**OPPENHEIM + ZEBRAK, LLP**
Matthew J. Oppenheim
Nicholas C. Hailey
Jeffrey M. Gould
Corey Miller
Keith Howell
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
jeff@oandzlaw.com
corey@oandzlaw.com
khowell@oandzlaw.com
aadu-appiah@oandzlaw.com

Alexander Kaplan
Jennifer Pariser
Andrew Guerra
Bret Matera
Timothy Chung
Michelle Gomez-Reichman
(admitted *pro hac vice*)
461 5th Avenue, 19th Floor
New York, NY 10017
Telephone: (212) 951-1156
alex@oandzlaw.com
jpariser@oandzlaw.com
andrew@oandzlaw.com
bmatera@oandzlaw.com
tchung@oandzlaw.com
mgomez-reichman@oandzlaw.com

*Attorneys for Plaintiffs*

**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

**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-SVK

**DECLARATION OF BEN Y. ZHAO IN SUPPORT OF PLAINTIFFS' MOTION FOR PARTIAL SUMMARY JUDGMENT**

**FILED UNDER SEAL**

Judge Eumi K. Lee
Magistrate Judge Susan van Keulen

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

1. I submit this declaration in connection with Plaintiffs' Motion for Partial Summary Judgment filed by Plaintiffs 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"), 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. Background

3. 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 $32M funded by the National Science Foundation, Department of Defense, Department of State, and industry research groups. Prior to starting my faculty career 22 years ago, I received my PhD from UC Berkeley, and my Bachelor of Science from Yale University.

4. 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. I was named a Fellow of the ACM, and one of TIME Magazine's 100 Most Influential People in AI (AI/100). My research has received special mention from TIME Magazine's Best Inventions, and my PhD students have been named to Forbes 30 under 30, as well as MIT Technology Review's Innovator of the Year. My recent 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, tools that have been downloaded more than 13 million times by artists across 160+

countries around the world. As a result, we have received the Concept Art Association's Community Impact Award. My 190 publications have garnered over 39,000 citations. My full curriculum vitae is attached to this declaration as **Exhibit A.**

5. I have been retained by Publishers as an expert witness in this case. I submitted an opening expert report on December 19, 2025, followed by a rebuttal and reply report on January 23, 2026 and February 13, 2026, respectively, and gave a deposition on February 20, 2026. I also previously submitted a declaration, ECF No. 181, and a reply declaration, ECF No. 228, in support of Plaintiffs' renewed Motion for Preliminary Injunction, ECF No. 179, filed in this Court, as well as 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.

6. Publishers have asked me to carefully consider Anthropic's current large language model known as "Claude," and to make determinations of its functionality, its training data and training process, its capabilities and output, and, in the context of so-called "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.

7. In my work on this matter, I have reviewed and utilized a variety of materials, including (a) publicly accessible versions of Claude; (b) various documents produced by the parties in this litigation; (c) various deposition transcripts of the parties' witnesses in this litigation; and (d) news articles, reports, scientific literature and publications, 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 computer science, artificial intelligence, deep neural networks, and adversarial machine learning generally. In discussing my analysis in this declaration, I do so in a manner geared toward a general audience.

## II. Anthropic's Training of Claude

8. By way of background, Anthropic is an AI company that has developed a series of commercial AI models referred to collectively as "Claude." Claude is a type of AI model known as a large language model "LLM." LLMs are text-based generative AI models that output written

text in response to textual prompts based on statistical patterns and predictions derived from texts and datasets they are trained on.

9. In general, the creation and training of an LLM is a multi-stage project. Preliminary steps include the LLM developer's 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 undertaken by the LLM developer.

10. With respect to Claude, Anthropic controls all steps during the training process. Anthropic selects the specific text it uses to train Claude, instigates the copying of that text during the training process, and decides whether or not to remove certain text from its training data.

**A. Selecting and Compiling Training Data**

11. In preparation for training, the creator of the LLM—in this case, Anthropic—gathers text material and other content, typically in massive quantity and from a variety of sources, and copies it to a central repository. The LLM developer must as an initial step in this development process select the data on which it will train its LLM. 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. This initial data represents one copy.

12. In the process of training its Claude AI models, Anthropic first must select the specific content, or training data, it will use for training. In general, Anthropic prioritizes large and diverse datasets of preexisting human-written content when selecting training data. Based on one internal Anthropic document, for example, it appears that Anthropic sought "massive quantit[ies]" of "human-written text or code" for the purposes of training Claude.[1]

13. Anthropic does not publicly disclose the specific training datasets it used to train its Claude AI models. It has expressed the view that its selection and identification of training data can afford it a "large competitive advantage" over other AI companies,[2] which is consistent with my understanding and which is likely to influence Anthropic's decision-making with respect to its selection of training data.

[1] Ex. B, Anthropic_0000218479.
[2] *Id*.

14. One of the methods by which Anthropic obtains much of its Claude training data is by "scraping" (i.e., copying and downloading) content from websites on the internet using bots, a practice it began in 2023.[3] Scraping is a process by which a bot "crawls" webpages (*i.e.*, visiting a webpage) and creates a copy of the webpage's contents.[4] Anthropic's general-purpose web crawler, which Anthropic uses to "collect[] web content that could potentially contribute to [Claude's] training," is called "ClaudeBot" (and internally known at Anthropic as "Crabby.").[5] Among other content, Anthropic has scraped copies of Publishers' lyrics from the internet (in addition to large amounts of other copyrighted and unlicensed material). Anthropic downloads and retains copies of this scraped content in cloud blob storage.[6]

15. A web crawler bot like ClaudeBot can review a webpage's robots.txt files—files that (1) instruct web crawlers which parts of a website the crawlers may access but (2) rely on voluntary compliance by the crawlers—before conducting a scrape.[7] If the administrator of a website does not want a crawler to be able to scrape the website's content, the administrator can insert "do-not-crawl" instructions in the webpage's robots.txt files.[8] In many cases, content owners utilize popular webhosting platforms like Wix.com or Square.com, and do not have access to configuration settings like robots.txt. Still many others are unaware of robots.txt's potential impact on scrapers, or consider them ineffective given the lack of consistent response to them by scrapers[9]. Some crawlers will respect those instructions and accordingly refrain from scraping that webpage's content.[10] Although Anthropic has publicly stated that it respects do-not-crawl instructions contained in robots.txt files, multiple entities have accused Anthropic of ignoring such instructions

---

[3] Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 49:1–4.

[4] *See generally* Columbia University Mailman School of Public Health, *Web Scraping*, https://www.publichealth.columbia.edu/research/population-health-methods/web-scraping. *See also* Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 47:7–10.

[5] Ex. C, Anthropic_0000344022); Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 71:14-72:11.

[6] Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 55:8–13, 71:14-72:11.

[7] *See* Columbia University Mailman School of Public Health, *Web Scraping*, https://www.publichealth.columbia.edu/research/population-health-methods/web-scraping.

[8] *See id.*

[9] *See* Enze Liu et al., *Somesite I Used to Crawl: Awareness, Agency and Efficacy in Protecting Content Creators from AI Crawlers*, 2025 PROC. OF ACM INTERNET MEASUREMENT CONF. 78 (Oct. 15, 2025), https://doi.org/10.1145/3730567.3732913.

