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

IN THE UNITED STATES DISTRICT COURT
FOR THE 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 [PROPOSED] SURRESPONSE TO PLAINTIFFS' RENEWED MOTION FOR PRELIMINARY INJUNCTION

REDACTED

Related Sealing Motion Pending at Dkt. 250

Hon. Eumi K. Lee

TABLE OF CONTENTS

I. QUALIFICATIONS AND ASSIGNMENT 1

A. Qualifications 1

B. Assignment and Materials Used 1

II. PLAINTIFFS' EXPERTS' NEW ARGUMENTS ARE UNSUPPORTED SPECULATION AND MISAPPLY ECONOMIC PRINCIPLES 1

A. Professor Smith Speculates About AI Firms' Agreements with Content Providers and Incorrectly Frames the Question of Whether a Licensing Market Could Exist 1

B. Mr. Newton-Rex's Discussion of the Quantity of Training Data Required to Train an LLM Is Economically Incomplete and Incoherent 5

I. QUALIFICATIONS AND ASSIGNMENT

A. Qualifications

1. My name is Steven R. Peterson. I previously submitted a declaration in this proceeding.[1] A complete statement of my qualifications is contained in my previous declaration.

B. Assignment and Materials Used

2. Counsel for Anthropic PBC has asked me to address the new arguments raised by Michael D. Smith and Ed Newton-Rex in their reply declarations. To prepare this declaration, I have reviewed the latest declarations submitted by Plaintiffs' experts, publicly available information, and relevant academic research.

II. PLAINTIFFS' EXPERTS' NEW ARGUMENTS ARE UNSUPPORTED SPECULATION AND MISAPPLY ECONOMIC PRINCIPLES

3. Neither Professor Smith nor Mr. Newton-Rex claims that a competitive market could coordinate the mass licensing effort needed to create training datasets for cutting-edge large language models ("LLMs"). Professor Smith speculates that agreements between AI companies and copyright owners allow the use of data for training, but he has not reviewed the actual terms of those agreements. Based on this assumption, he argues that as long as an AI company licenses some of the copyrighted material in its training dataset, that's enough to show a viable licensing market—but this argument violates basic economic principles for addressing the viability of a licensing market. Mr. Newton-Rex hypothesizes that it may be possible to train an LLM using less data than Anthropic claims, but he has not shown that training on less data is economically material to my conclusions.

A. Professor Smith Speculates About AI Firms' Agreements with Content Providers and Incorrectly Frames the Question of Whether a Licensing Market Could Exist

4. Professor Smith speculates about the nature and scope of agreements between AI firms and content owners that are not part of the record in this proceeding. Based on a review of news reports, Professor Smith asserts that agreements between OpenAI and content owners are

[1] Declaration of Steven R. Peterson, Ph.D. in Support of Defendant's Opposition to Plaintiffs' Renewed Motion for Preliminary Junction, August 22, 2024, ECF No. 210 ("Peterson Declaration").

market transactions for grants of training rights. As a matter of economics, no such conclusion is warranted.

5. A market is a mechanism for transferring ownership or other rights from one party to another, usually for money. While the descriptions of license agreements that Professor Smith cites sometimes reference uses of data for training, reports on the agreements Professor Smith describes vary and do not confirm a transaction for a grant of training rights.[2, 3, 4] Thus, without reviewing the actual agreements, Professor Smith has no economic basis to claim that the agreements transfer the rights to use content for training.[5] Moreover, reports on the

---

[2] The Reply Declaration of Michael D. Smith in Support of Plaintiffs' Motion for a Preliminary Injunction, September 12, 2024 ("Smith Reply Declaration"), cites to a number of press coverage on agreements, including: (1) News Corp's May 2024 deal with OpenAI, Smith Reply Declaration at ¶ 9, (2) The Associated Press's July 2023 deal with OpenAI, Smith Reply Declaration at ¶ 18, (3) *Times*' June 2024 agreement with OpenAI, Smith Reply Declaration at ¶ 18, (4) Dotdash Meredith's May 2023 deal with OpenAI, Smith Reply Declaration at ¶ 18, (5) LeMonde's March 2024 deal with OpenAI, Smith Reply Declaration at ¶ 18, (6) Taylor & Francis' May 2024 deal with OpenAI, Smith Reply Declaration at ¶ 19, (7) Axel Springer's December 2023 deal with OpenAI, Smith Reply Declaration at ¶ 20, and (8) Condé Nast's August 2024 deal with OpenAI, Smith Reply Declaration at ¶ 21. All of them either do not mention the use of data for training or include other rights like content display.

