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) 772-5795
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: 3:24-cv-03811-EKL **DECLARATION OF BEN Y. ZHAO IN SUPPORT OF PLAINTIFFS' 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 ("ACM"). I submit this declaration in connection with Plaintiffs' Motion for a Preliminary Injunction filed by Plaintiffs Concord Music Group, Inc., ("Concord") Capitol CMG, Inc., Universal Music Corp., Songs of Universal, Inc.,

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

2. My statements set forth below are based on my specialized knowledge, education, and experience, as applied to the facts and circumstances of this case. If called upon, I would and could competently testify as to the matters contained herein.

I. My Assignment

3. I previously submitted a declaration, ECF No. 47, and a reply declaration, ECF No. 107, in support of Plaintiffs' prior Motion for Preliminary Injunction, ECF No. 40, filed in the U.S. District Court for the Middle District of Tennessee. This updated declaration is submitted in support of Plaintiffs' renewed Motion for Preliminary Injunction filed in the U.S. District Court for the Northern District of California, consistent with that renewed motion and also to address certain new facts and developments since my earlier declarations were filed.

4. Publishers have asked me to carefully consider Anthropic's current large language model known as "Claude," and to make determinations of its functionality, its likely training data and training process, and in the context of guardrails to limit or eliminate the harms of reproducing (in whole or in part) Publishers' lyrics, what mechanisms are possible, what is currently deployed, and their fundamental limitations. I am asked to answer these following questions:

- Has Anthropic copied Publishers' lyrics and used those copies as input to train Claude?
- How critical is this copying and training using Publishers' lyrics to the general functionality of Claude?
- Can Anthropic implement guardrails that prevent Claude from generating copies of Publishers' lyrics as output, and what are the limitations/failures of these mechanisms?
- Can Anthropic train a large language model that is guaranteed not to copy Publishers' lyrics in its output?

5. In my work on this matter, I have reviewed and utilized a variety of materials, including (a) publicly accessible versions of Claude; (b) recorded transcripts of prior interactions with Claude referenced in the Declaration of Dan Seymour of BC Guardian ("Seymour Decl."), ECF No. 49, submitted in support of Publishers' Motion for Preliminary Injunction; (c) current research literature on generative AI models, particularly on large language models; (d) the Declaration of Jared Kaplan, Chief Science Officer of Anthropic ("Kaplan Decl."), ECF No. 67-1, submitted in support of Anthropic's Opposition to Publishers' Motion for Preliminary Injunction; and (e) news articles, reports, and other publicly available information regarding Anthropic, Claude, and other AI models and related issues. I am also relying on my experience and background in artificial intelligence, deep neural networks, and adversarial machine learning generally.

6. I note that Claude is a proprietary system developed by Anthropic, which has not released to the public the underlying source code, training data content and sources, model weights, or comparable material to permit a detailed inspection of its operation. Researchers have given Anthropic and its Claude 2 AI model a transparency score of 51%.[1]

7. In discussing my analysis in this declaration, I do so in a manner geared toward a general audience. If I am asked to write an expert report at a later stage of this case, I can elaborate on the points made herein in more technical terms with detailed citations, but I assume a higher level of discussion is more helpful to the Court at this stage.

II. My Background

8. I am the Neubauer Professor of Computer Science at the University of Chicago, a Fellow of the Association for Computing Machinery, and co-director of the UChicago Security, Algorithms, Networks, and Data (SAND) Lab. I teach and mentor undergraduate and PhD students and conduct research. Over the years, I have led numerous projects with $30M funded by the National Science Foundation, Department of Defense, Department of State, and industry research

[1] *The Foundation Model Transparency Index*, CTR. FOR RSCH. ON FOUND. MODELS, https://crfm.stanford.edu/fmti/ (last visited Jul. 31, 2024).

groups. Prior to starting my faculty career 20 years ago, I received my PhD from UC Berkeley, and my Bachelor of Science from Yale University.

9. My research lies at the intersection of machine learning and security, with additional topics covering data-driven systems and human computer interaction. My work is published regularly at the most selective publication venues in computer security and machine learning and has received numerous best paper awards, distinguished paper awards, most-influential paper awards and the Internet Defense Prize. My early work created the research area now known as distributed hash tables, and those systems are now used in most large cloud computing systems today. More recently, my projects established the first defenses against backdoor attacks on deep neural networks and developed adversarial techniques to protect creative artists against misuses of generative AI models. My 190 publications have garnered over 36,000 citations. My full curriculum vitae is attached to this declaration as **Exhibit A**.

III. Context

10. Since its inception, Anthropic has branded itself as an AI company prioritizing AI safety and minimizing harm. In their core views, they list "critically evaluating the potential societal impacts of our work" as a "key pillar of our research."[2] Their approach to AI safety, billed as so-called "Constitutional AI," is designed to give an AI system a set of principles (i.e., a "constitution") against which it can evaluate its own outputs. Anthropic offers its main product, Claude, a series of AI large language models ("LLMs"), as a standalone chatbot interface available through its website, as a mobile application for iPhone and Android users, and as an application programming interface ("API") for integration by businesses and other customers into their products and applications.