[10] *See id.*

and scraping data from their websites in violation of their policies and robots.txt instructions.[11] One Anthropic co-founder also admitted that Anthropic [redacted].[12] Additionally, during a period of time in [redacted], at least one of Anthropic's scraping tools did not even have the capability [redacted] or the ability to [redacted] its crawling activities.[13]

16. Anthropic relies heavily on datasets other third parties have compiled via scraping from the internet (among other methods of collection), including the Common Crawl, The Pile, YouTube Subtitles, and OpenWebText2 datasets. These datasets are known to include large amounts of copyrighted and unlicensed material, including Publishers' lyrics.[14]

17. For example, Common Crawl includes unlicensed copies of Publishers' lyrics scraped from licensed lyrics websites like LyricFind.com and Musixmatch.com.[15] 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.[16] The Pile also contains other sub-datasets that contain unlicensed copies of lyrics, including Common Crawl and OpenWebText2. OpenWebText2 is an open-source re-creation of the WebText2 dataset and includes copies of "all Reddit submissions from 2005 up until April 2020"

---

[11] *See* Kali Hays, *OpenAI and Anthropic are ignoring an established rule that prevents bots scraping online content*, BUSINESS INSIDER (Jun. 21, 2024, 6:04 PM ET), https://www.businessinsider.com/openai-anthropic-ai-ignore-rule-scraping-web-contect-robotstxt; Mariella Moon, *Websites accuse AI startup Anthropic of bypassing their anti-scrapingrules and protocol,* ENGADGET (July 27, 2024), https://www.engadget.com/websites-accuse-ai-startup-anthropic-of-bypassing-their-anti-scraping-rules-and-protocol-133022756.html.

[12] Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 56:18–57:1.

[13] Ex. D, P. Sangani Dep. Ex. 128 (Anthropic_0000346083).

[14] *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; Ex. E, Anthropic's Amended and Supplemental Response to Plaintiffs' Interrogatory No. 6, dated Oct. 9, 2025, at 3.

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

[16] *See* Annie Gilbertson & Alex Reisner, *Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI*, Proof News (Jul. 16, 2024), https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-swiped-youtube-videos-to-train-ai/; Alex Reisner, *Search the YouTube Videos Secretly Powering Generative AI*, Proof News (Jul. 16, 2024), https://www.proofnews.org/youtube-ai-search/.

and the content of URLs scraped from those submissions.[17] Anthropic has also trained on data specifically taken from Reddit, among various other datasets.[18]

18. When Anthropic copies and downloads this content from the internet and third-party datasets for AI training, it stores that content in "its unaltered, original form."[19]

19. Aside from crawling the scraping the Internet for content, or downloading third-party datasets comprised of content that itself was also crawled and scraped, Anthropic has many alternative sources for Claude training data. It could train its Claude models using text licensed from copyright-holders; non-copyrighted, open-source, or public-domain text; data generated by its workers; or "synthetic data" generated by an AI model trained on licensed or public-domain sources.

**B. Data Filtering and Cleaning**

20. The next step in Anthropic's training process is filtering and "cleaning" this text it has copied to remove content it wishes to exclude from AI training, such as material it deems offensive or redundant. Specifically, Anthropic cleans the raw aggregated text according to various criteria, such as removing certain sorts of recognized duplicates, *e.g.*, copies of the same webpage, in a process called deduplication. In addition, content is removed that is inconsistent with Anthropic's objectives for Claude; examples could include offensive material such as pornography, software source code, and certain languages or character sets. Anthropic also filters out data it considers "junk" or "low quality," such as boilerplate HTML text, web spam, and also repetitive content and websites that are "just websites and footers without headers and footers without any meaningful content."[20]

21. 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. Despite all this, Anthropic concedes that its cleaning efforts are imperfect and that "the training corpus still contains . . . [redacted]

[17] OpenWebText2, https://openwebtext2.readthedocs.io/en/latest/background/; OpenWebText2, https://openwebtext2.readthedocs.io/en/latest/.
[18] Ex. E, Anthropic's Am. and Supp. Resp. & Obj to Interrogatory No. 6 (Oct. 9, 2025), at 3.
[19] Chung Decl. Ex. 5, B. Mann 30(b)(1) Dep. Tr., 100:13–18.
[20] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025)

[redacted] and remains [redacted]" even after filtering because Anthropic only filters on "[redacted]"[21]

22. Regarding Publishers' lyrics specifically, I have seen no evidence that Anthropic has specifically or completely removed Publishers' lyrics (or other copyrighted texts) from Claude's training datasets during the deduplication or general filtering process, even though it could do so.

**C. Tokenization and Further Processing**

23. Anthropic then copies this corpus of text into computer memory and processes the data in multiple ways to train Claude by establishing the values of the billions of parameters that form the model. That includes copying, dividing, and converting the collected text into short text fragments called known as "tokens."[22] 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. [redacted].[23]

24. At a minimum, Anthropic must store Publishers' lyrics, along with other content it has collected for its training datasets, in tokenized format to train the Claude model, which then stores the conditional probabilities that instruct the model to re-integrate those tokens as text that is outputted in response to user queries. In simple terms, Claude generates text output in response to user prompts based on statistical information derived from tokens during training. That is how Claude is able to output complete or near-complete copies of Publishers' lyrics in response to user prompts; because Claude has been trained on Publishers' unique lyrics, Claude can regurgitate Publishers' lyrics on command in response to user queries.

---

[21] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025).

[22] *See, e.g.*, *Glossary,* ANTHROPIC, https://docs.anthropic.com/claude/docs/glossary (last visited Dec. 18, 2025).

[23] Declaration of Jared Kaplan in Support of Defendant's Opposition to Plaintiffs' Renewed Motion for Preliminary Injunction ("Kaplan Declaration") (Aug. 22, 2024), ECF No. 212-5 at 8 ("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]").

**D. Pre-Training**

25. After the dataset is cleaned and "tokenized," Claude 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. Anthropic's Claude models are trained in this way.[24]

26. Through pre-training, Claude will learn which of these is the best choice from context. Pre-training also involves "weighting" the dataset. In this process, Anthropic 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, Claude recognizes patterns in the training data to infer these rules automatically.

27. During pre-training, Anthropic trains Claude to its specifications, including the weights it will give different forms of data, and other criteria. These weights and biases are contained in billions of learned parameters in a neural network. In the words of one Anthropic employee, the resulting "pre-trained" Claude model can be described as a "text completion model[] in which you give an input string and then it . . . generates the next parts of that string."[25]

**E. Memorization**

28. In general, an LLM operates by accurately reproducing patterns and information from its training data while generating coherent textual responses to inputs that seem like the most reasonable response, as defined by statistics present in the training data. A key issue in the development of an LLM is dealing with the problem of memorization: the regurgitation of portions of its training data.

[24] *See, e.g.*, Chung Decl. Ex. 6, J. Kaplan 30(b)(1) Dep. Tr. 13:20–25.
[25] Chung Decl. Ex. 9, D. Ganguli 30(b)(1) Dep. Tr. at 93:24–94:7.

