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 SHAWN SHAN IN SUPPORT OF PLAINTIFFS' MOTION FOR PARTIAL SUMMARY JUDGMENT

FILED UNDER SEAL

Judge Eumi K. Lee
Magistrate Judge Susan van Keulen

I, Shawn Shan, 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 a visiting researcher and an incoming Assistant Professor of Computer Science at Dartmouth College. I will begin my appointment in July 2026. I received my Ph.D. in Computer Science from the University of Chicago in 2025. I also received Master of Science and Bachelor of Science degrees in Computer Science from the University of Chicago.

4. My research examines the technical foundations of artificial intelligence ("AI") systems, with a focus on evaluating their behavior, identifying limitations, and developing methods for AI security and accountability. My work involves deep technical investigations into generative models and their interaction with training data. My specific contributions include the following:

a. Analyzing Model Behavior and Provenance: Investigating the ability to identify the provenance of generative AI models and developing techniques to distinguish human-generated content from AI-generated content. I have also studied internal model phenomena, such as understanding "implosion" in the attention layers of generative models.

b. Developing Defensive Mechanisms: Leading the creation of Glaze, a tool that applies small, targeted perturbations to content to disrupt the ability of AI models to learn styles from training data.

c. Analyzing AI Data Ecosystem: I also study the efficacy of methods content creators can use to manage and control AI crawler access to their works, for example,

through protocols like robots.txt.

d. Other AI Security Research: My work extends to other areas of AI security, including protecting personal privacy against deep learning models (Fawkes), methods for identifying and mitigating unusual behavior in neural networks (Neural Cleanse), and developing defenses against query-based black-box attacks (Blacklight).

5. I have authored 18 peer reviewed publications on various subjects within computer science, including numerous articles on AI-related topics. They have garnered roughly 3,800 citations. I have also given numerous invited talks at conferences and symposia.

6. My expertise in AI data acquisition is reflected in, among other places, my recent publication, "*Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers*."[1] This work provides one of the first comprehensive technical analyses of the "crawling" infrastructure used by AI companies to scrape training data from the internet. In this study, my team and I conducted a longitudinal analysis of over 50,000 popular websites to measure the adoption of robots.txt protocols designed to manage AI crawlers. I performed both passive and active technical measurements to determine which commercial AI crawlers respect these blocking directives and which ignore them. This work directly informs my understanding of how AI models ingest data and the technical measures available to control that ingestion.

7. My research also extensively examines the security and robustness of AI models against adversarial exploitation, providing me with the technical expertise to evaluate the efficacy of "guardrails" and safety filters. I have authored multiple peer-reviewed studies on identifying and mitigating vulnerabilities where actors attempt to manipulate model behavior or bypass intended restrictions. For instance, my work on "Blacklight" developed scalable defenses against "query-based black-box attacks," where an outsider attempts to extract information or manipulate a model solely through its inputs and outputs. Similarly, my research on "Neural Cleanse" (one of

[1] Enze Liu et al., *Somesite I Used to Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers*, ARXIV: 2411.15091v2 (May 7, 2025), https://arxiv.org/pdf/2411.15091.

the most cited publications in AI security with over 2,100 citations) focused on identifying "backdoors" and hidden vulnerabilities in neural networks.

8. I am a member of the Glaze Project at the University of Chicago. The Project develops tools that assist various creators in protecting their work from unauthorized exploitation by generative AI. To date, the Glaze Project has developed several technologies primarily designed to protect visual arts, including Glaze, WebGlaze, and others. All of our tools are free for use by artists.

9. My AI-related work and research has been featured in major news outlets, including *The New York Times*, BBC, *The Wall Street Journal*, and NPR. I have received numerous awards in connection with my work, including *Forbes* 30 Under 30, the *MIT Technology Review* Innovator of the Year, and *MIT Technology Review* 35 Innovators Under 35. My research has also earned the USENIX Internet Defense Prize and several Distinguished Paper Awards at top conferences.

10. I have extensive experience with code as a software engineer. Over the course of my academic and industry career, I have written, reviewed, and debugged tens of thousands of lines of code related to AI systems, including large-scale data collection and preprocessing, model training pipelines, evaluation and debugging workflows, inference systems, and safety and guardrail mechanisms. I have trained and evaluated over 2,000 AI models across a range of architectures and applications, including image classifiers, text-to-image diffusion models, large language models ("LLMs"), and malware detection systems. I am involved in the DARPA GARD (Guaranteeing AI Robustness Against Deception) program as a primary evaluator, where I review and test research teams' AI models and systems for robustness and failure modes. I also regularly review and evaluate code written by others, including mentoring undergraduate and graduate students on research and course projects by reviewing their code for correctness, efficiency, reproducibility, and security, and providing technical guidance on software design and implementation.