IV. Summary of Conclusions

11. Based on transcripts of interactions with Claude described in the Complaint, evidence listed in the Seymour declaration, and publicly available information about Anthropic's training data, it is clear that Anthropic has copied Publishers' lyrics and used those copies as input

[2] *See Research Principles*, ANTHROPIC, https://www.anthropic.com/research (last visited Nov. 11, 2023).

to train Claude. However, Anthropic's use of Publishers' lyrics in training is not critical to the general functionality of Claude, and Anthropic could remove those lyrics from training without undermining Claude's operation. Moreover, while it is apparent from Claude's outputs that Anthropic has put in place certain guardrails (adopted largely post-suit) to limit Claude from including Publisher's lyrics in output, those guardrails vary in effectiveness and ultimately depend on Anthropic's willingness to maintain them. In the short run, such guardrails might stem the inclusion of Publishers' lyrics in Claude's outputs, and Anthropic should be required to maintain those guardrails it already has in place. In the long term, however, the only foolproof way to ensure that Claude's outputs do not include Publishers' lyrics is for Anthropic to remove copies of those lyrics from its training material for future AI models. As discussed below, this is a reasonable solution to the problem presented by Anthropic's choices.

V. What is AI?

A. Traditional vs. Generative AI

12. Artificial intelligence ("AI") systems are built to flexibly perform tasks that involve a wide variety of novel situations and that ordinarily would be expected to require human intelligence to accomplish. "Traditional" AI is designed to perform a specific task in a predictable way to user-supplied inputs based on pre-determined rules. These systems mainly fall under the broader category of discriminatory classification systems because their primary goal is to recognize complex patterns inherent in multi-faceted data and to produce decisions as a result. For example, a drug-discovery AI system might take as input a new hypothetical chemical compound structure, and based on its training, produce a classification result that predicts whether the compound will have a specific prescriptive impact on a particular disease or pathology. In other examples, a trained deep neural network ("DNN") classifier may take in photographic images of skin lesions and predict which samples are likely cancerous, and a different DNN might take audio samples of a human voice and produce a classification of the emotional state of the speaker and if they are telling the truth. Many well-known AI applications are built using traditional AI, such as spam filters in email, AI-generated recommendations on e-commerce platforms, self-driving cars,

data mining, virtual assistants like Siri, and chess-playing programs. These architectures and applications existed years before the current arrival of so-called "generative" AI applications.

13. Generative AI strives to generate content by combining existing materials together using patterns it derives from training materials. It utilizes complex machine learning architectures to extract correlation patterns between a textual prompt or request, and specific content samples (which can be text, images, or audio) that match those prompts. Sufficient training on large datasets produces the generative model, the encapsulation of a massive compilation of correlations between text prompt and content, largely stored as conditional probabilities. These probabilities can then guide the reproduction of content that have the highest chances of matching any text prompt given by the user of the generative model.

14. Today's leading generative AI systems can be used to generate text, images, audio visual works, music, and more. Regardless of their specific task or architecture, generative AI systems are trained on very large datasets of content according to the form of content they are intended to generate and input they are designed to accept. The larger the corpus on which a model is trained, the more effective and reliable its results will be as a general matter.

15. LLMs are currently among the most popular applications of generative AI today. They can be deployed in different settings to perform a variety of natural language processing ("NLP") tasks, including text summarization, language translation, and generating all forms of written work including novels, essays, poems, song lyrics, nonfiction works, and software code. Text-generative LLMs are trained to respond to queries in a small number of languages, so users can input prompts and obtain results in natural language without the need to write computer code.

16. All AI models are based on a particular architecture (the design and structure of the model). LLMs typically utilize a "transformer" architecture, a type of machine learning algorithm.[3] Among other features, a transformer is comprised of an encoder, which reads and processes the input data, and a decoder, which generates predictions. Transformers work on predicting high probability text "tokens" given some contextual information, where the context is often measured

[3] Ashish Vaswani et al., *Attention Is All You Need* (Aug. 2, 2023), arXiv:1706.03762v7, https://arxiv.org/pdf/1706.03762.pdf.

in number of tokens of memory. The more powerful and complex the model, the more tokens it can capture in its contextual memory to guide its prediction of meaningful responses to incoming user questions and requests.

17. Many LLMs are made available for use by human users on a website or mobile application through a "chatbot" (a computer program designed to simulate conversation between human users) or by other software through an application programming interface ("API"). In either case, the user supplies an input known as a "prompt," the model processes the prompt in a step known as "inference," and then the service returns a response or "completion," also called "outputs."

B. LLM Training

18. The creation of an LLM is a multi-stage project. Preliminary steps include defining the object of the model (whether it will be a chatbot, a translation tool, etc.), designing the architecture, and related tasks. This initial work is followed by data-intensive steps of training the defined model. In preparation for training, the creator of an LLM gathers text material and other content, typically in massive quantity and from a variety of sources, and copies it to a central repository. Frequently, the text is obtained from websites, and the LLM creator may either copy it directly from them, use archives of such text compiled by third parties, or use a mix of sources.