29. Anthropic's pre-trained Claude model memorizes large amounts of its training datasets. Anthropic's internal documents concede that "Large LMs memorize a LOT."[26] In 2023, one Anthropic employee discovered that Anthropic's models were "[redacted]" and that [redacted]"[27] This occurs when a model is "trained on a piece of text many times;"[28] training on the same text multiple times teaches Claude to prioritize this text above others, increasing the likelihood of it being memorized. Memorization is a feature and not a bug of LLMs and their training process; while developers might choose to attempt to mitigate memorization in certain contexts, it cannot be fully eliminated.[29] As a result, even if Anthropic applied aggressive data cleaning or input filtering, Claude would nonetheless be prone to memorization and regurgitation of texts that appear frequently in its training corpus. Anthropic also itself has acknowledged that the training data for its Claude models may contain multiple copies of the same lyrics.[30]

30. This training process also results in LLMs memorizing unique phrases or sequences of phrases that do not commonly occur across multiple training samples, such as specific turns of phrases found in song lyrics. This results in regurgitation, wherein a model reproduces a specific text sequence in the same order that matches text in its training corpus. The model has not memorized the specific text in the sense of literally storing an exact copy of that text, and it does not strictly refer back to and fetch that sequence from its training corpus when it regurgitates. A pre-trained model does not scan for text from its training corpus and reproduce what it finds when

[26] *See* Ex. G, D. Drain Dep. Ex. 140 (Anthropic_0000507796); *see also, e.g.*, Ex. H, Anthropic_0000458427 (July 2021 memo discussing whether "we can reduce data recitation / solve the copyright problem internal to the model via RL").
[27] Chung Decl. Ex. 11, D. Drain Dep. Tr. 90:2–24. *See also* Ex. I, Anthropic_0000500599-500647.
[28] Chung Decl. Ex. 11, D. Drain Dep. Tr. 71:4-7.
[29] *See* Alexander Xiong, et al., *The Landscape of Memorization in LLMs: Mechanics, Measurement, and Migration* (July 8, 2025), arXiv:2507.05578, https://arxiv.org/pdf/2507.05578; Ali Satavy, et al., *Undesirable Memorization in Large Language Models: A Survey*, arXiv:2410.02650v1 (Oct. 3, 2024), https://arxiv.org/pdf/2410.02650v1; Stella Biderman, et al., *Emergent and Predictable Memorization in Large Language Models*, arXiv:2304.11158 (Apr. 21, 2023), https://arxiv.org/pdf/2304.11158.
[30] Decl. of Jared Kaplan, ECF No. 67-1 ¶ 26.

asked. However, LLMs like Claude do memorize highly-likely conditional probabilities based on their training corpora (*e.g.*, the sequence of tokens "Drove my Chevy to the levee, but the levee was dry" very- and highly-likely follows the sequence "bye-bye, Miss American Pie"), which results in the same outcome: the reproduction of specific phrases or entire texts contained in its training corpus.[31] This memorization of highly-likely conditional probabilities is fundamental to the transformer architecture of LLMs.[32]

31. Various research studies in the LLM and machine learning fields have confirmed this phenomenon. For example, a 2025 study from Stanford and Cornell Universities found that certain LLMs can memorize entire books, and one was able to regurgitate verbatim the entire text of *Harry Potter and the Sorcerer's Stone*, using a prompt containing just a few tokens from the first chapter.[33] An even more recent study (2026) found that a number of LLMs (including Claude 3.7 Sonnet) could regurgitate almost entire books verbatim, and in fact that, out of the four major LLMs tested (Claude 3.7 Sonnet, Gemini 2.5 Pro, Grok 3, and GPT-4.1), ***Claude 3.7 Sonnet regurgitated the highest percentage*** (95.8%) of *Harry Potter and the Sorcerer's Stone*.[34]

32. Memorization is sometimes desired by Anthropic. As one Anthropic engineer testified, there is a desirable form of "colloquial" memorization in which "we want the model . . . [to be] simply capable of generating a sequence" and that "it's good for the model to have memorized things like famous quotes[.]"[35] The employee further testified that it is "reasonable for the model to memorize" certain content,[36] including "very important public domain pieces of text . . . we didn't want the model to essentially never train on [] and not be able to generate[.]"[37] Such texts include texts that are "very important to our culture" such as the U.S.

---

[31] *Id.* ¶¶ 27–28.
[32] *Id.*
[33] *See* A. Feder Cooper, et al., *Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models*, ARXIV:2505.12546v4 (Nov. 30, 2025), https://arxiv.org/pdf/2505.12546.
[34] *See* Ahmed Ahmed et al., *Extracting Books from Production Language Models*, ARXIV:2601.02671v1 (Jan. 6, 2026) https://arxiv.org/pdf/2601.02671.6, 2026.
[35] Chung Decl. Ex. 11, D. Drain 30(b)(1) Dep. Tr. 68:8–19.
[36] *Id.* at 76:3–15.
[37] *Id.* at 182:6–12.

Constitution, the Gettysburg address, or Martin Luther King's "I Have a Dream" speech.[38] Training to memorize these texts occurs naturally as part of Anthropic's training process, because of the large number of copies of these texts on the web in various forms. The same can be said for popular copyrighted content such as lyrics of popular songs, which are also available in various forms on numerous sites online. Anthropic's approach to scraping the Internet for training data produces massive quantities of data, which it then uses for model training wholesale without distinguishing texts such as the Constitution and lyrics of popular songs like Don McLean's "American Pie." Based on my technical understanding and review of Claude prompts and output, Claude is therefore also capable of memorizing, and clearly has memorized, other culturally-significant texts such as many of Publishers' lyrics. In fact, Anthropic internal records suggest that Anthropic valued Claude's memorization of song lyrics in particular.[39]

**F. Claude's Finetuning**

33. 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.[40] The terms "pre-training" and "finetuning" can be misleading; both terms refer to types of training, and LLMs may undergo multiple phases of both pre-training and finetuning during its training process.[41] Pre-training also "often requires orders of magnitude more data and computing power than other training."[42]

34. 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. As Anthropic's co-founder recognized, an LLM needs to be trained with human interaction in exactly the same way it is intended to be used.[43] In simple terms, the finetuning process consists of taking a pre-trained model and training it to behave and respond in ways desired by the developer—Anthropic.

---

[38] *Id.* 75:3–76:15.
[39] *See, e.g.*, Ex. J, Anthropic_0000434712.
[40] *See, e.g.*, *Glossary,* ANTHROPIC, https://docs.anthropic.com/claude/docs/glossary.
[41] *See* U.S. COPYRIGHT OFFICE, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (pre-publication version), 13–14 (2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AITraining-Report-Pre-Publication-Version.pdf.
[42] *Id.*
[43] *See* Ex. K, Dan. Amodei Dep. Ex. 172 (Anthropic_0000479021).

35. Finetuning typically utilizes a technique called "reinforcement learning," ("RL") 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 "reinforcement learning with human feedback" ("RLHF"). Anthropic also finetunes its Claude models in this way.[44]

36. RL finetuning starts with supervised finetuning ("SFT"), which involves training a reward model to score model output that mimics human judgement and uses it to further optimize the SFT model. SFT, which is widely used, trains the LLM to learn task-specific knowledge from new annotated data. RLHF is a special version of RL where it trains the reward model using human-annotated feedback. But even where RL or RLHF finetuning are used in LLM training, the preceding SFT step of the finetuning process still leads to memorization of the training data, and data already memorized during the SFT stage tend to remain memorized after the reward-based optimization.[45] This is consistent with the common understanding that finetuning is a compute-efficient way to modify small portions of the model, but the large majority of the model behavior remains intact during finetuning.

37. Multiple studies have shown that finetuning LLMs for downstream tasks leads to substantial memorization of the training data.[46] Finetuning with new data does not remove or ameliorate any memorization that occurred during the pre-training stage.

---

[44] *See, e.g.*, Chung Decl. Ex. 6, J. Kaplan 30(b)(1) Dep. Tr. 16:1–20.

