LATHAM & WATKINS LLP
Joseph R. Wetzel (SBN 238008)
joe.wetzel@lw.com
Andrew M. Gass (SBN 259694)
andrew.gass@lw.com
Brittany N. Lovejoy (SBN 286813)
britt.lovejoy@lw.com
505 Montgomery Street, Suite 2000
San Francisco, California 94111
Telephone: +1.415.391.0600

Sarang V. Damle (*pro hac vice*)
sy.damle@lw.com
555 Eleventh Street NW, Suite 1000
Washington, D.C. 20004
Telephone: +1.202.637.2200

Allison L. Stillman (*pro hac vice*)
alli.stillman@lw.com
1271 Avenue of the Americas
New York, New York 10020
Telephone: +1.212.906.1747

Attorneys for Defendant Anthropic PBC

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., ET AL., *Plaintiffs*, v. ANTHROPIC PBC, *Defendant*.	Case No. 5:24-cv-03811-EKL **DECLARATION OF STEVEN R. PETERSON, PHD IN SUPPORT OF DEFENDANT'S OPPOSITION TO PLAINTIFFS' RENEWED MOTION FOR PRELIMINARY INJUNCTION** Hon. Eumi K. Lee **REDACTED**

TABLE OF CONTENTS

I. QUALIFICATIONS AND ASSIGNMENT 1

A. Qualifications 1

B. Assignment and Materials Used 1

II. SUMMARY OF OPINIONS 2

III. ANTHROPIC AND THE CLAUDE MODEL 5

IV. THERE IS NOT AN INCIPIENT COMPETITIVE MARKET TO LICENSE TEXT TO TRAIN LLMS 7

A. The Hypothetical Competitive Market to License Training Data 9

1. Training Requires Large Amounts of Data Relative to the Size of Even a Large Publisher's Archive 10

2. The Availability of Reliable Information in the Hypothetical Competitive Market for Training Data Licenses Is Poor 12

3. License Rates for Training Data Would Be Low 14

4. Negotiation and Transaction Costs in the Hypothetical Market Would Be High and Would Be Elevated by Uncertainty Regarding the Value of Licenses 15

B. The Agreements Plaintiffs Have Identified Do Not Indicate that There Is an Incipient Market to License Training Data for LLMs 17

C. Using Song Lyrics to Train LLMs Does Not Threaten Plaintiffs' Existing Licenses 23

I. QUALIFICATIONS AND ASSIGNMENT

A. Qualifications

1. My name is Steven R. Peterson. I am an Executive Vice President at Compass Lexecon. Compass Lexecon is an economics consulting firm that specializes in the economics of competition, finance, and regulation, among other areas. I received my A.B. in economics from the University of California, Davis, in 1987 and my Ph.D. in economics from Harvard University in 1992. While at Harvard, my areas of specialization were economic theory and industrial organization. Industrial organization is the study of the interactions of larger firms that can strategically influence the markets in which they operate. I have also served as an adjunct faculty member in the Department of Economics at Northeastern University where I taught courses on the economics of antitrust, regulation, and public policy.

2. I have consulted on the economics of antitrust and competition, mergers, estimation of damages, valuation, regulation, and public policy. I have extensive experience with the policy and economics of copyrights. I testified on behalf of the Radio Music License Committee ("RMLC") on issues of market definition and market power in *Radio Music License Committee v. SESAC, Inc., SESAC, LLC, and SESAC Holdings, Inc.*, and I testified on public performance license rates for radio in the subsequent arbitration proceedings. I addressed similar issues in RMLC's litigations against Global Music Rights ("GMR") and Broadcast Music, Inc. I also testified before the United States Copyright Royalty Judges, Library of Congress, regarding appropriate rates for digital public performances of sound recordings in the *Web IV* (on behalf of the National Association of Broadcasters) and *Web V* (on behalf of Google) proceedings. I submitted comments on behalf of the National Association of Broadcasters addressing the Department of Justice's review of the American Society of Composers, Authors and Publishers ("ASCAP") and Broadcast Music, Inc. ("BMI") consent decrees.[1] A true and correct copy of my curriculum vitae is attached as Exhibit A.

B. Assignment and Materials Used

[1] ASCAP, BMI, SESAC, and GMR are performing rights organizations ("PROs") that aggregate and license public performance rights in music compositions controlled by publishers like the plaintiffs in this proceeding.

3. Counsel for Anthropic PBC has asked me to evaluate (1) whether there is an existing practicable market for licensing copyrighted materials used to train large language models ("LLMs") like Anthropic's Claude and (2) the effects of using Plaintiffs' song lyrics as training data for Anthropic's large language model ("LLM") on the potential market to license text and/or song lyrics to train LLMs. My analysis focuses on the use of song lyrics as inputs for training LLMs to generate new text. I understand the generation of new text to be the primary and prevalent use of LLMs like Anthropic's signature product, Claude.

4. To perform my analysis, I have reviewed materials provided in discovery to date,[2] Plaintiffs' materials submitted in connection with their original preliminary injunction motion and their renewed motion, briefs submitted by *amicus curiae* in connection with this and Plaintiffs' original motion, written testimony from Anthropic's Jared Kaplan, academic articles, and other public information.

II. SUMMARY OF OPINIONS

5. I previously submitted a declaration in support of Anthropic's opposition to Plaintiffs' original preliminary injunction motion on January 16, 2024. After reviewing Plaintiffs' and their *amici*'s responses to that declaration, my core opinions remain unchanged.

6. There is no current practicable market whereby an LLM provider like Anthropic has or could train a general, cutting-edge LLM like Claude using exclusively licensed or public domain materials. The limited examples of artificial intelligence models built exclusively on licensed or public domain materials cannot perform the same tasks as leading LLMs and, therefore, do not compete directly with state-of-the-art general AI LLMs in the marketplace like Claude, OpenAI's ChatGPT, Meta's LLaMA, Mistral, and others.