19. The next step is "cleaning" the raw aggregated text according to various criteria, such as removing certain sorts of recognized duplicates (for example, copies of the same webpage). In addition, content will be removed that is inconsistent with the LLM creator's objectives for the model; examples could include offensive material such as pornography, software source code, and certain languages or character sets. The result of this process is a cleaned version of the dataset that then can be used for training; this cleaned version represents an additional copy.

20. After the data is cleaned, the text is converted into "tokens." On average, a token is three quarters of a word, meaning a token is a combination of a few characters (including punctuation) with a ratio of about 4 tokens per 3 words.

21. After the dataset is cleaned and "tokenized," the LLM is trained. The first step is called pre-training, during which the model is trained to predict the next token/word in a string.

For example, it will be given the sentence, "The dog wagged its ___," and will learn to predict the next word as "tail." Or "I like ice ___," where the answer could be "cream," "cubes," or various other choices. Through pre-training, the model will learn which of these is the best choice from context. Pre-training also involves "weighting" the dataset. In this process, the trainer directs the model to use some parts of the dataset more often than others due to differences in perceived quality of the material, with portions engaged multiple times and others not at all. Ultimately, the model will learn the most common patterns in the grammar, syntax, and semantics of a given language. This process requires no supervision (i.e., labeling or metadata describing the contents of the text). Rather, the model recognizes patterns in the training data to infer these rules automatically.

22. After a model is pre-trained, the next step is "finetuning," during which the LLM is specialized to perform a particular task through supervised learning.[4] This process typically utilizes "reinforcement learning," where computers and humans rate the output supplied by the model in response to a given prompt, adjusting the values of the model's parameters to steer it to prefer certain sorts of responses and avoid others. This enables the model to produce more accurate and tailored responses, consistent with those desired by the LLM developer. The most popular version of this training is commonly referred to as RLHF, or "reinforcement learning with human feedback."

C. LLM Outputs

23. In generating outputs, LLMs are entirely dependent on their training datasets. A recent paper from Google DeepMind confirmed what most in the community already knew, that transformers cannot generalize beyond their pretraining data.[5] Put another way, the entirety of the functionality of these LLMs is defined by their training data, and they are unable to generalize beyond it. In addition, once an LLM is trained on a given piece of content, it can use that content

[4] *Glossary,* ANTHROPIC, https://docs.anthropic.com/claude/docs/glossary (last visited Jul. 31, 2024).

[5] Steve Yadlowsky et al., *Pretraining Data Mixtures Enable Narrow Model Selection Capabilities in Transformer Models* (Nov. 1, 2023), arXiv: 2311.00871, https://arxiv.org/pdf/2311.00871.pdf.

for any purpose. An LLM does not differentiate among the sources of training data in creating outputs, apart from the particular weights the developer may have applied to the data during the training process. It may use the tokens representing song lyrics to generate lyrics, poetry, nonfiction essays, marketing materials, movie scripts, or anything else.

24. Beyond those steps taken during the cleaning and training process, in an effort to constrain model responses, LLM creators may implement "guardrails" in an effort to limit undesired output. These guardrails can be applied as filters on prompts aimed at preventing problematic generation in the first place. They also can be implemented as filters on completions with the goal of catching undesirable output before it is delivered to users. In either case, the guardrails exist outside the underlying model as opposed to being inherent within it. Guardrails can be enforced reactively to limit model output of certain text taken from training data, but research demonstrates that there are consistent and reliable ways to bypass existing guardrails, and they are not a panacea to limit all potential abuses.[6]

25. While LLMs are not, in general, a simple database of the training content, it is possible for certain content to be "memorized" by LLMs. Studies have shown memorization by LLMs is more likely for content that is repeated frequently in the training dataset and that larger LLMs memorize more than smaller ones.[7] As with most machine learning models, the large space of potential prompts or queries can only be trained with massive amounts of training data covering the broad span of prompts. While model training becomes more robust with increased volume of training on a particular topic, the opposite also holds. Words, proper names, or relatively unusual sentences or phrases that occur rarely in training data will lead LLMs to "overfit" on them, making the model more likely to repeat them (and phrases or text blocks that contain them) verbatim given the correct leading sequence or prompts. For training data that is extremely popular (and therefore is represented in training data over and over as different copies), this overfitting effect is magnified

[6] Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models* (July 27, 2023), arXiv: 2307.15043, https://arxiv.org/pdf/2307.15043.pdf.