[3] The source used by Professor Smith to describe The Associated Press's July 2023 deal with OpenAI does not actually state that the agreement includes data for training. *See* Smith Reply Declaration at ¶ 18. It limits itself to the general statement about the amount of data used to train LLMs. *See* https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a (accessed 10/5/2024) ("OpenAI and other technology companies must ingest large troves of written works… to improve their AI systems known as large language models."). In fact, the press release on the deal only mentions that the parties agreed to share select news content and technology. *See* https://www.ap.org/media-center/press-releases/2023/ap-open-ai-agree-to-share-select-news-content-and-technology-in-new-collaboration/ (accessed 10/5/2024) ("The Associated Press and OpenAI have reached an agreement to share access to select news content and technology as they examine potential use cases for generative AI in news products and services.").

[4] Similarly, the agreements described in the plaintiffs' expert declaration do not pertain to data for training. *See* The Reply Declaration of Ed Newton-Rex in Support of Plaintiffs' Motion for a Preliminary Injunction, September 12, 2024 ("Newton-Rex Reply Declaration"), which cites to Condé Nast's August 2024 deal with OpenAI at ¶ 7. This agreement focuses on content display rights and does not mention the use of data for training.

[5] Economists widely agree that strong assumptions, like the ones maintained by Professor Smith regarding the content of the actual agreements, decrease the credibility of the inferences made with such assumptions. The issue is that assumptions, rather than data, drive results and the importance of the assumptions to the result is not clear and the quality of the assumptions is generally not testable. *See* Charles F. Manski, *Identification for prediction and decision*, Harvard University Press, 2007 at p.3 ("Credibility is a subjective matter, yet I take there to be a wide agreement on a principle (Manksi, 2003) that I have called *The Law of Decreasing Credibility* The credibility of inference decreases with the strength of the assumptions maintained.").

agreements indicate that they cover rights and obligations other than use of data for training.[6] Therefore, even if Professor Smith's speculation on training rights were correct, the amount paid for an agreement that bundles rights to use content for training rights with other rights to use content is not a proxy for the value of the former.

6. For example, OpenAI's agreement with *The Wall Street Journal* is described as including the right to display content.[7] Nevertheless, Professor Smith explicitly assumes that the reported potential value of OpenAI's agreement with *The Wall Street Journal* is the value of training rights. Professor Smith's inference is economically incorrect.[8]

7. Professor Smith attempts to explain away OpenAI's assertions in legal papers that training is fair use with the claim that OpenAI would have no need to sign deals with publishers if it were confident that training is fair use.[9] Of course, if the agreements cover uses of content that are regularly licensed, such as the right to reproduce and display content, or provide access to otherwise unavailable materials, metadata, or other consideration, there is good reason for OpenAI or another AI firm to enter a content agreement. Once again, Professor Smith's economic inference is simply incorrect as a matter of economics.

---

[6] As we discussed above, no agreements referred to by Professor Smith and Newton-Rex pertain exclusively to training. In fact, all of them pertain to content, and only News Corp, LeMonde, and Axel Springer agreements are reported to include training rights. Some content agreements definitively do not transfer training rights. As described in my August 22, 2024, declaration, [redacted] (*see* ¶ 7, [redacted]. Peterson Declaration at ¶ 42.

[7] News Corp's May 2024 deal with Open AI included rights to display content, *see* https://openai.com/index/news-corp-and-openai-sign-landmark-multi-year-global-partnership/ (accessed 10/5/2024) ("Through this partnership, OpenAI has permission to display content from News Corp mastheads in response to user questions…").

[8] The error in Professor Smith's logic is to assume that the part of the agreement that confers rights to use data for training is equal to the total payment under the agreement, which is also reported to include the right to display content. *See* Smith Reply Declaration, ¶ 11 ("Dr. Peterson surmises that a 100-year archive New York Times would likely yield 5.5 billion words and offered little marginal value to OpenAI. Peterson Decl. ¶ 25. If that were true, then the 1.6 billion works [*sic*] in *The Wall Street Journal* archive would not have come with a $250 million price tag to OpenAI.").

[9] Smith Reply Declaration at ¶¶ 25-26 ("[I]f OpenAI was confident in the fair use defense, there would be no need to sign these deals in the first place.").

8. Professor Smith's efforts to show the existence of some transactions for rights to train LLMs are in service of an effort to redefine the relevant questions from the existence and viability of a market to license training data to the question of whether some training data may be licensed. But Professor Smith does not assert that the comprehensive licensing of data for training datasets—i.e., the ability to license all copyrighted text required to train a general LLM—is feasible in a competitive market.[10] To circumvent the admitted limitation on a market's ability to mediate the necessary transactions, Professor Smith asserts without basis that the relevant question is whether there are *some* licenses to grant training rights.[11]

9. Professor Smith's recharacterization of the question of whether a market exists or could exist is unsupported by the evidence and wrong as a matter of economics. First, Professor Smith has assumed rather than demonstrated that the content agreements he describes include grants of training rights.

