OPPENHEIM + ZEBRAK, LLP
Matthew J. Oppenheim
Nicholas C. Hailey
Audrey L. Adu-Appiah
(admitted *pro hac vice*)
4530 Wisconsin Ave., NW, 5th Floor
Washington, DC 20016
Telephone: (202) 480-2999
matt@oandzlaw.com
nick@oandzlaw.com
aadu-appiah@oandzlaw.com

Jennifer Pariser
Andrew Guerra
Timothy Chung
(admitted *pro hac vice*)
461 5th Avenue, 19th Floor
New York, NY 10017
Telephone: (212) 951-1156
jpariser@oandzlaw.com
andrew@oandzlaw.com
tchung@oandzlaw.com

COBLENTZ PATCH DUFFY & BASS LLP
Jeffrey G. Knowles (SBN 129754)
One Montgomery Street, Suite 3000
San Francisco, CA 94104
Telephone: (415) 391-4800
ef-jgk@cpdb.com

COWAN, LIEBOWITZ & LATMAN, P.C.
Richard S. Mandel
Jonathan Z. King
Richard Dannay
(admitted *pro hac vice*)
114 West 47th Street
New York, NY 10036-1525
Telephone: (212) 790-9200
rsm@cll.com
jzk@cll.com
rxd@cll.com

Attorneys for Plaintiffs

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., ET AL. Plaintiffs, v. ANTHROPIC PBC, Defendant.	Case Number: 5:24-cv-03811-EKL **REPLY DECLARATION OF BEN Y. ZHAO IN SUPPORT OF PLAINTIFFS' REPLY IN SUPPORT OF MOTION FOR PRELIMINARY INJUNCTION** REDACTED Hon. Eumi K. Lee

I, Ben Y. Zhao, hereby declare pursuant to 28 U.S.C. § 1746:

1. I am the Neubauer Professor of Computer Science at the University of Chicago, and a Fellow of the Association for Computing Machinery.

2. I submit this Declaration in connection with Plaintiffs' Reply in Support of the Motion for Preliminary Injunction filed by Concord Music Group, Inc., Capitol CMG, Inc.,

Universal Music Corp., Songs of Universal, Inc., Universal Music – MGB NA LLC, Polygram Publishing, Inc., Universal Music – Z Tunes LLC, and ABKCO Music, Inc. (collectively, "Publishers" or "Plaintiffs"), in their lawsuit against Defendant Anthropic PBC ("Anthropic").

3. This Declaration incorporates by reference my previous declaration dated August 1, 2024, ECF No. 181 ("August 1, 2024 Declaration"), submitted in support of Plaintiffs' Motion for Preliminary Injunction. My earlier declaration contains important background and details not repeated here, including my curriculum vitae, areas of expertise, and initial conclusions relating to the Anthropic's use of Publishers' works in the Claude model.

4. In preparing this Declaration, I relied on my general background and training, and I reviewed Publishers' filings, my previous declaration, Anthropic's Opposition to Plaintiffs' Motion for Preliminary Injunction ("Anthropic's Opposition" or "Opp."), supporting filings, and other documents referenced herein.

Effectiveness of Claude's Guardrails

5. I have reviewed the Claude prompts and outputs collected by Ed Newton-Rex in the last month, as detailed in his Reply Declaration and the supporting exhibits. Declaration of Ed Newton-Rex ¶¶ 12-15. These results demonstrate that Claude's current guardrails remain vulnerable. An easy-to-use method has been publicized and is available to users who wish to obtain Publishers' lyrics or to use the model to create mashups, advertisements, or other unauthorized derivative works based on those lyrics. These results also suggest to me that Anthropic has not been able to expeditiously test its guardrails and patch gaps as they appear.

6. As I explained in my prior Declaration, output-based guardrails are inherently imperfect. Such guardrails must be continually updated and improved as vulnerabilities are discovered and, even then, they will still never be able to prevent in all cases output containing material that has been included in the model's underlying training corpus. This is why relying on output filters alone to fix a flaw in the training of an AI model is not a complete solution. August 1, 2024 Declaration ¶¶ 59-60. I believe the additional outputs collected by Mr. Newton-Rex, as discussed in his Reply Declaration, further illustrate my conclusions about guardrails in general and as applied to Claude specifically.

Training on Authorized Lyrics is Not Impracticable

7. Anthropic argues that "[i]t would not be possible for a generative AI platform like Anthropic to amass sufficient content to train an LLM like Claude in arm's length licensing transactions, at any price." Opp. 6. In support, Dr. Kaplan claims each version of Claude was trained on "trillions of tokens," and that future models will likely require "[redacted] to continue to show performance gains." Declaration of Jared Kaplan, ECF No. 209 ("Kaplan Decl.") ¶ 29–30. [redacted] Kaplan Decl. ¶ 60, with an approximate ratio of [redacted] and [redacted] Peterson Decl. ¶ 14.

8. ***First***, I disagree with Dr. Kaplan's suggestion that future models of Claude will likely require "[redacted] of tokens to continue to show performance gains." Kaplan Decl. ¶ 29–30. The premise that bigger is always better does not comport with observations and literature from the wider AI development community. Thus, Dr. Kaplan admits that models trained on fewer tokens than Claude "***potentially*** [have] fewer capabilities." *Id.* ¶ 25 (emphasis added).