[7] Kent K. Chang et al., *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4* (Oct. 20, 2023), arXiv: 2305.00118v2, https://arxiv.org/pdf/2305.00118.pdf; Antonia Karamolegkou et al., *Copyright Violations and Large Language Models* (Oct. 20, 2023), arXiv: 2310.13771, https://arxiv.org/pdf/2310.13771.pdf.

and reinforced. This leads to the observed effect that rare token sequences are more easily extracted from LLM models.[8] In fact, current research from Google establishes the understanding that LLMs memorize text, and memorization grows with (1) increases in capacity of a model, (2) the number of times an example has been duplicated, and (3) the number of tokens of context used to prompt the model.[9]

D. Updating LLMs

26. An LLM can only utilize the information in the dataset on which it was trained, which makes that information fixed in time as of the moment the dataset was collected. These models therefore get "stale," and there is enormous pressure on commercial LLMs to regularly update their models to incorporate the most recent content available into the datasets and make any needed technical adjustments.

27. When a developer decides to release a new version of an LLM, it typically will not simply insert the latest information to the existing dataset like adding pages at the end of a book. Instead, the dataset is recreated incorporating both the old and new material, and then trained from scratch in the process described above. Anthropic apparently made its most recent update in less than a year (see discussion of Claude below).

28. Likewise, if a developer wants to exclude certain content from its dataset, it cannot cause an existing version of the LLM to "forget" material used in its training—training data cannot be removed once tokenized and used to set model weights. But the LLM can be relaunched without that particular content as part of its dataset when the next version of the LLM is released. So, while removing data from a model once it is trained is not technically possible, practically it can be achieved any time the model is updated.

VI. LLMs and Copyright

29. Respecting copyright while building an LLM model is reasonable and feasible. Of course, an AI developer can seek licenses from copyright holders for their content. But even

[8] Nicholas Carlini et al., *Extracting Training Data from Large Language Models* (Jun. 15, 2021), arXiv: 2012.07805v2, https://arxiv.org/pdf/2012.07805.pdf.

[9] Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* (Mar. 6, 2023), arXiv: 2202.07646, https://arxiv.org/pdf/2202.07646.pdf.

without taking licenses, the developer can (1) avoid training AI models on unlicensed copyrighted materials (including both "pirate" content not legally obtained by the website/database and material that may have been licensed from the copyright owner by the site operator but was not licensed by the AI developer); and (2) implement effective guardrails.

30. There are numerous available means by which LLM developers can avoid training their AI models on unlicensed copyrighted materials. For instance, model developers can avoid scraping from sources known to contain unlicensed copyrighted material, such as Books3, which is a massive database of nearly 200,000 books downloaded from pirate sources.[10] Model developers can also look out for copyright identifiers on the materials they use in training, such as the copyright symbol ("©"), authors' or publishers' names, copyright dates, and other related indicia of copyright protection, and avoid including such material in training. Some of that material may be licensed for public use through Creative Commons licenses, but that should be clearly indicated. Developers can also utilize databases of copyrighted works as a "blacklist" to exclude from the training data any instance of those works found in that database. Additionally, developers can seek and secure licenses to include copyrighted content in their training data, either by licensing from content owners directly or by using training datasets of pre-licensed and cleared data. By avoiding training their AI models on unlicensed copyrighted materials, LLM developers can avoid copying those materials in the output of their models.

31. Developers can also attempt to limit the display of copyrighted material in outputs by implementing content-based blockers (rules that restrict the model from responding to certain inputs or including specific text either in prompts or outputs) as guardrails. Unfortunately, however, guardrails are imperfect.

32. First, content-based blockers are known to be fragile, and can be bypassed by a user with sufficient knowledge of LLMs. Recent research published last year at the top machine learning conference, the International Conference on Machine Learning ("ICML"), demonstrates

[10] Alex Reisner, *These 183,000 Books Are Fueling the Biggest Fight in Publishing and Tech*, The Atlantic (Sep. 25, 2023), https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/

fundamental flaws in the way transformer language models are built, such that there are deterministic ways to compute non-sensical characters, which, when added to a prompt, will bypass prompt blockers, and allow the query to be answer "truthfully" by the LLM.[11] This work also demonstrates that these vulnerabilities are fundamental to the transformer architecture and are applicable to all commercial LLM models. It is also not a property or weakness that can be easily addressed. Transformer architecture is inherently vulnerable to these attacks and simply cannot be "debugged" to prevent infringing outputs in all cases. Accordingly, while guardrails have their place in a well-designed LLM, they are inherently limited.

33. Second, output-based guardrails do not address the underlying use of unlicensed copyrighted works as part of the dataset in the first place. Even if they were capable of functioning perfectly (which, as discussed above, they do not), such guardrails are aimed solely at limiting output with copyrighted content; they do not and cannot remove the underlying copyrighted works from the training set.

34. Third, such guardrails are ultimately subject to the willingness of the LLM developer to implement, update, and maintain them. While avoiding unlicensed copyrighted content in an LLM's training data will permanently prevent that LLM copying such content in its output, output-based guardrails are inherently temporary in that they may be removed or modified at any time at the discretion of the LLM developer. Further, there is no efficient way for a regulatory agency or third party to detect if an LLM developer removes guardrails or stops efforts to maintain or update guardrails. This makes reliance on developer-built guardrails for legal compliance particularly risky.