7. The claim that there is an incipient market for training licenses implicitly assumes that such a market is economically viable if licenses were necessary to use data for training. A viable market needs both buyers and sellers. Therefore, one requirement of a viable market to license training data is that the AI firms seeking licenses could remain viable to participate in

[2] I understand that discovery is ongoing. I will update my analysis and opinions in this declaration as necessary to reflect my ongoing review of Plaintiffs' production.

that market. To remain financially and legally viable in a world where training data must be licensed, an AI firm exercising its best efforts to license training data would need to be able to acquire the needed data for a cutting-edge LLM with sufficient certainty to avoid copyright infringement. Economic analysis shows that the hypothetical competitive market for licenses covering data to train cutting-edge LLMs would be rendered impracticable as a method of acquiring licensing data by the low value of transactions, poor information, and high transaction costs. This conclusion is based on the following subordinate conclusions.

a) In a competitive market for licenses to train LLMs, the highest price that an artificial intelligence ("AI") company would pay is the increase in value of the model created by the addition of the licensor's content to the training dataset.

b) Incremental contributions to the value of the LLM by individual copyright owners would be small and would grow smaller as the training dataset grows.

c) The rates for licenses for data to train LLMs would be low in a hypothetical competitive market.

d) To successfully license data to train a cutting-edge LLM, AI firms would need to enter agreements with at least tens of thousands of publishers and potentially millions of copyright owners of material on the internet. The negotiation and administration costs of acquiring licenses in a hypothetical competitive market for licenses covering training data would be high relative to the value of licensors' individual contributions.

e) Each party's uncertainty regarding the value of a license for training data to the other will contribute to the failure of some negotiations, raising negotiating costs.

8. Information on copyrights and their owners is imperfect, making it costly or impossible to identify copyright owners in order to obtain licenses. In a world in which training data must be licensed, AI firms would be at risk of costly litigation for the training use of copyrighted works as the result of the inability to obtain licenses even with diligent effort.

9. Economists typically *assume* that the markets needed to affect desired transactions exist, but this assumption sometimes fails in practice. The reason that the

assumption fails is obvious and unsurprising. Clearly, potential markets with costs that are greater than their benefits are not viable and will not exist.[3] The result is that the benefits that the market would yield, if it existed, are unrealized in the competitive marketplace. Thus, while often overlooked, the problem of missing (incomplete) markets is a recognized form of market failure.[4]

10. In short, the claim that a market involving many thousands or millions of individual transactions to license huge volumes of copyrighted text covered by millions of copyrights would fail is unremarkable as a matter of economics. The quantity of information required to train cutting-edge LLMs is vast, and the practical difficulties of the mass licensing to acquire sufficient training data in a competitive market for a cutting-edge LLM has been found impracticable.[5] As a result, in a world in which AI firms had to license training data in a competitive marketplace, they would not be able to do so and would not exist. The implication is that one of the counterparties in the market that Plaintiffs envision would not be viable and would not be present to engage in the transactions Plaintiffs imagine would occur.

11. The content agreements that Plaintiffs' expert Professor Smith cites do not support the conclusion that there is an incipient licensing market for training data. Even when taken as presented by Professor Smith, the handful of agreements that appear to cover rights

---

[3] Dennis Carlton, "The Theory of Allocation and Its Implications for Marketing and Industrial Structure: Why Rationing is Efficient," *The Journal of Law & Economics*, October 1991, 34(2):231-262, p. 233 ("Aside from the fact that there are costs to setting up such markets, it is also true that many markets, once created, fail. Roughly one-half of all successful futures markets fail within ten years of their introduction. Since these markets clearly produce at least some benefits to their creators, the high failure rate must indicate the presence of significant costs of operation that outweigh these benefits. Furthermore, only a small fraction of products have ever had futures markets. Again, this emphasizes that there must be significant costs to the creation of these markets, otherwise the benefits flowing from them would suggest a proliferation of such markets.").

[4] Bernard Salanié, *Microeconomics of Market Failures*, The MIT Press, 2000, Chapter 13 and *see* p. 203 ("The theory of general equilibrium proceeded in the 1960s on the hypothesis that all markets necessary to the successful working of the economy could be opened. It escaped no one that this hypothesis was a bit heroic, that its negation would generally destroy the optimality of the equilibrium, but the technical tools that would permit one to treat the incompleteness of markets did not yet exist.").

[5] "Orphan Works and Mass Digitization," United States Copyright Office, June 2015, p. 1, available at https://www.copyright.gov/orphan/reports/orphan-works2015.pdf (accessed 8/21/2024).

much broader than training do not support the economic conclusion that AI firms are moving toward licensing, or could license, all of their training data. Moreover, a party to many of the agreements Professor Smith cites, OpenAI, continues to maintain in court filings that using copyrighted text as training is fair use, which indicates that there may be other reasons for entering those agreements.[6]

12. Similarly, the licensing of Plaintiffs' musical compositions by music streaming services, for example, does not demonstrate the viability of a market to license copyrighted works for training LLMs. The licensing of musical works is not analogous to licensing training data for LLMs. Music licensing is simplified by a combination of compulsory licenses and regulated rates for those licenses, and training datasets require greater diversity of works than streaming services. Moreover, the scale and complexity of licensing diverse training data is greater than found in music licensing and licensing efforts are hampered by incomplete information regarding ownership of copyrights.

13. The use of copyrighted materials to train general LLMs does not supplant or harm any existing markets for the works used to train the model. General intelligence LLMs like Claude are a new category of product that do not compete, and are not intended to compete, with the products and services that traditionally require licenses to musical works, such as sound recordings, music streaming services, karaoke services, film and television, or lyric websites.

III. ANTHROPIC AND THE CLAUDE MODEL

14. Anthropic is an AI company that develops AI models known as LLMs.[7] LLMs process large amounts of text in order to "understand" how words work together and use that "learning" to produce text of its own, typically in response to a user-generated prompt. When developing an LLM, engineers first create a "neural network," a type of software that analyzes large datasets and extracts statistical information from language to better understand how

---

[6] I understand that the agreements that Professor Smith references are not available for review in this proceeding.

[7] *See* https://www.anthropic.com/company (accessed 8/21/2024) ("Anthropic is an AI safety and research company.") and ("Anthropic is a collaborative team of researchers, engineers, policy experts, business leaders and operators, who bring our experience from many different domains to our work.")

language works.[8] LLMs "learn" how to recognize, interpret, and generate language through a "training" process that enables them to identify relationships and patterns between words.[9] Anthropic's LLM, called "Claude," is trained on data compiled from the Internet, non-public datasets that are commercially obtained, data from users and service providers, and internally generated data.[10] Anthropic estimates that the current versions of Claude, known as Claude 2, Claude 3, and Claude 3.5, were trained on datasets of billions of texts. These texts contained roughly [redacted] tokens for Claude 2, [redacted] tokens for Claude 3, and [redacted] tokens for Claude 3.5 Sonnet.[11] Tokens are fragments of words in the texts and are used in the training process. [redacted] tokens are roughly equivalent to [redacted] words, and [redacted] tokens are roughly equivalent to [redacted] words.[12]

15. Training datasets are large because the goal of training is for the model to identify patterns and relationships between words. Therefore, training depends on the volume and diversity of text.[13] Testing shows that LLMs trained on larger volumes of data were more adept at generative tasks (such as creating new expressions).[14]

16. Once trained, LLMs are able to perform a range of natural language processing tasks.[15] According to Anthropic "Claude tends to perform well at general, open-ended conversation; search, writing, editing, outlining, and summarizing text; coding; and providing helpful advice about a broad range of subjects … *Claude is designed to serve as a creative*

---

[8] Declaration of Jared Kaplan in Support of Defendant's Opposition to Plaintiffs' Motion for Preliminary Injunction ("Kaplan Declaration"), ¶ 20.