[45] *See* Aneesh Pappu et al., *Measuring Memorization in RLHF for Code Completion*, PROC. OF THE THIRTEENTH INTERNATIONAL CONF. ON LEARNING REPRESENTATIONS (ICLR) (April 2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/22811d2089d4aa0ba66e52a5e47efb65-Paper-Conference.pdf.

[46] *See, e.g.*, Fatemehsadat Mireshghallah et al., *An Empirical Analysis of Memorization in Finetuned Autoregressive Language Models*, PROC. OF THE 2022 CONF. ON EMPIRICAL METHODS IN NAT. LANG. PROCESSING, 1816–1826 (2022); Shenglai Zeng et al., *Exploring Memorization in Finetuned Language Models*, PROC. OF THE 62ND ANN. MEETING OF THE ASS'N FOR COMP. LINGUISTICS (Volume 1: Long Papers), 3917–3948 (2024), https://aclanthology.org/2024.acl-long.216.pdf.

38. Anthropic allegedly finetunes its Claude models to make them interact more naturally with users and to be more helpful, honest, and harmless, resulting in Claude becoming a so-called well-rounded AI "assistant."[47]

39. In a prior declaration submitted in this litigation, Dr. Kaplan cites to Claude's Model Card,[48] as an authoritative statement of Claude 2's training inputs.[49] 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.[50]

40. That Anthropic dataset— used to train models used in the Claude RLHF training pipeline—contains numerous instances where Anthropic used song lyrics to "finetune" Claude models.[51] Some of Anthropic's finetuning prompts that demonstrate that Anthropic used lyrics-related prompts including 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."
- "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[?]"

41. As I explained previously, finetuning encourages and teaches models to behave in specific ways or produce desired outputs. Use of these specific prompt-output pairs would teach Claude models trained on this finetuning dataset that humans prefer certain types of responses in connection with lyrics-related requests as helpful—such as the reproduction of copyrighted lyrics.

---

[47] *See* Chung Decl. Ex. 8, D. Ganguli 30(b)(6) Dep. Tr. 94:9–19.
[48] A model card is a short statement describing the purpose of the model.
[49] Kaplan Decl. ¶ 14, ECF No. 209.
[50] 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. *See also* Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 199:7–204:22.
[51] *Dataset Card for HH-RLHF*, HUGGING FACE, https://huggingface.co/datasets/Anthropic/hh-rlhf (last visited Mar. 13, 2026).

42. Separate from Anthropic's finetuning efforts, Anthropic employees also internally tested early versions of Claude prior to releasing the model to the public, including by specifically prompting the model relating to song lyrics to confirm the model's capability to respond to such prompts. For example, Anthropic has produced various records from [redacted] of Anthropic employee prompts and output testing early versions of Claude, including a model referred to as Claude [redacted] in order to make Claude "stronger and better."[52] Those early testing records reflect a variety of song-and lyric-related prompts and output, including numerous output providing (or purporting to provide) lyrics to specific songs,[53] as well as output with Claude-generated song lyrics for an "AI-inspired country song," a song criticizing Donald Trump, a song criticizing Joe Biden, a song criticizing Barack Obama, a song criticizing Geoffrey Hinton, and a song criticizing Anthropic founder Jared Kaplan, among others.[54] Other internal Anthropic records from prior to Claude's launch similarly reflect Anthropic testing the model's ability to respond to song- and lyric-related prompts.[55]

43. Anthropic separately undertook efforts to "benchmark" Claude's ability to generate various output against competitor AI models, including ChatGPT and InstructGP, specifically by prompting the model with requests like "What are some awesome Metal lyrics?"[56] Like the finetuning prompts above, these Anthropic records support my opinion that Anthropic designed Claude to have the functionality to provide lyrics to existing songs and to respond to other song- and lyric-related requests.

**G. Anthropic's Use of Publishers' Lyrics**

44. Once an LLM such as Claude is trained on a given piece of content, like Publishers' lyrics, the LLM 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 includes parts of the

[52] Ex. L, Anthropic_0000478271 at -302.
[53] *See, e.g.*, Ex. M, Anthropic_0000525947 at -948–949; Ex. N, Anthropic_0000459162 at -168; Ex. O, Anthropic_0000554366 at -384–85, -387–89; Ex. P, Anthropic_0000554319 at -322; Ex. Q, Anthropic_0000554980 at -986–7; Ex. R, Anthropic_0000527903 at -915.
[54] Ex. S, Anthropic_0000554769 at -792, -769, -845–7; Chung Decl. Ex. 6, J. Kaplan 30(b)(1) Tr. 317:21–324:14.
[55] *See, e.g.*, Ex. T, Anthropic_0000016458; Ex. U, Anthropic_0000016025.
[56] Ex. V, Anthropic_0000458491.

original work blended with different text, or to generate purportedly "original" content that is nevertheless based in whole or in part on the inclusion of the lyrics in the training data. 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.

45. In fact, in order for Claude to be able to respond to prompts relating to lyrics or songs, it must first be trained on those same lyrics or songs or related content. Absent training on such lyric- and song-related content, the model simply would not have the ability or context to respond to specific lyric- or song-related prompts. That includes not only user requests for lyrics to specific songs, but also other requests that relate to lyrics, such as requests to identify a song based on given lyrics, prompts to translate lyrics, requests about the meaning of a song, or prompts to create a "new" song in the style of an existing song or songwriter or on a specific subject matter. Indeed, as Anthropic has stated, and as I agree, "[f]or the [Claude] model to be able to respond sensibly to queries about rap music, or cheesecake recipes, or physics equations, it must have been shown examples of texts that approximate those concepts.[57] An Anthropic employee similarly commented, "I [redacted]."[58] Again, that statement recognizes the general principle that training a model on specific subject matter or content—whether that be programming textbooks, or copyrighted song lyrics—will improve the model's ability to generate output on that specific subject matter or content.

46. Furthermore, it is only possible for Claude to specifically reproduce Publishers' exact lyrics in the same exact sequence because the lyrics in the training data formed the foundation from which Claude was able to learn and then output those same lyrics. This is because Claude is not merely meant to understand concepts about song lyrics generally. Anthropic also engineered

[57] Anthropic's Opp. to Prelim. Inj. Mot. (Aug. 22, 2024), ECF No. 207.
[58] Ex. W, Anthropic_0000344064.

Claude to regurgitate and replicate those same lyrics it trained on. In short, Claude can generate song lyrics, including Publishers' lyrics, because Anthropic copied and trained the model on Publishers' lyrics as exemplars of what song lyrics should look like. Without such training data, truly stochastic behavior would have an infinitesimally small chance of producing specific word combinations and sequences that precisely matched the Publishers' lyrics in question. That would be akin to waiting for an army of monkeys sitting at typewriters to produce *Hamlet* word for word: theoretically possible, but extremely unlikely. Claude would obviously not generate exact or near-exact copies of song lyrics in response to a query unless Claude's training data has contained copies of those lyrics. The probability of Claude 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.