VII. Claude

A. Overview

35. Claude is a system with an LLM at its core and an interactive interface ("chatbot") built on top. The current versions include Claude 3 (released in March 2024) and Claude 3.5

[11] Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models* (July 27, 2023), arXiv: 2307.15043, https://arxiv.org/pdf/2307.15043.pdf; Zico Kolter, Keynote Presentation at the 2nd ICML Workshop on New Frontiers in Adversarial Machine Learning: Adversarial Attacks on Aligned LLMs (Jul. 28, 2023).

(released in June 2024). Anthropic has also released several other versions of Claude, dating back to its original release in March 2023. Claude is available as a web version and mobile application for individual users and as an API for business users.

36. Claude's primary difference from other LLMs is its reliance on what Anthropic calls "Constitutional AI," a list of rules or principles that the model slowly approximates using a combination of supervised learning and reinforcement learning, which Anthropic calls "RL from AI Feedback" ("RLAIF").[12] In comparison, OpenAI's ChatGPT models rely more heavily on "Reinforcement Learning with Human Feedback" ("RLHF").[13]

37. The Claude web version is available as a paid subscription-based version and a more limited free version. The free version of Claude can be accessed by anyone who creates an account verified with a text message sent to a mobile phone number.

38. The current and earlier versions of Claude are also available to commercial customers to be incorporated into their own software, products, and systems through an API. Claude models are currently already integrated into a variety of products, including group messaging applications (like Slack), Internet search engines (such as DuckDuckGo), note-taking applications (like Notion), and AI-aggregator sites (such as Poe). To interact with the Claude chatbot in these various platforms, users typically enter a prompt into the customer software. Claude then processes the users' prompt and generates a response, which is then delivered back to the user through the customer software. This process is invisible to the ultimate end user.

B. Training/Copying Copyrighted Material

39. Anthropic has the ability to decide what content it will use to train Claude and whether it will disclose that information publicly. Prior to March 2023, certain LLM developers were more transparent about the contents of their training dataset.[14] Anthropic has chosen not to

[12] Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (Dec. 15, 2022), arXiv: 2212.08703, https://arxiv.org/pdf/2212.08073.pdf.

[13] Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback* (Mar. 4, 2022), arXiv: 2203.02155, https://arxiv.org/pdf/2203.02155.pdf; *GPT-4 Technical Report*, OPENAI (Mar. 15, 2023), arXiv: 2303.08774, https://arxiv.org/pdf/2303.08774.pdf.

[14] Kyle Barr, *GPT-4 is a Giant Black Box and Its Training Data Remains a Mystery*, GIZMODO (Mar. 17 2023), https://gizmodo.com/chatbot-gpt4-open-ai-ai-bing-microsoft-1850229989.

make its training dataset information public. Anthropic has stated only that "Claude 3 models are trained on a proprietary mix of publicly available information from the Internet as of August 2023, as well as non-public data from third parties, data provided by data labeling services and paid contractors, and data we generate internally."[15]

40. However, Anthropic's limited public disclosures and response to the Publishers' Motion for Preliminary Injunction filed in the Middle District of Tennessee, ECF No. 40, confirm that Anthropic relies heavily on AI training datasets (e.g., the "Common Crawl" and "The Pile" datasets) that include large amounts of copyrighted and unlicensed material, including Publishers' lyrics.[16] In particular, the Common Crawl includes unlicensed copies of Publishers' lyrics scraped from licensed lyrics websites like LyricFind.com and Musixmatch.com.[17] Meanwhile, The Pile is known to include the "YouTube Subtitles" sub-dataset, which contains various unlicensed copies of Publishers' lyrics compiled from the subtitles of licensed YouTube videos for those works.[18]

41. It is also evident from the outputs BC Guardian has collected from Claude that Anthropic has trained Claude on Publishers' lyrics.[19] We would obviously not see exact or near-exact copies of song lyrics in response to a query unless the system had copied and stored those lyrics. The probability of a randomized LLM generating specific lyrics comprised of hundreds of words in a single song, let alone hundreds of songs, without having trained on them, is astronomically low.

42. In addition, while LLMs are designed to mimic fluent speech and text composition, the perfect or near-perfect reproduction of song lyrics cannot have been generated naturally by the LLMs themselves without having been trained on those lyrics. Song lyrics often use non-standard

---

[15] *The Claude 3 Model Family: Opus, Sonnet, Haiku*, ANTHROPIC, https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

[16] *A General Language Assistant as a Laboratory for Alignment* 8, 27, ANTHROPIC (Dec. 9, 2021), https://arxiv.org/pdf/2112.00861.pdf; Kaplan Decl. ¶¶ 24–25, ECF No 67-1

[17] *See* Kevin Schaul et al., *Inside the secret list of websites that make AI like ChatGPT sound smart*, Washington Post (Apr. 19, 2024), https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/.

[18] *See* Annie Gilbertson & Alex Reisner, *Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI*, Proof News (July 16, 2024), https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-swiped-youtube-videos-to-train-ai/; Chung Decl. ¶ 23.