[9] Kaplan Declaration, ¶ 20; "Public Comments of Anthropic PBC," Before the United States Copyright Office, Notification of Inquiry Regarding Artificial Intelligence and Copyright, October 30, 2023 ("Anthropic Comments"), p 6.

[10] Anthropic Comments, p. 5.

[11] Kaplan Declaration, ¶ 29.

[12] *See, e.g.*, "What are tokens and how to count them?" OpenAI, available at https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them (accessed 8/21/2024).

[13] Kaplan Declaration, ¶¶ 24-25.

[14] "Tracing Model Outputs to the Training Data," Anthropic, August 8, 2023 ("[P]atterns of generalization became more abstract with model scale.") available at https://www.anthropic.com/index/influence-functions (accessed 8/21/2024).

[15] *See* Kaplan Declaration, ¶¶ 11-13.

companion, to enable people to produce new works."[16] In addition, analysis of how Claude is used reveals that users rely on Claude to edit, rewrite, and brainstorm on original writing projects, such as novels and poems.[17] Notably, each of the use cases that Anthropic describes is *generative*—that is, Claude is not designed to simply regurgitate text or facts, but to use its "knowledge" to help users more efficiently access, understand, create, and convey information.[18]

IV. THERE IS NOT AN INCIPIENT COMPETITIVE MARKET TO LICENSE TEXT TO TRAIN LLMS

17. Plaintiffs allege that Anthropic's use of their text for training "will irreparably damage Publishers' competitive position and their ability to negotiate licenses with AI developers."[19] This claim presupposes the existence of a market for licenses covering text used to train LLMs. However, the claim that there is an incipient market for text to train LLMs is economically baseless without analysis of the characteristics of the *hypothetical competitive market* to license training material. The economics of the hypothetical market to license training text demonstrate that both the massive scale of the hypothetical market and the poor information surrounding copyright ownership would make it essentially impossible for AI firms to acquire the text required to train cutting-edge LLMs.

18. In order for a competitive market for training data to exist in a world in which AI firms had to license training data, both licensors and licensees must be viable when interacting through the hypothetical competitive licensing market – that is, AI firms that license training text diligently and in good faith through the hypothetical competitive licensing market would have to be both financially and legally viable. To be financially viable, AI firms would have to be able to license sufficient training data to build cutting-edge LLMs at total licensing costs that allow them to cover investments and ongoing research and development. In addition, an AI firm that

---

[16] *See also* Anthropic Comments, p. 2 (emphasis added); Anthropic Comments, pp. 3-4, describing examples of productivity-enhancing uses of Claude since its launch.

[17] Kaplan Declaration, ¶ 12.

[18] Kaplan Declaration, ¶¶ 21, 36.

[19] Plaintiffs' Notice of Motion and Motion for Preliminary Injunction; Memorandum of Law in Support of Plaintiffs' Motion for Preliminary Injunction, p. 27.

licensed diligently and in good faith in the competitive market would need to be able to acquire the quantities of text required to train the current and next generation of LLMs without using unlicensed copyrighted materials.

19. For a market of any kind to exist, the benefits of the market must more than cover its costs. Transaction costs preclude the establishment of many markets.[20] Organized markets, such as stock or commodity exchanges, require traders, market makers, clearinghouse employees, real estate to house workers, and so forth.[21] For markets that form organically, the operation of the market involves the meeting and interaction of buyers and sellers. Without an organized exchange, substantial resources may be spent identifying counterparties, and if the details of each transaction are unique, the parties may spend substantial resources negotiating over price and terms. Of course, if the value of the transaction is less than the cost of negotiating and completing it, the counterparties will not meet, and transactions will not take place.[22]

20. The economic analysis below shows that the hypothetical competitive market to license data for training would have a huge number of licenses between AI firms and copyright

---

[20] Dennis Carlton, "Contracts, Price Rigidity, and Market Equilibrium," *Journal of Political Economy*, October 1979, 87(5):1034-1062, p. 1035. Bernard Salanié, *Microeconomics of Market Failures*, The MIT Press, 2000, Chapter 13.

[21] Dennis Carlton, "The Theory of Allocation and Its Implications for Marketing and Industrial Structure: Why Rationing is Efficient," *The Journal of Law & Economics*, October 1991, 34(2):231-262, p. 233 ("One can learn a great deal about the cost of a price system by studying institutions whose business is to create such markets. For example, if one studies the Chicago Board of Trade, one observes the following: (1) a large building in an expensive part of town; (2) many people involved in each transaction (for example, brokers, floor traders, employees of a clearinghouse, consultants); (3) office buildings housing the people involved in each transaction; and (4) elaborate record keeping. Undoubtedly the greatest cost is the time cost of all the people involved. A significant fraction of the economy of the city of Chicago is devoted to the making of markets that clear by price. If a magic spell could be cast to make transactions costless, the Chicago economy would be devastated, at least in the short run. This emphasizes how far from costless the making of markets really is.").

[22] Dennis Carlton, "The Theory of Allocation and Its Implications for Marketing and Industrial Structure: Why Rationing is Efficient," *The Journal of Law & Economics*, October 1991, 34(2):231-262, p. 233 ("Aside from the fact that there are costs to setting up such markets, it is also true that many markets, once created, fail. Roughly one-half of all successful futures markets fail within ten years of their introduction. Since these markets clearly produce at least some benefits to their creators, the high failure rate must indicate the presence of significant costs of operation that outweigh these benefits. Furthermore, only a small fraction of products have ever had futures markets. Again, this emphasizes that there must be significant costs to the creation of these markets, otherwise the benefits flowing from them would suggest a proliferation of such markets.").

owners, such as book authors, journalists, and a wide variety of content publishers. To the extent individual posters on the Internet can assert copyright claims over their posts, the number of licenses needed could be in the tens or hundreds of millions.[23] The small contribution of each licensor's content to the value of the LLM would lead to low per-work license payments, even for relatively large licensors. The transaction costs to negotiate, track, and manage the required licenses would be high relative to their value, even when copyright owners are known. Moreover, information on copyright ownership is frequently poor or unavailable, making it essentially impossible to license a substantial share of works in a competitive market.