11. I previously worked at Meta from June 2019 to August 2019 and June 2020 to August 2020, where I was responsible for analyzing AI models in production and training pipelines in the context of protecting Meta platforms from malicious scraping. The models I trained at Meta

were not LLMs, but more traditional machine learning classifiers, which are discriminative AI models designed to categorize inputs into predefined labels (e.g., "malicious" or "benign"), as opposed to generative models like LLMs that synthesize new content by predicting and assembling sequences of text. The data used in those training pipelines were comprised of non-copyrighted material. In that capacity, I trained AI models and wrote thousands of lines of code that were deployed into production. Specifically, I was a member of Meta's Privacy team, which focused on protecting user privacy on the platform.

12. A detailed curriculum vitae showing my qualifications, experience, publications, awards and patents is attached to this report as **Exhibit A.**

13. I have been retained by Publishers to serve as an independent expert witness relating to their lawsuit against Anthropic. My opinions and analyses are detailed in full in the expert reports that I authored in this case, including my opening report dated December 19, 2025, my rebuttal report dated January 23, 2026, and my reply report dated February 13, 2026.

14. 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; (d) news articles, reports, and other publicly available information regarding Anthropic, Claude, and other AI models and related issues; and (e) source code for various Claude models and its training datasets. I am also relying on my experience, training, and background, as described above.

II. Inspection of Claude Training Data

15. Publishers asked me to conduct a series of inspections of certain source code and datasets relating to Anthropic's "Claude" AI models that Anthropic made available for inspection as part of this litigation. As part of these inspections, I was asked to analyze this data and report on my observations regarding, among other things, the composition and content of the datasets used to train Anthropic's AI models, including the presence of Plaintiffs' Works in Suit (the 499 works identified in Exhibit A to Publishers' First Amended Complaint, ECF No. 337-1) (hereinafter "Works in Suit") in those training datasets.

16. To perform this inspection and analysis, I relied on my knowledge, expertise, and research as a computer scientist working with AI systems and training data, as detailed above.

A. Overview of Training Data

17. Text-generative LLMs like Claude are a class of generative AI that are trained on and mimic natural human language in their written outputs.[2] Broadly speaking, generative AI refers to AI systems that analyze existing data, such as audio, images, or text, and generate content based on learned patterns derived from the datasets they are trained on. LLMs like Claude analyze the statistical relationship or conditional probability of tokenized words and phrases and are taught to generate written outputs based on preceding context.[3]

18. The first step of LLM development is referred to as "training," which occurs in two distinct phases: "pre-training" and "finetuning" (also called "post-training"). Pre-training consists of feeding the LLM a large corpus of training materials (the "training corpus") and training it to predict the next token in a given string. Finetuning consists of training the LLM to do specific tasks, often under a supervised learning process. During both stages, human intervention is required. During pre-training and finetuning, the developer must curate the datasets to feed the LLM. During finetuning, the developer also needs to (1) identify the tasks they want the LLM to perform, (2) annotate examples of correct outputs, (3) evaluate the LLM's behavior, (4) reward the model (if employing Reinforcement Learning from Human Feedback finetuning methods), and (5) validate the LLM's performance.[4]

19. LLM developers like Anthropic seek to train LLMs like Claude on a vast training corpus of human-written texts, often comprised of terabytes or petabytes of data and hundreds of billions or trillions of tokens. LLM developers, including Anthropic, typically source large portions of their training corpora by scraping text from webpages,[5] a process by which bots "crawl" through

---

[2] *See, e.g.*, W. Zhao et al., *A Survey of Large Language Models*, ARXIV: 2303.18223v16 (March 11, 2025), https://arxiv.org/pdf/2303.18223.

[3] *See, e.g.*, T. Brown et al., *Language Models Are Few-Shot Learners*, ARXIV: 2005.14165v4 (July 22, 2020), https://arxiv.org/pdf/2005.14165.

[4] *See, e.g.*, L. Ouyang, et al., *Training Language Models to Follow Instructions With Human Feedback*, ARXIV: 2203.02155v1 (March 4, 2022), https://arxiv.org/pdf/2203.02155.

[5] *See* Kevin Roose, *The Data That Powers A.I. is Disappearing Fast*, N.Y. TIMES (Jul. 19, 2024), https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html.

the internet and copy the contents of webpages they encounter.[6] LLM developers, such as Anthropic, may also acquire datasets constructed by third parties, such as "Common Crawl,"[7] "The Pile,"[8] and "OpenWebText2,"[9] which are themselves largely comprised of petabytes of content scraped from the Internet.

B. Claude Training Data Inspection Process and Environment

20. I performed three training data and source code inspections of Anthropic's "Claude" AI models, on September 9-11, 2025, October 2-3, 2025 and October 29, 2025, at Anthropic's office at 394 Pacific Avenue, San Francisco.

