 [https://perma.cc/7VDU-59SH].

221 *Id.* (As the feedback loop continues, "each new AI model learns more from AI-created content and less from human-created content . . . ." which "can lead to a gradual degradation in the quality and diversity of AI-generated content.").

222 Jiang et al., *supra* note 206, at 368.

If humanity comes to rely exclusively on AI-generated works for "the words we read, the art we see," and the music we hear, our cultural output may become an echo chamber of past creations rather than a space for genuine innovation.[223] Instead of pushing artistic and intellectual boundaries, we risk a future where creativity is reduced to an automated regurgitation of the *familiar*, rather than nurturing the expansion of dynamic creative expression.

C. More Than Just a © Symbol

A verdict favoring *The New York Times* aligns squarely with the foundational purpose of the Copyright Clause of the U.S. Constitution: "[t]o promote the Progress of Science and useful Arts."[224] The Copyright Clause, therefore, reflects a deliberate constitutional choice to empower creators with exclusive rights as a means to incentivize the continuous production of original works, thereby fostering a flourishing, innovative society.[225] The quid pro quo nature of copyright law—providing a time-limited monopoly in exchange for public benefit—depends on the enforceability of copyright protections; even, and especially, in the face of transformative technological shifts like generative AI.[226] If AI developers are permitted to harvest and commercially exploit copyrighted journalism without consent, compensation, or credit, the constitutional calculus underpinning copyright unravels.[227] Creators lose the incentive to invest the time, labor, and resources needed for high-quality content, and it is the public who ultimately suffers from a diminished cultural and informational landscape.[228]

Although enforcing copyright in the AI training context could contribute to algorithmic bias or selective representation,[229] these concerns cannot justify eroding the fundamental principles of copyright law. The Copyright Clause was never meant to guarantee equal data

---

223 *Id.*

224 U.S. CONST. art. I, § 8, cl. 8.

225 *Artl.S8.C8.1 Overview of Congress's Power Over Intellectual Property*, CONST. ANNOTATED, https://constitution.congress.gov/browse/essay/artI-S8-C8-1/ALDE_00013060/ (on file with the Oregon Law Review) (last visited Oct. 6, 2025).

226 *Id.* ("Without legal protection, competitors could freely copy such creations, denying the original creators the ability to recoup their investments in time and effort, reducing the incentive to create in the first place.").

227 *Id.*

228 *Id.*

229 *See supra* text accompanying notes 175–98.

access to private companies building profitable AI systems; rather, it is a guarantee of exclusive rights for authors, inventors, and other creators.[230] Moreover, the risks of bias or content gatekeeping arise not from respecting those rights, but from deeper issues like media monopolization and technology governance.[231] Consequently, honoring the *Times*'s exclusive right to its work in the AI context affirms a constitutional commitment to incentivize human creativity and compensate the labor of creatives, irrespective of the potential policy consequences.

CONCLUSION

The *Times*'s lawsuit against OpenAI raises fundamental questions about the intersection of copyright law, fair use, and the ethical responsibilities of AI developers. At the heart of this dispute lies the challenge of balancing AI innovation with the legal and economic protections afforded to authors, journalists, and other creators. Whereas OpenAI asserts that its use of copyrighted materials for training LLMs constitutes fair use,[232] the *Times* contends that such practices undermine the economic incentives that sustain creative industries.[233]

A fair use analysis of OpenAI's practices suggest that although AI training may satisfy certain fair use factors, the commercial nature of OpenAI's operations and the potential market harm to original creators' work weigh against a broad fair use defense.[234] The nature of AI-generated content—derivative yet transformative—complicates fair use jurisprudence, highlighting the need for courts to consider the unique attributes of machine learning models in assessing copyright claims.[235]

Beyond the legal arguments, this case carries significant implications for the broader media and technology landscape. If the *Times* prevails and courts mandate the destruction of OpenAI's training datasets, AI developers may be forced to reconstruct their models using either public domain works or selectively licensed materials.[236] This shift risks introducing new biases into AI-generated content, as reliance on public domain works skews toward older, historically privileged

230 *See* U.S. CONST. art. I, § 8, cl. 8.

231 *See supra* text accompanying notes 187–98.

232 Defendant's Memorandum in Support of Motion to Dismiss, *supra* note 9, at 3.

233 New York Times Complaint, *supra* note 4, at 14.

234 *See supra* Section IV.B.

235 *Id.*

236 *See supra* text accompanying notes 175–98.

perspectives, while selective licensing favors dominant, corporate-controlled narratives.[237] The economic and political consequences of algorithmic gatekeeping could further entrench existing power structures, limiting the diversity of perspectives represented in AI-generated knowledge.[238]

Conversely, if OpenAI's fair use defense is upheld, it could weaken copyright protections that ensure economic incentives for creators. Allowing AI companies to profit from models trained using copyrighted works without compensation or credit may undermine the sustainability of journalism, literature, and other creative industries.[239] Without adequate safeguards, AI-generated content could displace human authors, thereby reducing the financial viability of creative professions and diminishing the diversity of cultural and intellectual discourse more broadly.[240]

Ultimately, this case underscores the urgent need for legal frameworks that address the evolving relationship between AI and copyright law. This tension between technological advancement and the protection of creators' rights is not a new phenomenon.[241] Throughout history, groundbreaking innovations have disrupted existing creative industries, promoting legal and ethical debates about ownership, authorship, and fair use. For example, nineteenth-century painters and illustrators resisted the advent of the photograph, as they feared that mechanically produced images would devalue traditional artistic skills.[242] Similarly, the rise of the internet and digital media in the late twentieth and early twenty-first centuries led to widespread concerns over copyright infringement, as file-sharing platforms and

---

237 *Id.*

238 *Id.*

239 *See supra* Section V.B.

240 *Id.*

241 Eva Silva, *How Photography Pioneered a New Understanding of Art*, COLLECTOR (June 4, 2022), https://www.thecollector.com/how-photography-transformed-art/ [https://perma.cc/CVB5-79NE]. *See generally* Jean K. Chalaby, *The Streaming Industry and the Platform Economy: An Analysis*, 46 MEDIA CULTURE & SOC'Y 552, 554, https://journals.sagepub.com/doi/10.1177/01634437231210439#:~:text=SVoD%20services%20can%20acquire%20the,Turton%20and%20Opie%2C%202019 (on file with the Oregon Law Review); *The Evolution and Impact of Spotify on the Music Industry*, MEDIUM: DIGITAL MIRAI (Jul. 15, 2024), https://digitalmirai.medium.com/the-evolution-and-impact-of-spotify-on-the-music-industry-49a140cf9735 [https://perma.cc/3MH2-4WWF] ("Spotify has democratized music distribution, allowing independent artists to reach global audiences without the need for traditional record deals.").

242 Silva, *supra* note 241.

streaming services reshaped the ways creative works were distributed and monetized.[243] Most recently, the popular music platform Spotify disrupted the music industry's strict control over music distribution by making music widely accessible through a digital platform.[244] In all cases, legal frameworks and industry practices ultimately evolved to balance technological progress with the rights of creators. The debate over AI-generated content is just the latest iteration of this ongoing struggle, raising a fundamental question about how to foster innovation without undermining the very human creativity that fuels it.

243 *See generally* Chalaby, *supra* note 241.

244 *The Evolution and Impact of Spotify on the Music Industry*, *supra* note 241.