A. The Hypothetical Competitive Market to License Training Data

21. Training data is an input into the development of LLMs. In competitive markets for inputs, the price—or in this case the license rate—for each input is equal to the incremental value that it creates (*i.e.*, the marginal revenue product).[24] In the context of negotiations to license training data from copyright owners, this principle implies that the highest amount an AI firm would be willing to pay to add a copyright owner's data to its training database is the increase in the value of the model that results from the addition of that particular training data. In a viable, real-world market, licensing negotiations would settle on license rates below the AI firm's willingness to pay for the data and above the lowest amount that the licensor would accept to grant the license. In a practicable competitive market, each copyright owner must anticipate that it is licensing content to train an AI model that *will exist whether or not it licenses its content for training.* In this circumstance, each licensor's opportunity cost of granting the license is essentially zero because most uses of the model will be in areas unrelated to the licensed content and the contribution of the copyright owner's content to the model would have essentially no discernable effect on the operation of the model. Given licensees' low willingness to pay, however, licensors' transaction costs could exceed the competitive value of a license.

---

[23] *See, e.g.*, "Are Tweets Protected by Copyright?" Copyright Alliance, available at https://copyrightalliance.org/faqs/tweets-protected-copyright/ (accessed 8/21/2024).

[24] *See* N. Gregory Mankiw, *Principles of Microeconomics*, 7th ed., Cengage Learning, 2015, p. 377, ("Economists sometimes call this column of numbers the firm's *marginal revenue product*: It is the extra revenue the firm gets from hiring an additional unit of a factor of production.").

22. As discussed further below, a hypothetical market to license data for training LLMs would entail a large number of low-value negotiations between AI firms and copyright owners.

1. Training Requires Large Amounts of Data Relative to the Size of Even a Large Publisher's Archive

23. The hypothetical competitive market to license training data for a cutting-edge LLM would require tens of thousands of negotiations by each AI firm with large and small publishers – and potentially millions of negotiations with individual owners of material on the Internet.

24. For example, Claude 3.5 Sonnet was trained on approximately [redacted].[25] AI firms continue to search for more data to train the next generations of LLMs. As a result, the size of the training databases for future LLMs will be larger, with larger licensing requirements according to Plaintiffs.

25. The size of Anthropic's training dataset for Claude 3.5 Sonnet is significantly larger than the archives of even the largest publishers.

a) **NYT Archive.** If the *New York Times* averages, say, 150,000 words each day, then a 100-year archive of the *Times* would contain just under 5.5 billion words – less than [redacted] of the words used to train the current version of Claude. Put differently, more than [redacted] archives of the size of the *Times* would be needed to compile a text dataset equal to the size of the Claude training dataset, and still more would be needed to train the next version of Claude.

b) **Song Publishers.** The larger song publishers among the Plaintiffs claim to have between one million and 5.5 million songs in their catalogs.[26] Other song

---

[25] Kaplan Declaration, ¶ 29; "What are tokens and how to count them?" OpenAI, available at https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them (accessed 8/21/2024). For reference, one trillion is equal to one million multiplied by one million, or 10 raised to the power of 12 (*i.e.*, 10^{12}).

[26] "The Three Major Music Publishers Now Own or Control Over 10 million Songs between Them (Kind of)," Music Business Worldwide, October 20, 2022, available at https://www.musicbusinessworldwide.com/the-three-major-music-publishers-now-own-or-

publishers presumably have fewer songs in their libraries. If a typical song has 300 words,[27] then a large music publisher with a one-million song library could contribute 300 million words, or [redacted] of the database used to train the current version of Claude (assuming it owns the rights to enter such a license). More than [redacted] song publishers with one-million song libraries would be needed to create a database of the size used to train the current version of Claude. The administrative costs of licensing could be yet more costly as the result of many songs' copyrights being fractionally owned by multiple writers, publishers, or investors.

c) **All the Books in the World.** Google estimated in 2010 that there were approximately 130 million books in the world.[28] These books may not all have fully unique text – that is, some may be different copies of the same play with different commentary, for example. However, if each book were unique and had on average 100,000 words, roughly the number of words in a typical modern novel, all the books in the world would contain 13 trillion words. [redacted].[29]

---

control-over-10-million-songs-between-them-kind-of/ (accessed 8/21/2024). ("The 'kind of' disclaimer in our headline above is in there because these are the separate figures reported by each of the major music companies and their parent companies. What these numbers don't tell us is how many of these songs are *duplicates* i.e. [*sic*] are administered by two or even all three of the major publishers, due to them having multiple writers represented by different companies.").

[27] *See, e.g.*, William Wier, "Words Words Words, Are excessive lyrics ruining pop music?" Slate, March 11, 2008, available at https://slate.com/culture/2008/03/are-excessive-lyrics-ruining-pop-music.html (accessed 8/21/2024).

[28] Leonid Taycher, "Books of the world, stand up and be counted! All 129,864,880 of you," Google Book Search, August 5, 2010, available at https://booksearch.blogspot.com/2010/08/books-of-world-stand-up-and-be-counted.html (accessed 8/21/2024) (Discussing the difficulty of identifying unique books, "One definition of a book we find helpful inside Google when handling book metadata is a 'tome,' an idealized bound volume. A tome can have millions of copies (e.g., a particular edition of 'Angels and Demons' by Dan Brown) or can exist in just one or two copies (such as an obscure master's thesis languishing in a university library). This is a convenient definition to work with, but it has drawbacks. For example, we count hardcover and paperback books produced from the same text twice, but treat several pamphlets bound together by a library as a single book.").

[29] *See* Kaplan Declaration, ¶ 29.

26. These calculations show that, in a world in which training data had to be licensed in a hypothetical competitive market to license data for training LLMs were to exist, Anthropic and each of the other companies building LLMs would need to license data from tens or hundreds of thousands of publishers (assuming sufficient publishers exist). When the fact that many copyrights are held by individual authors[30] and authors' individual copyright interests in blog posts and other material on the internet are considered, AI firms would need to potentially license content from millions of individual copyright owners or more to acquire the data needed to train a cutting-edge general LLM.