9. Recent work on LLM training and development has demonstrated that by using better training or querying techniques, model trainers can produce LLMs of similar power and capabilities with much smaller model sizes and using smaller training datasets. For example, researchers at Microsoft were able to train a coding LLM, phi-1, using less than 10% training data than comparable models but still achieving state of the art performance. *See* Suriya Gunasekar et al., *Textbooks Are All You Need* (Oct. 2, 2023), arXiv:2306.11644v2, https://arxiv.org/pdf/2306.11644. *See also* Yuanzhi Li, *Textbooks Are All You Need II: phi-1.5 technical report* (Sep. 11, 2023), arXiv:2309.05463v1, https://arxiv.org/pdf/2309.05463; Cheng-Yu Hsieh, et al., *Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes*, FINDINGS OF THE ASS'N FOR COMPUT. LINGUISTICS: ACL 2023 8003 (2023), available at https://aclanthology.org/2023.findings-acl.507.pdf (applying "distilling step-by-step" mechanism during training resulted in outperforming LLMs with less training data). Similarly, Google DeepMind researchers were able to demonstrate how finetuning LLMs using synthetic data generated by "weaker but cheaper" models consistently outperformed

LLMs finetuned on more expensive model-generated data. Hritik Bansal et al., *Smaller, Weaker, Yet Better: Training LLM Reasoners via Compute-Optimal Sampling* (Aug. 29, 2024), arXiv:2408.16737v1, https://arxiv.org/pdf/2408.16737.

10. Dr. Kaplan admits Claude is trained on a "proprietary mix of publicly available information on the Internet, as well as non-public data from third parties, data provided by data labeling services and paid contractors" as well as "***data created internally***." Kaplan Decl. ¶ 28 (emphasis added). As others have suggested, *see, e.g.,* Benji Edwards, *The AI wars heat up with Claude 3, claimed to have "near-human" abilities*, ARS TECHNICA (Mar. 3, 2024), https://arstechnica.com/information-technology/2024/03/the-ai-wars-heat-up-with-claude-3-claimed-to-have-near-human-abilities/2/, this final category of data likely refers to synthetic data, which is data that has been generated by a generative AI model with similar properties to real-world data. Dr. Kaplan also admits "Anthropic may . . . develop[] synthetic data generated using our existing LLMs." Kaplan Decl. ¶ 28. Since synthetic data is produced by generative AI that itself has already been trained on real-world data, an LLM that is trained on synthetic data ***and*** real-world data requires less real-world data than a similarly-performing LLM trained fully on real-world data alone. I believe this is the case with Claude.

11. In other words, the "trillions of tokens" required to train Claude likely do not translate to "[redacted]" Kaplan Decl. ¶ 60. Assuming synthetic data is present in the training corpus, the amount of real-world text required to train each version of Claude is likely much less than implied by Dr. Kaplan. *Id.* ¶¶ 22, 28-30.

12. ***Second,*** Anthropic's argument vastly overstates the necessary quantities of additional copyrighted text required or available to train an LLM model like Claude. Consider the fact that AI developers have already trained on the majority of human-produced text available on the internet today, so much so that they will soon run out of data as early as 2026. *See* Alexandra Tremayne-Pengelly, A.I. Companies Are Running Out of Training Data: Study, Observer (Jul. 19, 2024), https://observer.com/2024/07/ai-training-data-crisis/. Simply put, there is a limited amount of existing, human-generated, copyrighted content currently available for ingestion by AI developers, and new copyrighted content is added slowly. The "trillions of tokens" to train

"models . . . like Claude," Kaplan Decl., ¶ 25, refers to generic, low quality data, and has little to do with high quality, human produced, copyrighted material. The latter could not grow [redacted] even if paid to do so. This is all the truer for Publishers' lyrics. So the idea that licensing costs would grow [redacted] with every iteration of the Claude model is absurd.

13. Thus, the scope of the problem as described by Anthropic—that "it would not be possible . . . to amass sufficient content to train an LLM like Claude in arm's length licensing transactions," Opp. 6—is likely overstated and does not properly reflect the true state of LLM developmental capabilities today.

Cost to Remove Plaintiffs Works from Training Dataset

14. Dr. Kaplan states that [redacted] Kaplan Decl. ¶ 61. Notably, Dr. Kaplan previously stated [redacted]. *See* Declaration of Jared Kaplan dated January 15, 2024, ECF No. 67-1 ¶ 44, Although Dr. Kaplan gives the surprising testimony [redacted] the reasoning and analysis set forth in my previous Declaration regarding this assertion remains the same: [redacted]. August 1 Declaration ¶ 64. [redacted] My understanding is that Publishers are not asking that Claude be retrained solely to eliminate their works from the training corpus.

I declare under penalty of perjury under the laws of the United States that the foregoing is true and correct to the best of my personal knowledge and belief.

Ben Y. Zhao

Executed in Chicago, IL, this 11 day of September, 2024.

SIGNATURE ATTESTATION PURSUANT TO CIVIL L.R. 5-1(h)

Pursuant to Civil L.R. 5-1(h), I hereby attest that concurrence in the filing of the document was obtained from the document's signatory. I declare under penalty of perjury that the foregoing is true and correct.

Dated: December 26, 2024

/s/ Timothy Chung
Timothy Chung