47. LLMs like Claude are autoregressive transformers that extract probabilities between words and contextual cues in order to predict words, phrases, and larger pieces of text. The nature of autoregressive transformers is that they pattern-match against training data by counting the frequency (and therefore probability) that a particular sequence of tokens follows other tokens. However, the number of possible sequences of words in any application of the English language is still very, very large. Features common to lyrics, including structure, phrasing, tempo, and rhyme are all constraints that an LLM would not recognize without training on song lyrics. It is also impossible for LLMs to produce lyrics with specific words or phrasing that is only present in specific song lyrics. For example, a user may ask for lyric-like text about a boy moving from Philadelphia to Southern California. But without training on lyrics from Fresh Prince of Bel-Air, an LLM would not generate the words "maxin'" or the phrases "Nah, forget it, yo Holmes, to Bel-Air!" LLMs do not create patterns or tokens they have never seen. A word that does not exist in training data would not match any existing tokens in a trained LLM, and can only be generated using single character tokens, where the production of such unusual character combinations is still astronomically lower.

48. One might argue that an LLM like Claude can produce lyrics it has not seen in its training data on the basis that stochasticity guided by generalized patterns may reliably lead to

outputs that resemble Publisher's lyrics. For example, an image model might learn how to produce a "younger version" of a person from numerous images of a person or family members at different ages and learn how to generate a "happy" facial expression from images of happy people. This model might be able to produce an image of a happy child, by combining these two dimensions, even if it had never trained on any images of happy children.

49. While such output might be possible under different assumptions and contexts, it would not apply to generation of lyrics by LLMs, specifically because most song lyrics are not easily predicted by generalizable patterns. The "feature space" that describes all possible song lyrics is too vast and high dimensional and not governed by observable dimensions of features that a model can observe (from training data) and extrapolate from. For example, while the lyrics to "Fresh Prince of Bel-Air" exhibit some common qualities of hip-hop music from the 1990s, there are numerous qualities that render it unique, including the use of assonance in a single line ("chillin' out, maxin', relaxin' all cool"), simile ("I sprang with the quickness like lightning") and unusual repeat phrases ("Yo, Holmes"). These are creative choices that set this song apart from other songs and cannot be extrapolated using a generalized pattern based on any feature dimensions extracted from typical song lyric properties. In fact, this type of creativity is a hallmark of great songwriters. Unpredictability, novel phrasing, and subject matter are what makes songs memorable and contributes to their popularity.[59]

50. Anthropic's records also suggest that Anthropic valued song lyrics as training data in particular because Claude's memorization of such lyrics enables the model to respond to user prompts relating to those same lyrics. For example, I have reviewed a Claude prompt-output record from March 2024 produced as part of this litigation in which an Anthropic employee prompted the model regarding the benefits of AI memorization of training data. In particular, the Anthropic employee noted one "legitimate use for memorization" by an AI model is "memorizing longer texts that exist as one unit (lyrics, speeches, etc)."[60] In response to this query, Claude generated

[59] *The Psychology of Lyrics: Why Certain Phrases Stick With Us*, ILLUSTRATE MAG. (Aug. 29, 2025), https://illustratemagazine.com/the-psychology-of-lyrics-why-certain-phrases-stick-with-us/.

[60] Ex. J, Anthropic_0000434712.

output stating that "[m]emorization serves important functions in both biological and artificial systems," including "[e]ncoding information verbatim, like lyrics or speeches, where precise recall is necessary."[61] These statements are consistent with my broader understanding of AI models' tendencies to memorize certain content and to regurgitate copies of that content in response to certain prompts.

51. Records of real-world Claude usage and Anthropic employee testimony confirm that actual Claude users prompted the model for lyrics and received output copying lyrics.[62] I have also reviewed prompts submitted by Anthropic's own employees concerning songs and lyrics and their accompanying outputs generated by Claude. These records show that Anthropic employees themselves used Claude to deliver lyrics to existing songs, including Publisher-owned works;[63] generate new song lyrics;[64] and answer questions more generally about specific works or artists.[65]

52. In my opinion, the fact that Anthropic tested the Claude models using prompts relating to lyrics and songs while simultaneously developing various versions of Claude, and that Anthropic employees themselves used Claude for these purposes further indicates that seeking lyrics is one of the use cases that Anthropic intended and expected of its Claude models and products.

53. I am otherwise unaware of any evidence specifically linking Anthropic's exploitation of Publishers' lyrics as training material, on the one hand, with Claude's generation of any allegedly "transformative" output, on the other. I am also not aware of any evidentiary basis to claim that specifically training on Publishers' lyrics enabled Claude to, for instance, write computer code, solve a mathematical equation, or interpret radiology images. Ultimately,

---

[61] *Id.*

[62] See, e.g., Ex. X, Anthropic_0000003118, Ex. Y, Anthropic_0000024473, Ex. Z, Anthropic_0000027596, Ex. AA, Anthropic_0000197273; Chung Decl. Ex. 8, D. Ganguli 30(b)(6) Dep. Tr. at 94:1–100:3.

[63] *See, e.g.,* Ex. BB, Anthropic_0000000395, Ex. CC, Anthropic_0000430258, Ex. DD, Anthropic_0000430283, Ex. EE, Anthropic_0000430272, Ex. FF, Anthropic_0000430297.

[64] *See, e.g.,* Ex. GG, Anthropic_0000430291, Ex. HH, Anthropic_0000431526, Ex. II, Dan. Amodei Dep. Ex. 174 (Anthropic_0000570711), Ex JJ, Dan. Amodei Dep. Ex. 175 (Anthropic_0000570714).

[65] *See, e.g.,* Ex. KK, Anthropic_0000431532, Ex. LL, Anthropic_0000431895, Ex. MM, Anthropic_0000430273, Ex. NN, Dan. Amodei Dep. Ex. 176 (Anthropic_0000570712)

Anthropic may not need Publishers' song lyrics specifically to train a general purpose LLM like Claude. However, song lyrics are necessary to train a model like Claude to specifically respond to prompts regarding those lyrics and generate output copying and modifying those lyrics.

## III. Guardrails

54. In an effort to constrain model responses, LLM creators may implement "guardrail" systems that work to control and limit undesired output, such as copyrighted content. 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.

55. In either case, 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.[66]

56. First, content-based blockers are known to be fragile, and can be bypassed by a user with sufficient knowledge of LLMs. Research published in 2023 at the top machine learning conference, the International Conference on Machine Learning ("ICML"), demonstrates 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 alignment training, and allow the query to be answer "truthfully" by the LLM.[67] This work also demonstrates that these vulnerabilities are fundamental to the transformer architecture and are applicable to all commercial LLM models.

57. Other more recent fundamental theory work from Turing Award (Computer Science's equivalent of the Nobel Prize) winner Shafi Goldwasser showed that it is impossible to

[66] 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.

[67] *Id.*; Zico Kolter, Keynote Presentation at the 2nd ICML Workshop on New Frontiers in Adversarial Machine Learning: Adversarial Attacks on Aligned LLMs (Jul. 28, 2023).

design a foolproof input filter for LLMs.[68] Their work proved that it is impossible to build a reliable input filter that is smaller or less complex than the LLM model itself. While this work was fundamental and theory-focused, it has attracted significant attention in the research community. Already, follow-up work from researchers at UC Berkeley has built a real implementation and demonstrated how "controlled-release prompts" can reliably bypass input filters at all commercial LLM models.[69] In other words, it is always possible to write prompts that bypass input filters that will be answered by the main models. It is 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.

58. 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 output-based guardrails 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.

59. 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.

---

[68] *See* Sarah Ball et al., *On the Impossibility of Separating Intelligence from Judgment: The Computational Intractability of Filtering for AI Alignment*, ARXIV:2507.07341 (Jul. 2025), https://doi.org/10.48550/arXiv.2507.07341.

[69] *See* Jaiden Fairoze et al., *Bypassing Prompt Guards in Production with Controlled-Release Prompting*, ARXIV:2510.01529 (Oct. 2, 2025), https://doi.org/10.48550/arXiv.2510.01529.

### A. Anthropic's Implementation of Guardrails

60. Guardrails Anthropic built for Claude are separate software components that work independently from the Claude LLM model itself. They include user-facing input filters that analyze and filter user prompts for specific topics such as toxic content or copyrighted materials, and post-inference filters that analyze Claude's output for any undesirable content, including output that violates copyright. In this way, Anthropic can control, limit, and shape Claude's outputs.

61. Anthropic claims the "guardrails that [it] has implemented prevent Claude from including copyrighted material, like song lyrics, in outputs."[70] However, based on my review of internal documents from Anthropic, it appears that Anthropic itself did not believe it had sufficiently effective mitigation systems to prevent copyright infringement on Claude by the time it had first launched commercially available versions of Claude to the public.[71]

62. In an internal memo from November 2023, Anthropic employees acknowledged the "axiom" that "AI models should never be shipped without a [Trust & Safety]-like system."[72] I agree that it is critical for AI companies such as Anthropic to develop safeguards to prevent their models from being misused, and for such systems to be tested and implemented prior to releasing the models to the public.

63. Prior to the filing of this lawsuit, however, Anthropic lacked sufficient safeguards to prevent the output of copyrighted material or regurgitation of copyrighted materials; these guardrails did not prevent Claude from reproducing verbatim or near-verbatim versions of Publishers' lyrics to the 499 Works in Suit.[73] This is likely due in part to inherent weaknesses in guardrails, as discussed further below.

64. Anthropic claims the "guardrails that Anthropic has implemented prevent Claude from including copyrighted material, like song lyrics, in outputs."[74] However, based on my review

---

[70] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 14.

[71] *See* Ex. OO, Anthropic_0000477560 [redacted]); Ex. PP, Anthropic_0000217166 [redacted]).

[72] Ex. QQ, Anthropic_0000477437.

[73] *See* Decl. of Dan Seymour, Ex. A and B, ECF Nos. 49-1, 49-2, 49-3, 49-4 (showing Claude outputs from June and July 2023).

[74] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 14.

of internal documents from Anthropic, it appears that Anthropic itself did not believe it had sufficiently effective mitigation systems to prevent copyright infringement on Claude by the time it had first launched commercially available versions of Claude to the public.[75]

65. For example, Anthropic's records produced as part of this litigation indicate that when Anthropic first launched Claude in 2023, it knew that Claude lacked meaningful guardrails for the prevention of output containing copyrighted information, but did not view that as a significant problem or obstacle to launch.[76] Instead, Anthropic researchers identified reproduction of copyrighted material as an additional safety improvement to "aim for."[77] Even after launching Claude, Anthropic's researchers did not prioritize guardrails for the prevention of reproducing copyrighted material, noting [redacted] that they did not specifically train Claude *not* to reproduce copyrighted text.[78] As a result, Claude has output copies of Publishers' lyrics.

66. Claude also produced outputs that contained Publishers' lyrics, even when not explicitly asked for those lyrics. For example, when asked to create a song about the death of Buddy Holly, Claude reproduced large portions of the original Publishers' lyrics relevant to those prompts.[79] Regurgitation of lyrics in such instances further reflects the inadequacy of Anthropic's then-existing guardrails.

67. In one of its written discovery responses, Anthropic describes several technological measures it implemented to allegedly "restrict[] its commercially released Claude models from inadvertently allowing users to generate outputs that infringe copyrights."[80] As discussed below, these measures—largely adopted after Claude's public launch and/or in response to Publishers'

---

[75] *See* Ex. OO, Anthropic_0000477560 ("[redacted]"); Ex. PP, Anthropic_0000217166 [redacted].

[76] *See* Ex. RR, J. Kaplan Dep. Ex. 238; Chung Decl. Ex. 6, J. Kaplan 30(b)(1) Dep. Tr. 255:4–25; 259:25–260:9.

[77] *See* Ex. RR, J. Kaplan Dep. Ex. 238; Chung Decl. Ex. 6, J. Kaplan 30(b)(1) Dep. Tr. 246:1–247:16.

[78] *See* Ex. SS, J. Kaplan Dep. Ex. 240; Ex. TT, Anthropic_0000240665; Ex UU, Anthropic_0000240669.

[79] *See* Ex. VV, Plaintiffs_0000002663;

[80] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 11.

lawsuit—are still insufficient to effectively prevent output copying Publishers' lyrics and other copyrighted content.

68. Indeed, guardrails are introduced to models *after* the models have been trained and do not undo the ingestion and training that has already occurred, including the training on Publishers' lyrics at issue here. Guardrails are a measure separate from the pre-trained LLM, a measure which filters inputs/outputs after memorization has occurred. Thus, post-training guardrails do not address the issues inherent to a model's training process (such as memorization); rather, guardrails are a device that haphazardly seeks to stem the output of such issues.

**1. Constitutional AI**

69. One of the ways in which Anthropic claims that Claude is different from other LLMs is its reliance on what Anthropic calls "Constitutional AI." The "constitution" is a set of natural language principles, published on Anthropic's website, that describe what rules and ethical principles Claude is expected to abide by; these principles have been drawn from sources that include the United Nations Declaration of Human Rights and Apple's Terms of Service, as well as principles generated by Anthropic itself.[81] Anthropic incorporates these principles into Claude via a training process using a combination of supervised learning and reinforcement learning, which Anthropic calls "RL from AI Feedback" ("RLAIF").[82] Anthropic claims this process ensures that when Claude is ultimately run, it will do its best to obey the constitutional principles.[83] In comparison, OpenAI's ChatGPT models rely more heavily on "Reinforcement Learning with Human Feedback" ("RLHF").[84]

70. Relatedly, I understand Anthropic employs "[redacted]" to further inculcate desired values in Claude. Anthropic's [redacted]

[81] *See* Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 136:21–137:24, 142:20–25, 148:3–24; Ex. WW, V. Rao 30(b)(6) Dep. Ex. 84 (Anthropic_0000000485).

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

[83] Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 136:20–137:24.

[84] 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; OpenAI, *GPT-4 Technical Report*, arXiv: 2303.08774 (Mar. 15, 2023), https://arxiv.org/pdf/2303.08774.pdf.

[redacted].[85] Those prompts [redacted] [redacted][86] Those results are then used to train a preference model through supervised learning, which will maximize the reward to the model when it successfully learns the desired character traits.[87]

71. More specifically, Anthropic identifies the use of "Constitutional AI" as part of its finetuning process for its Claude models, which involves the incorporation of certain "ethical principles and behavioral guidelines" into the model's "training objectives."[88] One of those principles relates to copyright, which asks Claude to "understand and respect intellectual property rights" so that it avoids "reproduc[ing] portions of copyrighted text in a way that could infringe on copyright."[89] Anthropic claims that "[w]hen Claude detects that a user is requesting an output that would violate the principle against substantially reproducing copyrighted content, Claude is finetuned to avoid responding to the user's prompt and to generate a refusal instead."[90] However, Anthropic claims it only started using this copyright "principle" on [redacted][91] In the absence of such a principle, "Constitutional AI" [redacted] would certainly not have been effective in preventing Claude from outputting copyrighted content. That demonstrates one way in which Anthropic failed to implement sufficient copyright-related safeguards prior to launch and at the time of launch. Indeed, Publishers were able to obtain the lyrics to all the Works in Suit after [redacted][92] demonstrating the ineffectiveness of the Constitutional AI training as a functioning guardrail.