21. The inspections were conducted pursuant to the Joint Stipulation Regarding Training Data and Source Code Inspection Protocol in this case (ECF No. 386) ("Inspection Protocol"). As specified in the Inspection Protocol, the inspection environment consisted of two computers in a secure room with no internet connectivity. The first computer connected to a cloud-based storage system hosted by Amazon Web Services containing certain software code and datasets previously requested by Publishers. It was equipped with various Python packages necessary to review that material. Pursuant to the Inspection Protocol, Publishers were permitted with sufficient notice to request and/or supply particular code and files to be made available on this computer for use during the inspection.

22. The second computer supplied by Anthropic in the inspection environment was an iPad with basic functionality that permitted limited note taking. Pursuant to the Inspection Protocol, Publishers were permitted to request electronic copies (referred to as "print requests" in the Inspection Protocol) of limited portions of the code and data, as well as copies of any notes taken on the second computer. If the print request was acceptable to Anthropic, the requested material was produced to Publishers by Anthropic after each inspection concluded.

---

[6] *See generally* UNITED STATES 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-AI-Training-Report-Pre-Publication-Version.pdf.

[7] *Overview*, COMMON CRAWL, https://commoncrawl.org/overview (last accessed Mar. 19, 2025).

[8] *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, THE PILE, https://pile.eleuther.ai/ (last accessed Mar. 19, 2025).

[9] *OpenWebText2*, https://openwebtext2.readthedocs.io/en/latest/ (last accessed Mar. 19, 2025).

23. Prior to the first inspection, I requested through Publishers' counsel that Anthropic provide the follow materials to aid my work:

a. An AWS ["Amazon Web Service"] EC2 instance with at least 64 GB RAM, 16+ CPU cores, and 512 GB local storage, connected to an S3 bucket containing the inspection data.

b. Installation of Python 3.10 or later, along with necessary Python packages for reading and analyzing the data (pyarrow, pandas, numpy, gzip, regex, or any other packages required to read the data).

24. Anthropic confirmed that it would supply the python3 Ubuntu package (at version 3.10.6) and the Python packages pyarrow, pandas, numpy, and regex at the latest versions published to PyPI. I also requested that Anthropic allow our automated script to run for up to five days after the onsite inspection, with the script appending records to a text file that would be submitted for a print request. Anthropic confirmed that it would not interfere with any programs left running for a reasonable period of time. Finally, as permitted by the Inspection Protocol, I sent to Anthropic through Publishers' counsel several text files containing the lyrics to the Works in Suit as well as certain other compositions owned by Publishers, and software code I had written to search for these songs in the training data.

25. During the inspection, Anthropic made available training data corresponding to Claude models from Claude 1 through 3.7 Sonnet (inclusive). This training data was organized into a series of distinct folders, each containing compressed data files. The folders themselves did not always bear self-explanatory names. Anthropic also provided a digital "inventory" mapping file on the inspection computer that correlated specific dataset folders to specific Claude models. This mapping allowed me to distinguish which datasets—ranging from clearly labeled public crawls (*e.g.*, [redacted]") to internal designations (*e.g.*, "[redacted]")—were utilized for which models. I reviewed this inventory to confirm the association between the data and the models; however, Anthropic made clear that this inventory list would not be produced or printed from the secure environment.

26. During the first inspection, I determined that the AWS computer was insufficiently powered with 16 CPU and my scripts would take nearly two weeks to complete. At my request, Anthropic subsequently replaced this computer for one with 64 CPU.

27. The environment during my next two inspections was identical to the first inspection apart from additional text files I sent to Anthropic to be made available for reference during the inspections. Anthropic also made additional Claude code and data available during the subsequent inspections pursuant to negotiation between counsel.

28. To examine and analyze the presence of the Works in Suit in Claude's training data, I developed and executed a Python script designed to scan Anthropic's massive training datasets for the Works in Suit and associated Copyright Management Information (CMI). The methodology proceeded in two technical phases:

29. **Preparation of Reference Data (The "Fingerprints"):** First, I ingested the text files containing the lyrics for all 499 Works in Suit provided by Counsel, which I understand match the "genuine lyrics" to those works as reflected in Exhibit B to Publishers' First Amended Complaint, ECF No. 337-2. I converted these lyrics into a set of "phrases" to create a digital fingerprint for each song. To ensure accuracy and reduce false positives, I applied a specific text-processing steps:

i. Normalization: All text was converted to lowercase and split into a set of individual lyrics (one line of the lyrics) referred to as a "segment."

ii. Filtering: I filtered out extremely short or extremely long segments (keeping only those between 2 and 30 words) and removed non-alphabetic characters from the segments.

iii. