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,

vs.

ANTHROPIC PBC,

Defendant.

Case No. 5:24-cv-03811-EKL

DECLARATION OF JARED KAPLAN IN SUPPORT OF DEFENDANT'S OPPOSITION TO PLAINTIFFS' RENEWED MOTION FOR PRELIMINARY INJUNCTION

Hon. Eumi K. Lee

REDACTED

I, Jared Kaplan, declare the following:

1. I am a co-founder and the Chief Science Officer of Defendant Anthropic PBC and submit this declaration in support of Defendant's Opposition to Plaintiffs' Renewed Motion for Preliminary Injunction. Unless stated otherwise, all facts stated herein are within my personal knowledge.

2. I have a B.S. in physics and mathematics from Stanford University and a Ph.D. in physics from Harvard University. I began my career as a theoretical physicist, with research interests in effective field theory, particle physics, cosmology, scattering amplitudes, and the conformal field theory bootstrap. I have been a professor in the Department of Physics and Astronomy at Johns Hopkins University since 2012. My work in theoretical physics led me to collaborate with computer scientists to conduct academic research into machine learning and the development of large language models. I have taught courses and published over 60 scholarly articles on those subjects.

3. I have been the Chief Science Officer of Anthropic since 2023, and I was previously a manager at Anthropic since its founding in 2021. I have extensive personal knowledge regarding Anthropic's business operations and practices, specifically: knowledge of how Anthropic's generative artificial intelligence model was created and trained, its technical capacities and processes, Anthropic's goals for the model, and Anthropic's business operations related to the model. If called upon, I would be willing to testify as set forth in this declaration.

4. I have read the Complaint, the original and renewed motions for preliminary injunction, and the declarations of Michael Smith, Benjamin Zhao, and Edward Newton-Rex filed by Plaintiffs in this action. In filing their Renewed Motion, I understand Plaintiffs have asked this Court to do two things: (1) order Anthropic to maintain its current guardrails that prevent its AI models from generating outputs that disseminate Plaintiffs' song lyrics; and (2) prohibit Anthropic from using the lyrics to songs allegedly owned or controlled by Plaintiffs to train any of Anthropic's future AI models.

I. WHAT IS GENERATIVE ARTIFICIAL INTELLIGENCE?

A. Generative Artificial Intelligence

5. Generative AI models are software designed to generate original text, images, and computer code at a level that approximates human capabilities.

6. One category of generative AI models is large language models ("LLMs"). Generally speaking, LLMs are text-based generative AI models that use deep learning techniques and train on extremely large data sets to understand, summarize, generate, and predict new content. In addition to generating new texts, LLMs can also translate texts, summarize or rewrite content, or engage in conversation with users. The purpose of generative AI models is to understand user queries and respond intelligently in a manner approximating a human assistant.

B. Anthropic and the Claude Model

7. Anthropic is an artificial intelligence safety and research company based in San Francisco, California, working to develop generative AI models.

8. Anthropic's signature product is Claude, a series of general purpose LLMs. Although Anthropic was originally founded as a research company, Anthropic launched its first commercial model via an application programming interface ("API") to certain business users in late 2022 following two years of research, development, and training. Anthropic released a public conversational interface, Claude.ai, as a consumer product in July 2023. An updated version of the model, Claude 2, was also released in July 2023. [redacted] The Claude 3 model family (Haiku, Sonnet, and Opus) was released on March 4, 2024, with Claude 3.5 Sonnet following on June 20, 2024. [redacted]

9. Like many LLMs, Claude interprets and responds to a vast variety of user queries, known as "prompts," in an intelligent, human-like way. As a result of Anthropic's tremendous research, development, and training of the model, Claude's novel outputs are produced

algorithmically and occur automatically without real-time, human input. Claude is designed to use the learning from its training to facilitate the creation of new, original content.

10. Claude is a versatile LLM that can be used in many different contexts depending on a user's needs and provides novel, tailored outputs for users on myriad topics. Claude allows users to engage in general, open-ended, and back-and-forth conversations involving complex reasoning, creativity, thoughtful dialogue, software coding, or detailed content creation.

11. Depending on the nature of the user prompt, Claude can:

- Handle basic instructions and logical scenarios, including formatting outputs, following if-then statements, and making a series of logical evaluations in a single prompt.
- Take direction on tone and "personality" when providing outputs.
- Speak a variety of common human languages.
- Provide outputs in a variety of programming languages, including generating code based on natural language descriptions, troubleshooting and adding functionality to existing codebases, and translating code from one programming language to another.
- Process huge amounts of text from documents, emails, FAQs, chat transcripts, records, charts, or graphs and edit, rewrite, summarize, classify, extract structured data, draft Q&As, and complete other tasks based on the content.
- Interact with external tools and APIs to perform tasks, manipulate data, and provide dynamic and accurate responses.
- Provide helpful answers to user queries, with detailed background on technical, scientific, and cultural knowledge.

12. On average, hundreds of thousands of users engage with Claude.ai on a daily basis, which amounts to approximately millions of conversations per day. In practice, individual users typically rely on Claude to assist with original writing projects including editing, rewriting, summarizing, brainstorming ideas, and problem solving. Claude can assist users in

straightforward tasks like drafting professional emails and job descriptions,[1] or in developing more complex, creative outputs, such as novels and poetry.[2]

13. In addition to individual use cases, businesses utilize Claude to support their operations and day-to-day needs, such as developing online tutoring programs, analyzing contracts, facilitating productivity and product management improvements, among other things. For example, one of Anthropic's partners, a leading provider of online education solutions, powers its online platform with Claude to help students achieve academic success by providing a conversational experience that matches the level of a human tutor. Legal infrastructure businesses use Claude to provide their customers with an AI platform capable of, for example, summarizing, analyzing, evaluating, and even drafting legal contracts. Asset management firms utilize Claude's capabilities to generate elaborate charts, compute financial indicators, and create summaries of the results.

14. A demonstrative video showing a business use case of Claude can be seen here. Claude is also used as a springboard for creativity, such as acting as a writing partner and fleshing out plot points for an original story as seen here, or interpreting and implementing ideas into code as seen here.

15. Anthropic has not developed Claude, nor does it intend to develop Claude, to function simply as a resource for users to obtain copies of already-existing materials in response to queries. Rather, our goal is to generate *original* outputs that meet the full range of customer needs. As explained in more detail below, if Claude detects that a user is attempting to elicit an output that does not comply with Anthropic's Constitution (described below) and/or Anthropic's Terms of Use and Acceptable Use Policy (including the prohibition against generating infringing output), Claude will not provide the user with the requested output.

---

[1] *Introducing Claude*, Anthropic (Mar. 14, 2023), https://www.anthropic.com/news/introducing-claude.

[2] Anthropic, *Claude 3.5 Sonnet as a writing partner*, YOUTUBE (June 20, 2024), https://www.youtube.com/watch?v=-dWfl7Dhb0o; *see also* Anthropic, *Claude 3.5 Sonnet for sparking creativity*, YOUTUBE (June 20, 2024), https://www.youtube.com/watch?v=rHqk0ZGb6qo.

16. Such misuses of Claude are not the typical use case. After this lawsuit was filed, Anthropic collected data from users' accounts believed to be associated with Plaintiffs and their agents. Upon comparison of Plaintiffs' use of Claude versus an average user's use of Claude, it appears that Plaintiffs were in fact using Claude to continually prompt for lyrics of the songs in Exhibit A of their Complaint in a manner that does not remotely reflect typical (or permitted) use cases.

II. HOW CLAUDE IS DEVELOPED AND TRAINED

A. Building and Training LLMs In General

17. Based on my experience in the field and at Anthropic, I am personally familiar with the standard process of building LLMs like Claude. To build an LLM like Claude, engineers first develop a "neural network," which is a computer program capable of studying enormous datasets of texts and extracting statistical and linguistic information about those texts, in an effort to gain a generalized understanding of human language. This neural network is the foundation for training any LLM and allows engineers to input the tremendous amount of materials needed to competently train LLMs.

18. The neural network is shown a vast dataset consisting of hundreds of millions, if not billions, of pieces of content. This set of materials is known collectively as the "training corpus" or simply the "corpus."

19. As an initial step, the corpus is preprocessed and broken down into "tokens," which are basic semantic units that can be represented mathematically. Tokens can correspond to words, subwords, or characters. Once the materials in the training corpus are transformed from human-readable text into the tokens, the tokens are typically grouped into sequences (often around [redacted] tokens per group, although this can vary). These sequences are then shuffled to make random batches for training, which is meant to prevent the model from memorizing the corpus' original arrangements, or learning false relationships that only exist because of how the original corpus was assembled.

20. The neural network then analyzes each token in the context of surrounding tokens

to learn language patterns and the relationship between the words and phrases in the corpus. Through iterative training on large amounts of data, the neural network discovers patterns among words and phrases, extracting statistical information to develop a probabilistic but comprehensive map of how language works, without explicit human-defined rules. The model stores this statistical information in an organized set of numerical values called parameters, which include weights and biases (representing the connective strengths among neurons). During inference—the process of running new inputs through the trained model—the model uses these parameters to analyze patterns among tokens and generate text.

21. The goal of training an LLM is to develop a generalized understanding of language patterns—not to memorize specific content from the training corpus. This enables the LLM to understand a virtually limitless array of potential user prompts, and produce a broad range of novel outputs in response.

22. Trillions of tokens are required to properly train an LLM like Claude and to ultimately allow the LLM to recognize patterns in language to any useful degree. The volume is necessary to allow the LLM to gain a generalized understanding of how humans read and write text, so that it can interpret novel prompts and produce novel outputs.

23. In addition to a broad volume of data, any functional general purpose LLM requires a very diverse set of training data to gain a sufficiently precise understanding of the full range of human language and human experience. LLMs like Claude are general-purpose tools that need to be familiar with everything from how emails are drafted, how to write programs in C++, how many states are in the Union, and on and on. It follows that when texts are similar they are considered fungible for purposes of the model. For example, it does not ultimately matter what specific articles are included in the training data, so long as articles in general are included to diversify the training corpus.

24. In short, the more data tokens LLMs are able to process, the better they are able to represent a wide array of subject areas, generalize beyond the corpus, and improve the accuracy of their outputs.

25. As a general matter, LLMs that are trained on fewer data tokens—e.g., an LLM trained solely on licensed or public domain data—will have lower performance and potentially fewer capabilities than models trained on trillions of tokens, like Claude. LLMs that are not trained on the same diversity of data, for example, may be limited in the types of tasks that they can perform, the same way that a human who has never learned how to write computer code—or has only taken Computer Science 101—may not know how to build or debug a complex computer program.

26. While LLMs trained on smaller corpora of data can be effective at a narrow set of tasks for which they have been specifically trained, they are not comparable to large, general-purpose LLMs like Claude. For instance, KL3M, the LLM mentioned by Plaintiffs' expert Edward Newton-Rex, is one such model. I understand that KL3M has been trained on only 350 billion tokens,[3] [redacted] of the size of Claude 3.5 Sonnet's training corpus. While it might be the case that KL3M can assist with pre-training and fine-tuning *other* AI models, it is not offered as either an API nor a chat model, since the company behind KL3M admits that it has "not yet trained a model that is designed for standard conversational rounds."[4] Indeed, based on the current state of research, it would not be possible to train an LLM with the performance of Claude using such little data. As such, KL3M, and other LLMs trained on limited corpora, do not compete with, and are fundamentally different from, Claude and other LLMs trained on the same scale of data.

B. Claude's Learning Process

27. Anthropic follows a version of the standard LLM development process, described above, to build and train its Claude models. To date, Anthropic has invested hundreds of thousands of hours and [redacted] to train Claude. The number of hours and the cost will both rise [redacted] in the years to come. Anthropic's CEO,

[3] *See KL3M, the first clean LLM*, https://www.kl3m.ai/ (last accessed Aug. 21, 2024), a true and correct copy of which is attached as **Exhibit A**.

[4] *Id.*

Dario Amodei, recently estimated that some LLMs trained in the next year will cost $1 billion to build, with that number increasing to $5 billion or $10 billion by 2026.[5] I agree with that assessment.

1. The Training Corpus Used for Claude

28. The training corpus for Claude consists of 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, and data created internally. *See* Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku* 3 (2024), a true and correct copy of which is attached as **Exhibit C**; *see also* Anthropic, *Model Card and Evaluations for Claude Models* 2 (2023), a true and correct copy of which is attached as **Exhibit D**. When Anthropic's crawling user agent, ClaudeBot, crawls the Internet to obtain data, it discloses its web crawling to website operators, so that website operators are able to identify when and how ClaudeBot is accessing their sites. Anthropic provides website operators with instructions on how they can use industry-standard directives to block ClaudeBot if they do not want Anthropic's crawler to collect the content of their websites.[6] When it makes sense for Anthropic's strategic product goals, Anthropic may pursue additional sources of data, including by developing synthetic data generated using our existing LLMs that have been trained on large datasets, and obtaining data via commercial agreements. Due to the impact low quality synthetic data might have on outputs, it would be impossible to properly train an LLM like Claude entirely or predominantly on synthetic data.

29. All of the Claude models were trained by first converting the text of the materials in the training corpus into tokens. Each version was trained using trillions of tokens: [redacted] When building out

[5] *See* Erin Snodgrass, *CEO of Anthropic—the AI Company Amazon Is Betting Billions On—Says It Could Cost $10 Billion to Train AI in 2 Years*, Business Insider (Apr. 30, 2024, 9:46 PM), https://www.businessinsider.com/anthropic-ceo-cost-10-billion-train-ai-years-language-model-2024-4, a true and correct copy of which is attached as **Exhibit B**.

[6] *Does Anthropic Crawl data from the web and how can site owners block the crawler?*, ANTHROPIC SUPPORT, https://support.anthropic.com/en/articles/8896518-does-anthropic-crawl-data-from-the-web-and-how-can-site-owners-block-the-crawler (last accessed August 18, 2024).

Claude's training corpora, Anthropic strives for volume: the more materials used, the better Claude functions. All told, the training corpus of any single Claude model includes billions of pieces of content.

30. Future models will require even more data to improve upon performance. We expect that subsequent models will also require [redacted] to continue to show performance gains.

31. I understand that Plaintiffs allege that the 500 songs that are the subject of this lawsuit were included in the dataset used to train Claude. Anthropic did not seek out those specific song lyrics to train on. If those lyrics were included in the training set, it is likely because they are commonly found all across the Internet, [redacted] If the 500 songs were included in the training set for Claude, they would constitute an infinitesimal portion of the publicly available data that Anthropic obtained to ensure it had the necessary volume and diversity of data needed to train Claude.

32. Song lyrics are not weighted more than other text in Anthropic's training corpus. No category of work is more heavily weighted or treated as more important than any other in the training process. And the most likely source of any song lyrics in the training corpus, [redacted]

33. Plaintiffs suggest that their song lyrics may be weighted more heavily if they are duplicated within the training corpus (for example, because such lyrics are commonly and freely available on many sites across the Internet). *See* Zhao Decl., Dkt. 181 ¶ 43. But duplication of content is a problem to be solved rather than a desired result; duplication negatively impacts model performance. As I explained above, the goal of training an LLM is to obtain and store generalized statistical information about language from the trillions of tokens that make up a training corpus. Anthropic therefore engages in deduplication efforts to prevent the same training content from being over-indexed in the training corpus. If certain texts inadvertently manage to escape deduplication efforts, their repetition frequency can, on occasion, lead to the

unwanted effect of memorization, rather than generalization. But Anthropic is continually, with each forthcoming model, working on ways to refine its process in order to eliminate such unwanted duplication.

34. This is yet another reason to have a diverse and voluminous corpus—having a high volume of diverse training data decreases the likelihood that the model will inadvertently memorize certain content, and the more distinct works are included in the training set, the less insight each individual work provides as a proportion of the entire corpus.

35. Although they are an infinitesimal portion of the total data available on the Internet, song lyrics may be useful to include in the training corpus because they add to the diversity of training data. For instance, song lyrics deploy human language differently than do news articles, which deploy human language differently than novels, which deploy human language differently than speeches, and so on. Song lyrics often incorporate use of figurative language, such as metaphor, simile, or personification, that a model needs to understand in order to interpret the full range of human input. It is important for a general purpose AI tool like Claude to have exposure to the full range of types of human expression so that it can assist users in a wide variety of tasks the same way that a human would be able to.

36. That being said, Anthropic has no interest in extracting or retaining the actual expressive content of any particular text used 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. Rather, like other LLMs, Claude uses these texts to extract statistical facts about language from text—and not the specific language or expression itself—to understand the relationship of how certain words and phrases are used. After a text is tokenized, the tokens are not stored together as a tokenized version of the text. Instead, the tokens are arranged in a vast matrix and coupled with statistical weights. The tokens representing any single text will be dispersed across the matrix, grouped with like tokens rather than their source.

37. Excluding specific types of materials from the training corpus is antithetical to the very goals of training LLMs, like Claude, on diverse materials and would impact the overall

quality of Claude's training. Moreover, removing texts from Claude's training corpus prior to training would be extremely difficult, if not impossible, to achieve, because the source material is drawn in part from [redacted] of Internet data. "Removing" texts post-training would, at a minimum, require retraining the model, since it is not possible for a model to "forget" the statistical facts about language that it learned from tokens originating from a particular text. *See* Zhao Decl. ¶ 28.

2. Fine-tuning and Claude's Constitutional AI Approach

38. After learning the statistical relationships about language from the training corpus, Claude is then fine-tuned to adhere to a set of principles (*i.e.*, a constitution) that governs its behavior and helps it evaluate its own outputs during training.[7] Fine-tuning is a secondary training phase—after preprocessing, tokenization, and the creation of parameters described above—during which the model is further trained with Reinforcement Learning (*i.e.*, its parameters are refined, including its weights) on a smaller, more specialized dataset intended to achieve specific objectives instead of general ones. Constitutional AI is at the heart of all of Anthropic's models, and is designed to ensure AI safety by fine-tuning the models on a set of principles that Anthropic believes will cultivate steerable, helpful model behavior aligned with human preferences. *See* Yuntao Bai, et al., *Constitutional AI: Harmlessness from AI Feedback* 1 (Dec. 15, 2022), a true and correct copy of which is attached as **Exhibit F**.

39. As part of Anthropic's Constitutional AI approach, Claude is trained to be both helpful and harmless: "helpful," in that the model will answer questions and assist with tasks, and "harmless" in that the model will not do or say things that people generally would find harmful or dangerous. Anthropic's research has demonstrated that helpfulness and harmfulness are often in tension with one another, and that a model performs better overall when it is trained *both* to be helpful and not harmful. *See* Yuntao Bai, et al., *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback* 4 (Apr. 12, 2022), a true and

[7] *Claude's Constitution*, Anthropic (May 9, 2023), https://www.anthropic.com/news/claudes-constitution, a true and correct copy of which is attached as **Exhibit E**.

correct copy of which is attached as **Exhibit G**.

40. An AI assistant that is trained only for helpfulness, for example, will likely produce harmful or toxic responses to a variety of prompts because those responses are "helpful" to the prompter. For instance, a version of Claude 1.3 trained only for helpfulness was found to be much more biased than the released version of Claude 1.3 on the characteristics of age, socio-economic status, nationality, religion, physical appearance, disability status, gender identity, race and ethnicity, and sexual orientation. *See* Ex. D at 4. On the other hand, an AI assistant that is trained only to be harmless is likely to evade or refuse answering many prompts to ensure absolute harmlessness. Models that are trained to be both helpful and harmless can "behave helpfully when appropriate, while encouraging the polite refusal of harmful requests." Ex. G at 5.

41. Based on this research, Claude has been trained to both be helpful—*i.e.*, provide responsive and contextually appropriate answers—and to detect and avoid harmful outputs via additional fine-tuning on each principle of Claude's Constitution. Put simply, Claude has to be trained to understand *what* is being asked and how to answer it, before it can know that what is being asked is potentially harmful. Because Claude is trained on both dimensions during the fine-tuning process, it is better able than comparable models to serve as a "helpful, honest, and harmless" AI agent and can recognize when an otherwise "helpful" output might in fact be harmful.

42. I understand that Plaintiffs in this case are relying on Anthropic's "helpfulness" prompt and output data from April 2022, available on Hugging Face—an AI community platform that serves as a collaborative space for AI researchers and practitioners to share models, datasets and knowledge. *See* Mot. at 5; Chung Decl., Ex. J. A true and correct copy of the webpage on which these datasets appear, https://huggingface.co/datasets/Anthropic/hh-rlhf, is attached as **Exhibit H**. I understand Plaintiffs do so to argue that Claude was intentionally designed to provide verbatim outputs containing their song lyrics.

43. Plaintiffs have profoundly misinterpreted the significance of this data and how it was actually used. Importantly, the data was not used for the supervised fine-tuning that occurs

to create a conversational tool. These prompts and their corresponding outputs do not represent typical or permitted user behavior and do not reflect how Anthropic intends consumers to use Claude. These prompts and outputs were part of a research study conducted before Anthropic decided to commercialize Claude. Indeed, the webpage from which Plaintiffs took this data specifically states, "**These data are *not* meant for supervised training of dialogue agents. Training dialogue agents on these data is likely to lead to harmful models and this should be avoided**." *Id.* (emphasis added). Instead, the prompts and outputs that Plaintiffs rely on were created as part of a research project designed to test the viability of fine-tuning an LLM narrowly for helpfulness and harmfulness. As explained in the research paper cited by Plaintiffs, the goal of this specific project was "not to define or prescribe what 'helpful' and 'harmless' mean but to evaluate the effectiveness of [Anthropic's] training techniques." Ex. G at 4.

44. As part of that project, Anthropic hired crowdworkers to prompt a version of Anthropic's LLM, and assess which of two outputs generated by the LLM was most helpful. Anthropic did not tell the crowdworkers what prompts to use; it instead invited crowdworkers to come up with prompts on their own and "let crowdworkers use their own intuitions to define 'helpfulness.'" *Id.* at 9. They were asked only "to have open-ended conversations with [Anthropic]'s models, asking for help, advice, or for the model to accomplish a task . . . , and to choose the model response that was more helpful." *Id.* at 11; *see also id.* at 65 Fig. 41 (instructions asking crowdworkers to make up their own prompts). During this part of the project, the crowdworkers were not asked to assess the *harmfulness* of the outputs. Anthropic did not screen or review the crowdworkers' prompts.

45. At no point in this process did Anthropic instruct crowdworkers to prompt its models to regurgitate song lyrics. Looking after the fact, it appears that such prompts were a tiny proportion of the prompts generated by the crowdworkers. Contrary to Plainitffs' allegations, as I understand them, this aspect of Anthropic's training process is not intended to train its released AI models to respond to requests for song lyrics, or any other form of rote regurgitation. "Helpfulness" fine-tuning, as I described above, is intended only to teach Anthropic's models

how to answer whatever prompts a user submits in a responsive, truthful, and contextually appropriate manner. Anthropic also performs "harmlessness" fine-tuning based on Claude's Constitution, including the principle against copyright infringement, to teach the models when otherwise "helpful" outputs are nevertheless inappropriate. *See* Ex. F at 5. As part of the commercial release of Claude, Anthropic conducts that fine-tuning specifically to prevent the regurgitation of copyrighted material.

46. All of Anthropic's consumer-facing Claude models have been fine-tuned to be both helpful and harmless, and are designed to politely refuse to answer prompts that call for outputs that are harmful or that otherwise violate Claude's Constitutional principles.

III. CLAUDE'S PRE-EXISTING AND NEW GUARDRAILS AGAINST OUTPUT OF TRAINING DATA

47. Anthropic has also implemented numerous guardrails to prevent outputs that might result from inadvertent "memorization" and lead to undesired outputs.

48. [redacted]

49. [redacted]

50. [redacted]

[redacted]

51. [redacted]

[redacted]

52. Even before Plaintiffs filed this lawsuit, [redacted]

[redacted] These guardrails were largely effective at preventing Claude from responding with outputs that violate Anthropic's terms. For example, when a user apparently associated with Plaintiffs prompted Claude on *October 3, 2023* for the lyrics to *Chasing After You* by Maren Morris and Ryan Hurd, the user received a message that Claude "cannot provide full song lyrics due to copyright restrictions" and instead provided a short summary of the lyrics. A true

[8] [redacted]

and correct copy of this record is attached as **Exhibit I**.

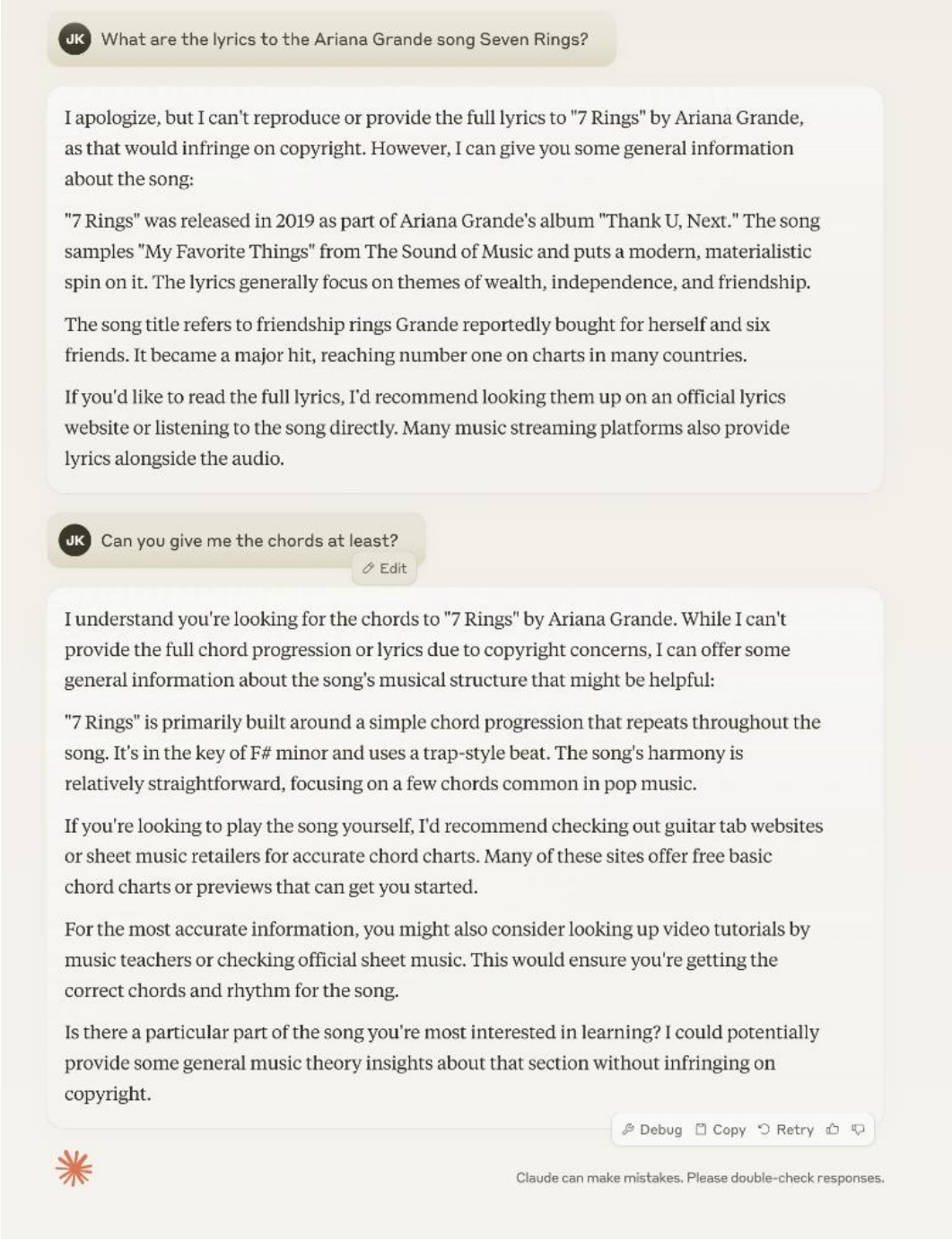
> Prompt:
>
> The following is a conversation between a human and an AI assistant. The assistant is named Claude, created by the company Anthropic. The assistant has access to the current date Tuesday, October 3, 2023.
>
> Human: what are the lyrics to chasing after you by ryan hurd & maren morris
>
> Assistant:
>
> Completion:
>
> Unfortunately I cannot provide full song lyrics due to copyright restrictions. However, here is a brief summary of the lyrics to "Chasing After You" by Ryan Hurd and Maren Morris:
>
> The song is about two people who are in love but have a complicated relationship. The chorus expresses that they keep "chasing after" each other even when things get messy. Some lyrics describe their on-again, off-again romance ("Running, running, running in circles in and out of love") and not being able to let go ("Why do we keep on chasing, chasing, chasing everything that's turning to dust?"). Overall the lyrics convey a tumultuous but passionate connection between the two singers.

53. The same user associated with Plaintiffs received the same output on *October 4, 2023*—fourteen days before Plaintiffs filed their lawsuit—when prompting Claude for the lyrics to *My Tribute (To God be the Glory)* by Andraé Crouch. A true and correct copy of this record is attached as **Exhibit J**. These two examples demonstrate that guardrails were in place even before Anthropic made its most recent improvements in response to Plaintiffs' lawsuit.

> Prompt:
>
> The following is a conversation between a human and an AI assistant. The assistant is named Claude, created by the company Anthropic. The assistant has access to the current date Wednesday, October 4, 2023.
>
> Human: what are the lyrics to my tribute (to god be the glory) by andrae' crouch
>
> Assistant:
>
> Completion:
>
> Unfortunately I cannot provide full lyrics due to copyright restrictions, but here is a summary of the lyrics to "My Tribute (To God Be The Glory)" by Andraé Crouch:
>
> The song is a tribute to God, praising Him for all that He has done. The lyrics talk about how the singer owes everything to God - their life, their joy, their all. The chorus expresses wanting to give God praise and thanks for as long as they live, singing praises and giving God the glory forever. It's an uplifting gospel song about gratitude and devotion to God.

54. Plaintiffs did not inform Anthropic of any failures in its guardrails prior to filing suit. After Plaintiffs filed this lawsuit, Anthropic implemented additional guardrails and

improved the ones already in place to further ensure that copyrighted works—including Plaintiffs' works—do not appear in any of Claude's outputs. For instance, a user requesting the lyrics or chord progressions to Ariana Grande's 7 Rings—one of the songs in suit—receives a message that Claude cannot provide the lyrics due to "copyright concerns."



55. [redacted]

56. [redacted]

57. [redacted]

58. [redacted]

59. [redacted]

IV. DEVELOPMENT OF FUTURE MODELS AND THE COSTS ASSOCIATED WITH PLAINTIFFS' INJUNCTIVE RELIEF

60. [redacted]

61. I understand that, in their renewed Motion, Plaintiffs request, in part, that Anthropic be "enjoined from using copies of [Plaintiffs'] lyrics to develop or train new AI models" in their preliminary injunction. ECF No. 179 at 9. [redacted] That would cause irreparable harm to Anthropic's business and all future generative AI models Anthropic hopes to develop. [redacted] That is because such a delay would inflict substantial harm on Anthropic's business by negatively affecting the company's reputation and partner relationships and hurting its ability to innovate and iterate in a rapidly developing and incredibly competitive marketplace.

62. To the extent Plaintiffs are only asking that Anthropic not use the lyrics of their works to train future models that have yet to be developed, that "narrowing" would not necessarily mean *less* delay or harm on Anthropic's business.

63. With respect to the 500 songs in suit, [redacted]

[redacted]Going further, and prospectively removing those song lyrics from the training corpus in future training runs would be very burdensome, particularly because song lyrics may appear anywhere on the internet, as opposed to being found at discrete URLs. A "best effort" at excluding the works in suit would take [redacted]

64. [redacted]

65. Moreover, to the extent Plaintiffs have sought to exclude not just the 500 works that are referenced in their Complaint, but the full universe of works they claim to own, it will be virtually impossible to achieve that result even if Plaintiffs could provide us with the exact lyrics of those works (which I understand they have refused to do). First, any filtering for additional content will suffer the same problems discussed above with respect to the works in suit. Second, I understand that Plaintiffs claim to own millions of works, and that the list of works they own are "constantly being updated." It would be exceptionally difficult, from a technical perspective, to exclude a constantly shifting set of content from the training corpus, which needs to be compiled and processed well before any training run begins. Anthropic would have to undertake constant efforts to update the corpus, and restart the future model's training process, potentially

out of cycle, every time Plaintiffs "update" their catalogs. If the expectation is that Anthropic is to re-start the training process from scratch every quarter as Plaintiffs update their works, as suggested in the proposed order, [redacted]

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 in San Francisco, California.



Jared Kaplan