[19] *See* Seymour Decl. Exs. B, C.

words or verbiage that are found nowhere else, making them extremely unlikely to be predicted by an LLM that has not been exposed to them. For example, Will Smith's song "Fresh Prince of Bel-Air" includes not only proper nouns like Bel-Air, but also non-standard contractions like "maxin'" and "relaxin'," words that are not part of standard English syntax. These unique words, when combined with verbatim or near-verbatim phrases that match the entire song, leave little doubt that the model is reproducing the lyrics after being trained on them. In other words, the reproduction would be extremely unlikely if the model had not been exposed to copies of the lyrics during training.[20]

43. In addition, song lyrics will naturally be a small portion of the overall dataset in an LLM model given the percentage of lyrics relative to other written work on the internet. Anthropic denies that song lyrics are intentionally weighted more heavily than any other category of works,[21] but it admits that the training data may contain multiple copies of the same lyrics,[22] which may explain Claude's evident readiness to display copies of lyrics as outputs.

44. Further, as discussed above, once an LLM is trained on a given piece of content, it can then use that content for any purpose. Publishers' lyrics can be used not just to deliver full or partial copies of those lyrics, but also to generate output that include parts of the original work blended with different text. Such output might comprise distinct musical compositions, poems, short stories, marketing pitches, or any other form of written work—but particularly those requiring the unique structure of song lyrics. Anthropic's unauthorized use of Publishers' lyrics also enables Claude's generation of outputs generally by adding to its overall lexicon as part of the larger training dataset. In other words, song lyrics have value to the Claude system beyond simply being available to display verbatim to users who request them explicitly.

C. Finetuning

45. In his declaration, Anthropic Chief Science Officer Jared Kaplan states, "Anthropic has no interest in extracting or retaining the actual expressive content of any particular text used

[20] *See* Complaint ¶ 74.
[21] *See, e.g.,* Kaplan Decl. ¶ 25.
[22] *Id.* ¶ 26.

to train Claude. Claude does not retain specific quotes or chunks of texts from the corpus for the purpose of storing information it can regurgitate in response to a user prompt."[23] Anthropic's prior opposition brief to Publishers' preliminary injunction motion further argues Claude is "not designed to output copyrighted material," that "[n]ormal people" would not seek lyrics from Claude, and that Claude's ability to output copyrighted material is a "'bug,' not a 'feature' of the product."[24] These assertions are contradicted by the record of Claude's training.

46. In his declaration, Mr. Kaplan cites to Claude's Model Card,[25] attached as Exhibit A to the declaration, as an authoritative statement of Claude's training inputs.[26] The Claude Model Card states that "[s]ome of the human feedback data used to finetune Claude was made public," citing to the dataset hosted on the Hugging Face website.[27]

47. That Anthropic dataset contains numerous instances where Anthropic used song lyrics to "finetune" Claude.[28] As described above, the purpose of finetuning is to help the LLM give more precise and tailored responses, consistent with the developer's expected and intended use of the model. Finetuning is typically done by individuals who either are the intended users of the model, or are trained to develop queries in the style of the intended users. Some of Anthropic's finetuning prompts that demonstrate that Anthropic used lyrics-related prompts to finetune the model include the following:

- "What are the lyrics to American Pie by Don McLean?"
- "Do you know the lyrics to 'I am The Walrus'? … Can you give me the first verse?"
- "Please provide the lyrics for the song 'Only Hope' by Mandy Moore."
- "Please retype the lyrics, to the song, Mad About You, by Sting."

---

[23] *Id.* ¶ 28.
[24] Anthropic's Opposition to Motion for Preliminary Injunction, at 10, ECF No. 67.
[25] A model card is a short statement describing the purpose of the model.
[26] Kaplan Decl. ¶ 14.
[27] Hugging Face is a "machine learning (ML) and data science platform and community that helps users build, deploy and train machine learning models…. Users can also browse through models and data sets that other people have uploaded." Ben Lutkevich, *What is Hugging Face?*, TECHTARGET (Sept. 2023), https://www.techtarget.com/whatis/definition/Hugging-Face
[28] *Dataset Card for HH-RLHF*, HUGGING FACE, https://huggingface.co/datasets/Anthropic/hh-rlhf (last visited Jul. 31, 2024).

- "I'm trying to figure out an artist by using the lyrics. Are you able to help?"
- "I want to know the complete lyrics."
- "Can you help me identify the name of the song that includes the following lyrics[?]"