2. The Availability of Reliable Information in the Hypothetical Competitive Market for Training Data Licenses Is Poor

27. A market to license training data would require readily available, reliable information on copyright ownership and the value of licenses to operate efficiently.[31] However, the market for licenses for training data would have poor information on ownership and on the value of licenses, which may not be determined or verified at a cost the potential transaction would justify. Many existing training datasets come from sources on the Internet.[32] Studies show that the copyrights in existing training datasets containing text from the Internet are not well documented, which makes licensing the data in existing datasets impracticable or effectively impossible.[33] Beyond the problem of poor documentation, it may be impossible to

---

[30] Book publishers, for example, may not control the rights required to provide training licenses, and negotiations with individual copyright owners would be needed to acquire rights to their works. Lucio Lanucara, "Who Owns the Copyright to Published Works?" The Michelson Institute for Intellectual Property, September 27, 2021, available at https://michelsonip.com/who-owns-the-copyright/ (accessed 8/21/2024) ("In the United States, the Copyright Act (Title 17 US Code) states that intellectual property belongs to the author, unless otherwise specified in a publishing contract.").

[31] Robert S. Pindyck and Daniel L. Rubinfeld, *Microeconomics*, 7th ed., Pearson, 2009, p. 316 ("Market failure can also occur when consumers lack information about the quality or nature of a product and so cannot make utility-maximizing purchasing decisions.").

[32] *Jesse Dodge, et al.*, "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus," Paul G. Allen School of Computer Science & Engineering, University of Washington, September 2021 (arXiv:2104.08758v2).

[33] *Shayne Longpre, et al.*, "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI," Data Provenance Initiative, November 2023, (arXiv:2310.16787v3), p. 2; *Jesse Dodge, et al.*, "Documenting Large Webtext Corpora: A Case

identify and contact the copyright owners of many works because a work may be "orphaned," meaning its copyright owner cannot reliably be identified or contacted even with diligent effort.[34] The implication is that using text available on the Internet for text and seeking licenses is not a feasible approach to satisfying the licensing obligations that Plaintiffs seek to impose on AI firms. This implies that AI firms would need to first identify potential licensors, and then directly obtain and license text from them in order to avoid infringing copyrights if training were not fair use.

28. Even when copyright owners can be identified and contacted, the volume of licenses required makes mass licensing of copyrights impracticable. For example, the U.S. Copyright Office found that licensing for mass digitization "is essentially impossible, not necessarily because of the lack of identifying information or the inability to contact the copyright owner, but because of the sheer number of individual permissions required."[35] Thus, the scale of the licensing required can make it impracticable, even in the instances where the necessary information is available.

29. The lack of information regarding copyright owners and the scale of the required licensing effort makes it essentially impossible to license training data for a cutting-edge LLM through a competitive market with negotiations between copyright owners and AI firms. This is

---

Study on the Colossal Clean Crawled Corpus," Paul G. Allen School of Computer Science & Engineering, University of Washington, September 2021 (arXiv:2104.08758v2).

[34] Expert Report of Gloriana St. Clair, *The Authors Guild, Inc. v. Google Inc.*, No. 1:05-cv-08136 (S.D.N.Y.), Dkt. 1042-1 ¶ 29 (experiment showed that "only half of the publishers… responded to [license] request" despite efforts to send initial request letters and follow-up letters); *see also* "Orphan Works and Mass Digitization," United States Copyright Office, June 2015, p. 1, available at https://www.copyright.gov/orphan/reports/orphan-works2015.pdf (accessed 8/21/2024) ("a user's ability to seek permission or to negotiate licensing terms is compromised by the fact that despite his or her diligent efforts, the user cannot identify or locate the copyright owner. Second, in the case of mass digitization – which involves making reproductions of many works, as well as possible efforts to make the works publicly accessible – obtaining permission is essentially impossible, not necessarily because of the lack of identifying information or the inability to contact the copyright owner, *but because of the sheer number of individual permissions required* (emphasis added)."); p. 38 ("Studies of library collections of printed, published books and similar works estimate that between 17% and 25% of published works and as much as 70% of specialized collections are orphan works.").

[35] "Orphan Works and Mass Digitization," United States Copyright Office, June 2015, p. 1, available at https://www.copyright.gov/orphan/reports/orphan-works2015.pdf (accessed 8/21/2024).

particularly the case because of the low value of the licenses that result from the small contribution that even large publishers would make to the value of the LLM.

30. Other features of the copyright licensing marketplace make the large-scale licensing that Plaintiffs claim AI firms should undertake still more difficult to address. Frequently, a copyright has multiple owners. Such co-ownership is a well-documented challenge when it comes to licensing musical compositions. Owners of musical compositions and their agents have been vocal about their practice and expectation that licensees procure licenses from all co-owners of musical compositions, meaning that several licenses may be required to secure the required rights to one song and the informational burden of the transaction is higher as well.[36]

3. License Rates for Training Data Would Be Low

31. Competitive rates for licenses covering training data would be low in a hypothetical competitive licensing market. Training LLMs depends primarily on the quantity and diversity of text available.[37] The texts from different copyright owners used to train LLMs are part of a massive database, and each copyright owner makes only a tiny contribution to the whole, and in addition, copyright owners' works are fungible with other owners' works.[38] Put differently, an AI firm could exclude one copyright owner's data from a training set, or exclude the copyright owner's data and replace it with alternative data, without materially affecting the value of the LLM.[39] This implies that AI firms place a low value on incremental data for their training datasets – that is, they have a low willingness to pay for training data at the margin.

32. A Plaintiff with a one-million song catalog could offer Anthropic [redacted] (*i.e.*, [redacted]) of the content needed to train the latest version of Claude.[40] The incremental value

---

[36] Comments by the National Music Publishers Association Submitted in Response to U.S. Department of Justice Antitrust Division September 22, 2015, Solicitation of Public Comments Regarding PRO Licensing of Jointly Owned Works, available at https://www.justice.gov/atr/public/ascapbmi2015/ascapbmi22.pdf (accessed 8/21/2024).

[37] Kaplan Declaration, ¶¶ 17-26.

[38] Kaplan Declaration, ¶ 23.

[39] Kaplan Declaration, ¶ 23; *see also* "OpenAI and journalism," OpenAI, January 8, 2024, available at https://openai.com/blog/openai-and-journalism (accessed 8/21/2024).