10. Second, the economic question of whether a market exists must be addressed in a coherent framework—that is, all market participants have to be viable in the proposed outcome. For the reasons discussed in my previous declaration, AI firms would not remain viable under Professor Smith's framework.[12] An AI firm that licensed only a subset of its data would not be operating legitimately if Plaintiffs' contention that training requires a license is correct. Moreover, AI firms would be subject to litigation by the owners of all unlicensed data in the training dataset.

---

[10] Smith Reply Declaration at ¶ 27 ("The proper question is whether there is a market for licensing or seeking authorization for copyrighted data in training corpora.").

[11] Smith Reply Declaration at ¶ 27 ("Additionally, the argument that it would be economically impractical for LLMs to negotiate licenses or seek authorization from all rightsholders and therefore it is economically impractical for LLMs to negotiate licenses with any rightsholders, *see* Peterson Decl. ¶ 45, not only suffers from the fallacy of composition—what is true for one member of the group does not make the same true for the entire group—but it also suffers from the empirical problem that plenty of rightsholders have negotiated multi-million-dollar licensing agreements with LLMs.").

[12] Peterson Declaration at § IV.

B. Mr. Newton-Rex's Discussion of the Quantity of Training Data Required to Train an LLM Is Economically Incomplete and Incoherent

11. Mr. Newton-Rex asserts that the "impediments to licensing or seeking authorization from copyright owners in building high-performing LLMs are exaggerated."[13] Even if correct, this claim does not support a conclusion that a market to license rights to use data for training is viable. As explained below, Mr. Newton-Rex's argument regarding reductions in the use of potentially copyrighted data for training is economically incoherent.

12. Mr. Newton-Rex claims that LLMs can be trained, at least in part, using synthetic data. As a result, the need for potentially copyrighted training data can be reduced, and the magnitude of non-synthetic data required to train a "high-performing LLM" cannot be determined.[14] A portion of this argument is related to the computer science of training LLMs, which I do not address. However, Mr. Newton-Rex's argument fails as a matter of economics. Even assuming his argument regarding the use of synthetic data were correct, Mr. Newton-Rex does not provide any information on the reduction in potentially copyrighted information that may be possible from the use of synthetic data, and he performs no analysis to demonstrate that a competitive market could coordinate the licensing of the required non-synthetic data. Thus, Mr. Newton-Rex does not address the economics of whether a market to license training rights for LLMs is economically viable.

13. Mr. Newton-Rex's argument is also economically incomplete. I understand that synthetic data for training an LLM would come from another LLM.[15] Mr. Newton-Rex does not address the requirements for or sources of data to train the LLM(s) that will generate synthetic data, which would presumably include human-generated and potentially copyrighted data.[16]

---

[13] Newton-Rex Reply Declaration at ¶ 9.

[14] Newton-Rex Reply Declaration at ¶ 8. Mr. Newton-Rex does not describe what he means by "high-performing LLM."

[15] Reply Declaration of Ben Y. Zhao in Support of Plaintiffs' Reply in Support of Motion for Preliminary Injunction, September 11, 2024 at ¶ 10 ("[S]ynthetic data is produced by generative AI that itself has already been trained on real-world data"); *see also* Surresponse Declaration of Jared Kaplan in Support of Defendant's Opposition to Plaintiffs' Motion for Preliminary Injunction, October 7, 2024 ("Kaplan Surresponse Declaration") at ¶ 9.

[16] Kaplan Surresponse Declaration at ¶ 9.

Thus, Mr. Newton-Rex does not adequately account for potentially copyrighted training data in the hypothetical training process he proposes. A discussion of the need for human-generated potentially copyrighted data to train an LLM that does not account for all of the potentially copyrighted data used in the proposed training process is economically incomplete and incoherent.

14. Mr. Newton-Rex also does not directly address the impediments to using a competitive market to coordinate the licensing of grants to use data to train a cutting-edge LLM. Moreover, he does not adequately demonstrate or describe the reduction in training data that he speculates is possible, and like Professor Smith, he does not demonstrate the existence of transactions for grants of training rights. Mr. Newton-Rex's declaration provides no economic basis to find that a market to license training rights is economically viable.

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 October 7, 2024, at Arlington, MA.



Steven R. Peterson