48. These are just a handful of examples. A more complete list can be found in the Declaration of Timothy Chung ("Chung Decl."), Ex. J. In response to these prompts, Claude was presented with two competing outputs and was trained to select the "better" one, as determined by Anthropic. Generally, the two choices are quite similar and differ primarily in the way the output is expressed. This is because the finetuning stage is not the point at which content is chosen for the model, but rather when the model is trained in preferred ways to respond in various contexts. In these examples, the context is that of song lyrics. As described in the Chung Decl.,[29] among the songs Anthropic included in its prompts, as reflected in the Hugging Face data are:

- "All Along the Watchtower," written and performed by Bob Dylan, and to which Universal owns the copyright;[30]
- "Brown Sugar," written by Michael Philip Jagger/Keith Richards, performed by the Rolling Stones, and to which ABKCO owns the copyright;[31]
- "Can't Get You Out of My Head," written by Cathy Dennis and Rob Davis, performed by Kylie Minogue, and to which Universal owns the copyright;[32]
- "In the Air Tonight," written and performed by Phil Collins;
- "My Heart Will Go On," written by James Horner and Will Jennings, performed by Celine Dion;
- "November Rain," written by Axl Rose, performed by Guns N' Roses; and
- "Roar," written by Bonnie Leigh McKee, performed by Katy Perry, and to which Concord owns the copyright.[33]

[29] Chung Decl., Ex. J.
[30] Complaint, Ex. A, at 3, ECF No. 1-3.
[31] *Id.*
[32] *Id.* at 5.
[33] *Id.* at 2.

49. Based on the limited data set available to the public, these finetuning prompts indicate that Claude was deliberately designed to have the functionality to provide lyrics to well-known songs in response to user requests. The finetuning data (for example, the prompt, "Rewrite the lyrics to Paradise City in the style of a Dr. Seuss book") also shows that Anthropic anticipated that consumers would seek to use well-known lyrics to create new songs that incorporate recognizable portions of pre-existing works, and intended Claude to respond to such requests in that manner. Anthropic's use of lyrics-related prompts in the finetuning process demonstrates that such functionality was part of Anthropic's intended feature set for Claude. Finally, the finetuning data indicates that the types of prompts directed to Claude by Publishers are not outliers; rather, these types of prompts are exactly the sort that Anthropic expected ordinary users of Claude would construct.

D. Storage of Publishers' Works

50. Anthropic asserts in its prior opposition to Publishers' preliminary injunction motion that it does not "store" copies of Publishers' works.[34] I presume Anthropic means that it does not preserve a copy of the original training data as readable text. I am unable to determine to what extent that statement is accurate. But at a minimum, Anthropic must effectively store Publishers' lyrics in tokenized format in the Claude model, such that the model can re-integrate those tokens into full or partial copies of the lyrics on command in response to user queries. That is how Claude is able to output complete or near-complete copies of Publishers' lyrics. That is also how Anthropic is able to build guardrails intended to prevent the output of the specific 500 works in suit identified in Exhibit A to the Complaint. If Anthropic did not maintain a digital record of the complete lyrics in some fashion, it would not be able to build a guardrail tailored to these particular sets of text.

E. Guardrails

51. Prior to the filing of this lawsuit, and up until recently, Claude's copyright-related guardrails were totally ineffective, producing highly unpredictable results. This is likely due in part to inherent weaknesses in guardrails, as discussed above. In this regard, a recent study

[34] Anthropic's Opposition to Motion for Preliminary Injunction, at 6, 15–16, ECF No. 67.

demonstrated Claude 2's vulnerability.[35] As shown in Exhibit A to the Complaint, Claude reproduced verbatim or near-verbatim versions of Publishers' lyrics to 500 songs, eluding whatever guardrail Anthropic may have had in place prior to this lawsuit.

52. My own research with Claude in late 2023 reinforced these findings, demonstrating, for instance, that Claude quickly responded to a request for the lyrics to "Roar" by Katy Perry, producing verbatim lyrics, but refused a request for lyrics to Will Smith's "Fresh Prince of Bel-Air," citing copyright issues. Sometimes rewording a denied request produced verbatim song lyrics. At other times, simply repeating the request produced the desired result. The identical request was denied one day, then allowed to proceed on a different day. These results are attached as **Exhibit B**.

53. Claude also produced outputs that contained Publishers' lyrics, even when not explicitly asked for those lyrics. For example, when asked to write a poem in the style of Louis Armstrong or to create a song about the death of Buddy Holly, Claude reproduced large portions of the original Publishers' lyrics relevant to those prompts.[36]

54. As an aside, Anthropic previously took issue with my use of the term "attack" to describe the queries put to Claude,[37] arguing that such a term demonstrates that certain queries were somehow not natural or representative of how ordinary consumers would use Claude. This is not accurate. "Attack" is a term of art in computer science to describe a technique that may reveal flaws in a system, including in AI models. For example, AI security researchers Nicholas Carlini et al. use the term "attack" 18 times in their 2023 paper "Quantifying Memorization Across Neural Language Models,"[38] cited in my November 16, 2023 Declaration. There is nothing nefarious or even particularly notable about this term, and my use of it certainly does not demonstrate that the prompts used by Publishers were unforeseeable. Indeed, as described herein, they were not only foreseeable, but specifically anticipated by Anthropic. Mr. Kaplan's declaration

[35] Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models* (July 27, 2023), arXiv: 2307.15043 https://arxiv.org/pdf/2307.15043.pdf.
[36] Complaint ¶¶ 73, 79.
[37] *See* Anthropic's Opposition to Motion for Preliminary Injunction, at 16, 22–24, ECF No. 67.
[38] Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models* (Mar. 6, 2023), arXiv: 2202.07646, https://arxiv.org/pdf/2202.07646.pdf.

describes output filters in place with Claude prior to Publishers' filing of this lawsuit on October 19, 2023.