[40] The marginal value of a publishers' data may be substantially smaller than the average value of content. Competitive input markets, of course, compensate suppliers based on the marginal value of the inputs provided.

of the content held by a large music publisher is smaller than the quantities of data that could be obtained from other publishers, but even they may have too little data to justify opening negotiations. For example, OpenAI reports that it explained to *The New York Times* in the course of their negotiations "that, like any single source, their content didn't meaningfully contribute to the training of our existing models and also wouldn't be sufficiently impactful for future training."[41] Individual authors with copyrights in books could license still smaller shares of the data required to train an LLM. Moreover, the interchangeability of copyright owners' data puts the copyright owners in competition with each other to sell training licenses. With copyright owners facing competition among themselves and facing opportunity costs that are essentially zero for training uses of their content, negotiated license rates would be low.

4. Negotiation and Transaction Costs in the Hypothetical Market Would Be High and Would Be Elevated by Uncertainty Regarding the Value of Licenses

33. Negotiating licenses to acquire training data would be administratively costly. The costs involved entail at least the cost of negotiating each agreement and the costs of continuing to track content and administer the resulting licenses.

34. *The New York Times* negotiated with OpenAI for months regarding real-time display of *New York Times* content with attribution on ChatGPT.[42] Thus, both OpenAI (and Microsoft) and *The New York Times* apparently invested substantial time in their effort to reach an agreement – and still failed. If *The New York Times* does not have sufficient data to make it worthwhile for OpenAI to expend the cost and effort to negotiate a license to train using *The New York Times'* archives, the vast majority of publishers would not have content that is sufficiently valuable to justify the cost of entering into a training license. Moreover, any dataset large enough to be worth licensing has no value if the rest of the data needed to train the LLM is

---

[41] "OpenAI and journalism," OpenAI, January 8, 2024, available at https://openai.com/blog/openai-and-journalism (accessed 8/21/2024). OpenAI describes its negotiations with The Times as not addressing training uses. Negotiations were "focused on a high-value partnership around real-time display with attribution in ChatGPT[.]"

[42] First Amended Complaint, *The New York Times Company v. Microsoft Corporation, et al.*, No. 1:23-cv-11195 (S.D.N.Y.), Dkt. 170, ¶ 7.

not available in the hypothetical competitive market.

35. Beyond the cost of negotiating licenses (*i.e.*, agreeing to the terms of the deal) is the administrative burden of tracking the content that licenses cover. Fox Corp. reportedly "launched a blockchain platform called Verify … to help media organizations monitor how their content is used online."[43] "Fox Corp. intends to use the Verify Protocol to negotiate deals licensing content from Fox networks to AI firms."[44] The implication is that Fox invested in a software tool in order to track what it has licensed to AI firms. The need for and benefits of a software tool indicate the complexity and cost of administering licenses. The costs of tracking licenses are both the cost of developing and maintaining the software tool and the cost of inputting data and managing the data in the tool (*e.g.*, updating records to reflect changes of ownership of content). Addressing the administrability of managing copyrights, Anthropic's Chief Science Officer also notes that excluding an ever-expanding corpus of text from Claude's training database would be extremely difficult in the course of Anthropic's business.[45]

36. A further complication in the hypothetical competitive market to license training data is that the incremental contribution to the value of an LLM created by a particular publisher's data is not known or verifiable for the parties to a hypothetical licensing negotiation. Uncertainty of this kind (*i.e.*, when negotiators do not know the value of a deal to their negotiating partner), makes negotiations more likely to fail even when there are gains from reaching an agreement.[46] Thus, poor information regarding the competitive value of a publisher's training data to an AI firm implies that negotiations could fail to reach agreement.

---

[43] Natalie Korach, "Fox corp. Launches Verify Tool to Check Authenticity of Content, Negotiate with AI Firms," The Wrap, January 9, 2024, available at https://www.thewrap.com/fox-corp-verify-tool-ai-negotiate-media-companies/ (accessed 8/21/2024).

[44] Natalie Korach, "Fox corp. Launches Verify Tool to Check Authenticity of Content, Negotiate with AI Firms," The Wrap, January 9, 2024, available at https://www.thewrap.com/fox-corp-verify-tool-ai-negotiate-media-companies/ (accessed 8/21/2024).

[45] Kaplan Declaration, ¶ 37.

[46] *See* Andreu Mas-Colell, Michael D. Whinston, and Jerry R. Green, *Microeconomic Theory*, Oxford University Press, 1995, p. 895 ("Whenever gains from trade are possible, but not certain, there is *no* ex post efficient social choice function that is both Bayesian incentive compatible and satisfies these interim participation constraints.").

Time spent on negotiations that do not lead to agreement would make the hypothetical negotiations for licenses more costly and reduce the amount of training data available through the competitive market.

37. In short, given the vast amount of information required to train LLMs and the practical difficulties associated with licensing enough individual works to amass such information, cutting-edge LLMs would likely not exist if required to license the works that make up their training datasets. This conclusion is validated by the testimony of Mr. Newton-Rex, in which he admits that an LLM trained on licensed and public domain data will not match the performance of today's cutting-edge LLMs because of the reduction in available training data.[47] For example, the KL3M large language model does not have "conversational chat" capability, among other limitations, according to its creators.[48] Plaintiffs' own witness recognizes and agrees that today's cutting-edge LLMs could not be trained in the legal framework that Plaintiffs advance. More generally, none of the "fairly trained" AI models that Mr. Newton-Rex describes perform the functions that cutting-edge LLMs perform and would not directly compete with them.

B. The Agreements Plaintiffs Have Identified Do Not Indicate that There Is an Incipient Market to License Training Data for LLMs

38. The hypothetical competitive market for training data would entail tens or hundreds of thousands, if not millions, of licenses for each AI firm. The existence of a few licenses— some of which address non-text data for image or music AI models and may be driven by idiosyncrasies of licensors and differences in the amounts of training data required for those products—does not demonstrate the practicability of a competitive market to license data to train LLMs.

39. Plaintiffs' expert, Professor Smith, states that "Licensing agreements by various

---

[47] Declaration of Ed Newton-Rex in Support of Plaintiffs' Motion for Preliminary Injunction, August 1, 2024, ¶ 29.