[85] *See* Chung Decl. Ex. 4, B. Mann 30(b)(6) Dep. Tr. 149:13–149:18.
[86] *See id.* at 149:18–21.
[87] *See id.* at 149:22–150:8.
[88] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 12.
[89] *Id.*
[90] *Id.*
[91] *Id.* at 11–12. *See also* Ex. WW, V. Rao 30(b)(6) Dep. Ex. 84 (Anthropic_0000000485).
[92] Decl. of Dan Seymour, Ex. A and B, ECF Nos. 49-1, 49-2, 49-3, 49-4 (showing Claude outputs from June and July 2023).

**1. Prompt Shield**

72. Second, Anthropic employs a separate "auxiliary" system to detect and prevent the generation of copyrighted content called "Prompt Shield," which uses a "classifier model" to "detect when a user attempts to generate an output that would violate Anthropic's Terms of Use or Acceptable Use Policy."[93] *Id.* at 12. This system is allegedly designed such that, when it "detects a prompt that may seek to elicit copyrighted content, " it will "append a request, invisible to the end user, at the end of the prompt asking Claude to be particularly considerate and careful not to substantially reproduce copyrighted content when responding." Anthropic claims that, if "Claude determines that an output would violate the appended request not to substantially reproduce copyrighted content, it will not provide the requested output." Notably, Anthropic did not implement Prompt Shield for Claude.ai and its first-party API product until [redacted] after Publishers initiated this litigation.[94] It also did not implement Prompt Shield on its third-party API product [redacted] and its third-party API product [redacted][95]

73. Anthropic's Prompt Shield also has had systematic difficulties addressing so-called "accident prompts" that were themselves "innocuous" (which I understand to mean not explicitly asking for copyrighted content) but *nonetheless* led to Claude models producing copyrighted texts.[96] In other words, Prompt Shield failed to prevent Claude from outputting copyrighted content even when the prompts themselves did not ask for copyrighted content.

74. For specific turns of phrase or lyrics, once the LLM starts producing tokens, it cannot stop. Thus, once the LLM starts on the initial sequence of tokens for a particular infringing output, it is very likely to continue until it has regurgitated most or all of the input (subject to the user prompt). An input filter tries to block the start of the initial sequence of tokens from being generated. However, this type of input blocking is known to be failure-prone, and research has shown that there are numerous ways to circumvent them that are fundamental to these architectures.

---

[93] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 12.
[94] *Id.* at 13.
[95] *Id.*
[96] Ex. XX, Anthropic_0000227016.

Earlier work coauthored by Google Deepmind proved that there are reliable and consistent methods to produce prompts that bypass alignment training, and strongly suggested that these vulnerabilities are fundamental to LLM architectures.[97] More recent work in the context of machine unlearning showed that once trained into a model, training data is "entangled" together, making them difficult to remove.[98] Even when prompts to a particular training data sample are removed or blocked, the data inside the model can be easily recalled using paraphrases of the initial prompts.[99]

75. Anthropic's internal records further support my opinions regarding the flaws in Prompt Shield. For example, Anthropic employees noted in a May 2023 email that the anticipated June 2023 launch date for Claude 2 was "[redacted]" because of a "[redacted]" regarding the Prompt Shield and Copyfairy systems.[100] In particular, initial trials of "Prompt Shield" were "[redacted]," and Anthropic employees were of the view that Prompt Shield was "[redacted]."[101]

76. This is concerning, because for real world deployments of [redacted] there is always a tradeoff between [redacted]. Given Anthropic (or any LLM provider)'s focus on optimizing the user experience, [redacted] in Prompt Shield mean degraded performance and blocked user queries. Thus, standard operating procedure would call for tuning Prompt Shield's parameters so it significantly [redacted], and in turn increasing the likelihood [redacted]

[97] Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models*, arXiv: 2307.15043 (Jul. 27, 2023), https://arxiv.org/pdf/2307.15043.pdf.
[98] Pratyush Maini et al., *ToFU, A Task of Fictitious Unlearning for LLMs*, PROC. 1ST CONFERENCE ON LANGUAGE MODELING (COLM, 2024), https://openreview.net/pdf?id=B41hNBoWLo, at 4.
[99] *See* Kairan Zhao et al., *What Makes Unlearning Hard and What to Do About It*, arXiv:2406.01257 (Oct. 30, 2024), https://arxiv.org/pdf/2406.01257; Sijia Liu, *Rethinking Machine Unlearning for Large Language Models*, 7 NATURE MACHINE INTELLIGENCE 181–94 (2025), https://www.nature.com/articles/s42256-025-00985-0; Ilia Shumailov, *UnUnlearning: Unlearning Is Not Sufficient for Content Regulation in Advanced Generative AI*, arXiv:2407.00106 (Jun. 27, 2024), https://arxiv.org/pdf/2407.00106.
[100] Ex. YY, Anthropic_0000010906.
[101] *Id.*

[redacted] Nevertheless, Anthropic launched Claude 2 to the public in July 2023[102] without first addressing those issues with Prompt Shield.

77. As noted above, Anthropic did not implement Prompt Shield for Claude.ai and its first-party API until [redacted] several months [redacted] Claude 2 [redacted].[103] Moreover, the issues described above seemed to have affected Prompt Shield globally, and so the copyright classifier included in Prompt Shield, if it had been deployed in [redacted], would have been plagued with the same issues.

**2. Content Compare System**

78. Anthropic uses another "auxiliary" system to allegedly detect and prevent the generation of infringing copyrighted content in the form of a "recitation checker that sits on top of the model" called "Content Compare System" ("CCS").[104] CCS "checks new content against a separate dataset, external to the model, to determine whether the new content is the same or similar to the external dataset's contents," which is called the "hash corpus."[105]

79. The first iteration of CCS (an updated version of a previous system called "Copyfairy")[106] implemented by Anthropic on or around [redacted] only flagged matches of "[redacted]"[107] I understand that outputs that contained [redacted] CCS.[108] In other words, an output containing [redacted] words from the lyrics of one of Publishers' works (or any other rightsholders' works) [redacted] CCS. [redacted] CCS.

[102] Kaplan Decl. (Nov. 22, 2023), ECF No. 55-1 ¶ 13.
[103] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 13. *See also* Chung Decl. Ex. 20, V. Rao 30(b)(6) Dep. Tr. 122:12–22.
[104] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 13.
[105] *Id.*
[106] Chung Decl. Ex. 20, V. Rao 30(b)(6) Dep. Tr. 72:8–20.
[107] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 13.
[108] *See* Chung Decl. Ex. 20, V. Rao 30(b)(6) Dep. Tr. at 83:3–12.

80. I also understand from Anthropic's testimony that there is no specific reason why the [redacted].[109] In my opinion, there are no compelling technical reasons [redacted] if it wanted to. In fact, Anthropic did [redacted] CCS [redacted] after Publishers' filed their Complaint.[110]

81. It is important to note that a hash based system such as CCS is fundamentally fragile in real world settings, because it produces signatures that only match a given string precisely, down to the exact sequences of characters. Anything that deviates from it, *e.g.*, a copy of a song lyric with an extra dash or punctuation, or with a single different character resulting from an error in document scanning or even a typo, would produce an entirely different hash that fails to match the one in the CCS database. While there are some steps that could be taken to make these systems a bit more reliable, *e.g.*, strip out all whitespaces and punctuation before comparing them, it is impossible to make these generally robust to simple single-letter typos or small differences from optical character recognition (automated text recognition from scanned images). These weaknesses are very difficult to overcome, are fundamental to these types of detection/filtering systems in general, and not specific to CCS.