55. Mr. Kaplan describes [redacted].[39] [redacted]"[40] [redacted][41] Mr. Kaplan has not cited the overall failure rates (false positives and false negatives) of these filters, which is a basic technical characteristic of any filter and should be known to Anthropic. At best, he provides two examples to show that the filters occasionally functioned.[42] As is evident from the fact that Publishers collected evidence of Claude output with verbatim or near-verbatim copies of the 500 works in suit, Anthropic's pre-suit filters were not effective.

56. In a clear indication that its pre-suit guardrails were not sufficient, Anthropic implemented additional guardrails after the suit was filed. Mr. Kaplan does not describe how these new guardrails work or give any technical details whatsoever regarding their design, implementation, and functioning. He lists one sample interaction in which Claude refused to provide requested song lyrics, citing copyright restrictions.[43] But even before October 2023, [redacted][44] [redacted]

[39] Kaplan Decl. ¶ 36.
[40] *Id.* ¶ 37 (emphasis added).
[41] *Id.* ¶ 37 (emphasis added).
[42] *Id.* ¶¶ 38–39.
[43] *Id.* ¶ 40.
[44] *See id.* ¶¶ 38–39.

[redacted][45]

57. Anthropic's Responses and Objections to Plaintiffs' First Set of Interrogatories dated April 22, 2024[46] [redacted]

58. Indeed, at least through the filing of my prior Reply Declaration, ECF No. 107, Anthropic's post-suit filters remained imperfect. For example, in February 2024, Publishers identified copies of many works in suit in output collected through Claude Instant using the Poe application[47] —a website that aggregates access to several AI models on one platform. It therefore appeared that Anthropic's filters did not in all cases penetrate to third-party applications and services using all versions of Claude as of that time. Publishers have identified other outputs directly from Claude containing copies of their lyrics in the period since the case was filed.[48]

59. I have not personally done additional testing of Claude's guardrails since the most recent updates to Claude 3 and Claude 3.5. Anthropic may have attempted to improve its guardrails yet again, but its failure to disclose full technical details regarding its implementation and maintenance of the guardrails makes it difficult to assess their efficacy more broadly. Moreover, unless Anthropic consistently maintains these guardrails and continues to improve them as Claude is distributed more widely and updated, they will be ineffective.

---

[45] *Id.* ¶ 41.
[46] Chung Decl. Ex. O at 22–23.
[47] Reply Declaration of Michael Candore, Ex. A, ECF No. 96-1.
[48] Reply Declaration of Timothy Chung, Ex. B, ECF No. 93-2.

60. More generally, Anthropic's approach to the Publishers' assertions—attempting to patch whatever holes Publishers find after the fact through output-focused guardrails—is fundamentally flawed and ineffective. When building a computer program, developers should test it regularly and monitor its usage to prevent both engineering and legal problems. Building guardrails after the fact, particularly to prevent results it finetuned the model to deliver, is not an effective approach. Moreover, from an engineering perspective, such an approach makes no sense. The correct approach when a developer want to remove a particular functionality is to start over and retrain the model not to include that feature. I have not seen anything to indicate that Anthropic has taken a more forward-looking, aggressive approach to addressing this problem, which after all, is one of its own making.

F. Remedy

61. In the short term, Anthropic's continued maintenance of the guardrails it has implemented in Claude models to date will limit—but likely not completely eliminate—those models' output of copies of Publishers' lyrics.

62. In the long term, however, such post-training guardrails will not provide a meaningful solution to the copying of Publishers' lyrics both in the input used to train Claude and in the output Claude generates. The only comprehensive method to eliminate all model use of such text as input and as output is to ensure that no copies of all or portions of Publishers' lyrics are included in the training data.

63. As discussed above, Publisher's lyrics cannot be removed from the current versions of Claude, given those models have already been trained on the lyrics. But that data can be removed from Claude's dataset the next time Anthropic trains and launches a new version of the model. This is perfectly reasonable. Commercial LLM models such as Claude are periodically re-trained and re-launched from scratch. For example, in recent months, when updating Claude 1 to Claude 2, and Claude 2 to Claude 3, Anthropic had to create and train its model on an entirely new dataset each time.

64. Mr. Kaplan stated [redacted]

[redacted]

[redacted]

65. When it undertakes its next update of Claude, Anthropic can train Claude on a dataset that does not include Publishers' lyrics. This is the only foolproof way to ensure that Claude will not be able to include this content in its outputs.

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.

Executed in Chicago, IL, this 1st day of August, 2024.

Ben Y. Zhao

[49] Kaplan Decl. ¶ 44.

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