[48] *See* https://www.kl3m.ai/ (accessed 8/21/2024) ("While our pretraining data does include a number of conversational sources, we have not yet trained a model that is designed for standard conversational rounds."); Kaplan Declaration, ¶ 26.

other AI companies in recent years, both for written and non-written creative content, further demonstrate the development of a viable content licensing market for training AI."[49] Notably, Professor Smith does not appear to have reviewed the agreements themselves. Based on publicly available information, the agreements that Professor Smith enumerates either are not related to training LLMs or their terms are not clear and appear to encompass more than the licensing of training data. As a result, they do not support an inference that there is an incipient market to license all copyrighted textual data required for training general LLMs.

a) **Stability AI.** Stability AI trained its audio generation model with data from AudioSparx and claims that by partnering with a licensing company it "has permission to use copyrighted material."[50] The article that Professor Smith cites does not describe the nature of the partnership or the license, and specifically does not address whether it covered training. Moreover, the audio generator appears to have been trained on a single dataset of audio recordings, which indicates the data requirements required to train that model are far lower than the requirements faced by AI firms building cutting-edge general LLMs.

b) **Generative AI by Getty Images.** Getty worked with Nvidia to launch Generative AI by Getty Images. "Generative AI by Getty Images (yes, it's an unwieldy name) is trained only on the vast Getty Images library, including premium content, giving users full copyright indemnification."[51] This example does not provide information relevant to licensing text from third parties to train LLMs. First, Getty trained the image generator *using only its own data*. Thus, the volume of information needed to train the AI model to generate images appears to be much lower than in the case of the latest general LLMs. Second,

---

[49] Declaration of Michael D. Smith in Support of Plaintiffs' Motion for a Preliminary Injunction ("Smith Declaration"), 8/1/2024, ¶ 77.

[50] Emilia David, "Stability AI Releases AI Audio Platform," The Verge, September 13, 2023, available at https://www.theverge.com/2023/9/13/23871635/stability-ai-generative-audio-model-platform (accessed 8/21/2024).

[51] Emilia David, "Getty made an AI generator that only trained on its licensed images," The Verge, September 25, 2023, available at https://www.theverge.com/2023/9/25/23884679/getty-ai-generative-image-platform-launch (accessed 8/21/2024).

Getty assures users that they have copyright protection, which indicates the possibility that *output* from the model could infringe one of Getty's images. Any licensing or other considerations regarding potentially infringing output are separate from the issues raised using data solely for training.

c) **OpenAI-Associated Press.** OpenAI has licensed news stories from Associated Press. According to OpenAI and Associated Press "'The arrangement sees OpenAI licensing part of AP's text archive, while AP will leverage OpenAI's technology and product expertise[.]'"[52] The description of the license does not indicate that OpenAI is paying to train its LLMs using the Associated Press's articles, as opposed to accessing archival materials not otherwise available to it. In fact, the description of the deal is substantially broader, with direct benefits to AP's efforts to use AI.

d) **OpenAI-Shutterstock.** OpenAI has licensed images[53] and agreed to a license for "access to additional Shutterstock training data including Shutterstock's image, video and music libraries and associated metadata."[54] Thus, the agreement to access Shutterstock's data covers images, videos, music, and their associated metadata. The license is not reported to cover a text database. As described above, even if the licenses explicitly covered use of the images and music for training, these licenses are not informative about the hypothetical market for text data to train general LLMs.

40. Professor Smith also raises recent licenses between OpenAI, which creates LLMs

---

[52] Matt O'Brien, "ChatGPT-maker OpenAI signs deal with AP to license news stories," July 13, 2023, available at https://perma.cc/HZF6-S7KF (accessed 8/21/2024).

[53] "Shutterstock Partners with OpenAI and Leads the Way to Bring AI-Generated Content to All," October 25, 2022, available at https://www.shutterstock.com/press/20435 (accessed 8/21/2024).

[54] "Shutterstock Expands Partnership with OpenAI, Signs New Six-Year Agreement to Provide High-Quality Training Data," July 11, 2023, available at https://investor.shutterstock.com/news-releases/news-release-details/shutterstock-expands-partnership-openai-signs-new-six-year (accessed 8/21/2024).

comparable to Claude, among other models, and owners of copyrighted text data.[55] These content agreements are not available for review in this matter. However, the reports on them do not support the view that there is an incipient licensing market for *training data*.

41. There are several reasons that AI firms would "license" data. **First**, AI firms may license data for purposes that are consistent with their views that training LLMs with copyrighted data is fair use. Indeed, OpenAI has entered content agreements with publishers while publicly maintaining its position that training LLMs with copyrighted data is fair use.[56] In fact, OpenAI has recently affirmed this position in its litigation with *The New York Times* and *Daily News*.[57] The existence of a handful of licenses between OpenAI and content providers does not support the inference that OpenAI has come to the conclusion that it must license training data.

42. In fact, Universal International Music, B.V., ("Universal") produced [redacted].[58] Thus, the existence of a content

---

[55] Smith Declaration, ¶ 76.

[56] *See, e.g.*, *Silverman v. OpenAI, Inc.*, No. 23-cv-03416-AMO (N.D. Cal.), Dkt. 32 at 2-3; *Tremblay v. OpenAI, Inc.*, No. 23-cv-03223-AMO (N.D. Cal.), Dkt. 33 at 2-3; *see also* "OpenAI and Journalism," OpenAI, January 8, 2024, available at https://openai.com/blog/openai-and-journalism (accessed 8/21/2024) ("Training is fair use, but we provide an opt-out because it's the right thing to do"); OpenAI Letter Re: Notice of Inquiry and Request for Comment (Docket No. 2025-06), October 30, 2023, p. 11 ("OpenAI believes that the training of AI models qualifies as a fair use, falling squarely in line with established precedents recognizing that the use of copyrighted materials by technology innovators in transformative ways is entirely consistent with copyright law.").

[57] *See, e.g.*, Memorandum of Law in Support of OpenAI Defendants' Motion to Dismiss, *The New York Times*, No. 1:23-cv-11195 (S.D.N.Y.), Dkt. 52; Reply Memorandum of Law in Further Support of OpenAI Defendants' Motion to Dismiss, *Daily News, LP et al. v. Microsoft Corporation, OpenAI, et al.*, No. 1:24-cv-03285 (S.D.N.Y.), Dkt. 106.

[58] *See* UMPG_0000000510.

agreement does not imply that the parties reached agreement regarding training uses of copyrighted content.

43. **Second**, some agreements with copyright owners supplying data to AI firms allow for the output of the AI firm's models to reproduce the training data to a degree that is beyond fair use. For example, OpenAI recently "agreed to pay German media conglomerate Axel Springer, which publishes *Business Insider* and *Politico*, to show parts of articles in ChatGPT responses."[59] Thus, some agreements between AI companies and copyright owners cover uses other than training on publicly available content.[60]

44. **Third**, a license for access to data, or a broader deal allowing the display of copyrighted material, with a copyright owner that threatens to sue may be less expensive than defending a lawsuit, even if the AI firm expects to ultimately prevail on the training issue.