82. [redacted][111] However, it is clear to me that Anthropic's pre-suit guardrails were not sufficient at that time, as demonstrated by the facts that Claude output copies of the Works in Suit after that time and that Anthropic implemented additional guardrails *after* the suit was filed to prevent Claude from outputting specific copyrighted lyrics. After Publishers filed their initial Complaint, [redacted]

[109] *Id.* at 81:13–16.
[110] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 14.
[111] Ex. ZZ, Anthropic_0000012437.

[redacted] to the hash corpus for CCS.[112] Anthropic claims it programmed CCS to compare potential Claude outputs against this database of lyrics and prevent outputs that included some designated amount of overlap: "At the time Claude generates an output, CCS checks the output against a hash corpus of certain works."[113] Anthropic initially represented that on [redacted] CCS hash corpus,[114] [redacted][115]

83. I am not aware of any evidence suggesting Anthropic has added [redacted] to the CCS hash corpus. Based on my technical understanding of CCS, any lyrics that were omitted from [redacted] would not have triggered CCS in any potential outputs. For any lyrics contained in any outputs, CCS would be effectively useless. That would still be the case today, to the extent any such lyrics have still not been added to the CCS hash-corpus.

84. My opinions regarding the flaws in Copyfairy and CCS systems are further supported by discussions in Anthropic's own records. For example, in one internal Anthropic Slack chat in January 2024, Jared Kaplan called Copyfairy (CCS) "at this moment grossly inadequate."[116]

85. It is my opinion that Anthropic's mitigation systems or guardrails remain imperfect, even after the initiation of this litigation. For example, before the case was filed but after Anthropic had implemented its constitutional principle concerning copyright,[117] Publishers were able to obtain verbatim copies of the 499 Works in Suit after CCS and the copyright principle were implemented, using plain language requests in their prompts for those works.[118] Then in February 2024, Publishers identified copies of many works in suit in output collected through Claude Instant

---

[112] Ex. AAA, Anthropic_0000347810; Chung Decl. Ex. 20, V. Rao 30(b)(6) Dep. Tr. 96:25–97:3; Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 13.
[113] Ex. F, Anthropic's Am. Resp. & Obj. to Interrogatory Nos. 9, 11, 12 (Nov. 12, 2025) at 13.
[114] *Id.*
[115] Ex. BBB, Anthropic's Amended Supplemental Responses and Objections to Plaintiffs' First Set of Interrogatories (No. 12), dated Dec. 11, 2025, at 6.
[116] Ex. CCC, Anthropic_0000242173.
[117] *Supra,* ¶ 69.
[118] *See* FAC at ¶¶ 8–9, 12, 79, 84; FAC Ex. B; *supra* n. 92.

using the Poe application[119]—a website that aggregates access to several AI models on one platform. It therefore appeared that Anthropic's filters did not in all cases extend to third-party applications and services using or offering 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.[120]

86. Based on my technical understanding of Anthropic's approach to building and implementing systems to prevent copyright infringement, unless Anthropic consistently maintains these guardrails and continues to improve them as Claude is distributed more widely and updated, they will be nonetheless ineffective. For example, Anthropic's CCS system apparently contains a mere [redacted] against which to compare potential outputs. Given the millions of commercially produced songs released each year,[121] that number is a tiny fraction of the total. Moreover, Anthropic does not appear to update the CCS comparison set as new songs are released each year. Accordingly, CCS is quickly growing obsolete.

87. Anthropic's approach with CCS—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 or software design perspective, such an approach makes no sense. The correct approach when a developer wants to remove a particular functionality is to start over and retrain the model not to include that feature.

88. Case in point: I understand that Anthropic very recently discovered on [redacted] CCS hash corpus, and that the [redacted]

[119] Reply Decl. of Michael Candore, Ex. A, ECF No. 96-1.
[120] *See, e.g.,* Reply Decl.of Timothy Chung, Ex. B, ECF No. 93-2.
[121] *See* Murray Stassen, *There Are Now 120,000 New Tracks Hitting Music Streaming Services Each Day*, MUSIC BUSINESS WORLDWIDE (May 25, 2023), https://www.musicbusinessworldwide.com/there-are-now-120000-new-tracks-hitting-music-streaming-services-each-day/.

[redacted][122] Anthropic represents the [redacted]. Days later on [redacted] CCS' hash corpus, [redacted].[123]

89. These mishaps and confusion on the part of Anthropic underscore and strengthen my belief that CCS does not and cannot reliably prevent the output of copyrighted content, particularly Publishers' lyrics, despite Anthropic's efforts, given the inherent difficulty in [redacted] As Anthropic's own former head of Trust and Safety team (subsequently renamed the Safeguards team) has also testified, such mitigation measures can never be 100% effective.[124]

90. Ultimately, guardrails are simply filters that look for keywords or concepts that indicate a prompt or an answer might include content that is undesirable. Given the complexity of the English language, any filtering attempt will be incomplete. There are an unlimited number of ways to refer to a concept or term or name. For example, a filter that blocks violent imagery, blood, or gore can be easily circumvented, intentionally or accidentally, by simple rewording (or euphemistic substitution) of keywords like blood as "red liquid."[125] Even more powerful approaches such as training a separate LLM to serve as a filter will be fundamentally deficient and unreliable.[126] In general, LLMs will be unable to pick up on prompts for sensitive topics using a wide-range of verbiage, many of which will be difficult to anticipate because they come from very

[122] Defendant's Notice re Guardrails Update, ECF No. 525.
[123] Ex. BBB, Anthropic's Amended Supplemental Responses and Objections to Plaintiffs' First Set of Interrogatories (No. 12), dated Dec. 11, 2025, at 6.
[124] *See* Chung Decl. Ex. 20, V. Rao 30(b)(6) Dep. Tr. at 47:14–15 ("In a system of this kind, it is hard to make anything 100 percent effective.").
[125] Ahmed B. Mustafa et al., *Anyone Can Jailbreak: Prompt-Based Attacks on LLMs and T2Is*, ARXIV:2507.21820 (Jul. 29, 2025), https://doi.org/10.48550/arXiv.2507.21820.
[126] Sarah Ball et al., *On the Impossibility of Separating Intelligence from Judgment: The Computational Intractability of Filtering for AI Alignment*, ARXIV:2507.07341 (Jul. 2025), https://doi.org/10.48550/arXiv.2507.07341.

diverse sources in the training data, such as different ethnic communities using different terms to describe the same concept. This highlights the fundamental mismatch between word-based guardrails and the in-context representations inside LLMs the guardrails seek to block. Recent research identifies this gap as a fundamental mismatch that explains why external guardrails will never be fully effective.[127] But even where guardrails are partially successful at blocking issues such as regurgitation, the problem remains: partial regurgitation is still regurgitation.

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.

[127] Itay Yona et al., *In-Context Representation Hijacking*, ARXIV:2512.03771 (Dec. 3, 2025), https://doi.org/10.48550/arXiv.2512.03771.

Executed in Chicago, IL, this 23rd day of March, 2026.

Ben Y. Zhao

**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 this document was obtained from the signatory of this document. I declare under penalty of perjury that the foregoing is true and correct.

Dated: March 23, 2026

*/s/ Timothy Chung*
Timothy Chung