45. In short, none of the licenses Professor Smith describes supports the conclusion that a hypothetical competitive market to license sufficient training data to train today's cutting-edge LLMs, such as Claude or ChatGPT, is developing. In fact, such a claim is at odds with the relevant economics. If OpenAI's content agreements, for example, indicated that OpenAI believes that a market for training licenses is developing and that licenses were legally required for training, the handful of licenses that Professor Smith lists would be woefully inadequate to license OpenAI's training corpus and provide the necessary legal protections against copyright

---

[59] Will Oremus and Elahe Izadi, "AI's future could hinge on one thorny legal question," The Washington Post, January 4, 2024, available at https://www.washingtonpost.com/technology/2024/01/04/nyt-ai-copyright-lawsuit-fair-use/ (accessed 8/21/2024).

[60] *See, e.g.*, "OpenAI and journalism," OpenAI, January 8, 2024, available at https://openai.com/blog/openai-and-journalism (accessed 8/21/2024) (explaining OpenAI's approach to licensing); "SearchGPT Prototype," OpenAI, July 25, 2024, available at https://openai.com/index/searchgpt-prototype/ (accessed 8/22/2024); "News Corp and OpenAI Sign Landmark Multi-Year Global Partnership," News Corp, May 22, 2024, available at https://investors.newscorp.com/news-releases/news-release-details/news-corp-and-openai-sign-landmark-multi-year-global-partnership (accessed 8/22/2024) ("OpenAI has permission to display content from News Corp mastheads in response to user questions and to enhance its products"); "Le Monde and Open AI Sign Partnership Agreement on Artificial Intelligence," Le Monde, March 13, 2024, available at https://www.lemonde.fr/en/about-us/article/2024/03/13/le-monde-signs-artificial-intelligence-partnership-agreement-with-open-ai_6615418_115.html (accessed 8/22/2024) ("The agreement … provides for references to Le Monde articles to be highlighted and systematically accompanied by a logo, a hyperlink, and the titles of the articles used as references.").

infringement. OpenAI would need to license data from thousands (or millions) of additional authors and publishers to cover the text in its training database. Dr. Smith does not contend that OpenAI or any other AI company with a cutting-edge LLM is engaged in such an effort.

46. Professor Smith attempts to support the view that there is an incipient market to license training data by arguing that previous disruptive technologies, such as Internet radio and music streaming originally asserted that they could not afford licenses for music, but have since entered into licenses with song publishers and music labels.[61] The examples that Professor Smith cites, however, come from the music industry and are not analogous to the economics underlying the hypothetical licensing of training data for LLMs like Claude.

47. The examples Professor Smith cites involve products designed to reproduce, distribute, or perform musical works to the public. General LLMs like Claude do not stream or perform musical works and are designed to create new text in response to a wide range of user prompts. Regurgitation of training materials is considered a bug to be addressed, and not a feature, of products like Claude.[62]

48. In particular, Professor Smith fails to note that the licensing he cites does not take place in a competitive marketplace. Music licensing takes place in highly regulated markets. Some rights are subject to compulsory licenses with rates established by the Copyright Royalty Board. Performance rights are controlled by performing rights organizations ("PROs"), the two largest of which are subject to antitrust consent decrees that are enforced by the U.S. Department of Justice. The consent decrees require the subject PROs to provide a rate to potential licensees on demand and gives licensees the right to litigate rates in federal court.[63] Moreover, as shown above, the scale of licensing required for a cutting-edge LLM is several orders of magnitude (*i.e.*, several factors of 10) larger than the scale of music licensing. The fact that relatively small-scale music licensing requires a substantial regulatory framework involving compulsory licenses and

---

[61] Smith Declaration, ¶¶ 72-73.

[62] Kaplan Declaration, ¶ 48.

[63] *See United States v. Am. Soc'y of Composers, Authors & Publishers*, Civ. Action No. 41-1395 (WCC) (S.D.N.Y. Jun. 11, 2001) and *United States v. Broadcast Music, Inc.*, No. 64-Civ-3787 (S.D.N.Y. Nov. 18, 1994).

regulated rates to function, if imperfectly, validates the conclusion that a competitive market would not be able to coordinate the much larger set of licenses needed for an AI firm to license a training dataset for a cutting-edge general LLM, if such licenses were required. In fact, Professor Smith's analysis is insufficient to show that even some form of "regulated" marketplace for training rights would be practicable.

C. Using Song Lyrics to Train LLMs Does Not Threaten Plaintiffs' Existing Licenses

49. Claude's use of lyrics combined with other inputs to create new content does not threaten existing licenses with licensed lyric websites, or any other licensees of Plaintiffs' musical compositions for that matter. Training uses input data to teach an LLM how to process language and create new outputs. Such uses do not threaten Plaintiffs' existing licenses that permit outputs containing their musical works. An LLM such as Claude does not compete with music streaming services or Internet radio, for example. Thus, using lyrics for training does not threaten the licenses covering such public performances of music. Licenses to existing lyric websites might be threatened by services that reliably regurgitate a wide array of song lyrics for a sustained period of time.[64] However, training Claude or another LLM using text data is intended to create non-infringing outputs and does not use lyrics, or any other works, for the purpose of regurgitating training data as outputs.[65] Moreover, Plaintiffs recognize that Anthropic has implemented guardrails to prevent Claude from providing infringing output.[66] The production of non-infringing outputs is, of course, economically beneficial.

---

[64] It would be cheaper and easier to obtain accurate song lyrics from a licensed site than to attempt to get an LLM to regurgitate a potentially incorrect or incomplete set of lyrics, assuming that the song of interest was "memorized" by the LLM and that potentially infringing outputs are not effectively filtered before presentation to the user.

[65] Kaplan Declaration, ¶ 36.

[66] Plaintiffs' Notice of Motion and Motion for Preliminary Injunction; Memorandum of Law in Support of Plaintiffs' Motion for Preliminary Injunction, August 1, 2024, p. 29 ("Publishers are willing to take Anthropic at its word" that guardrails are effective "and seek simply to maintain these already-implemented guardrails"); *see also* Plaintiffs' Objections and Responses to Defendant's First Set of Requests for Admission (Nos. 1–15), April 22, 2024, p. 7 (admitting that "[Publishers] do not dispute Anthropic implemented guardrails seeking to prevent Claude from providing allegedly infringing outputs").

I declare under penalty of perjury that to the best of my knowledge, information, and belief, the foregoing statements are true and correct.

Executed on August 22, 2024, at Arlington, MA.



Steven R. Peterson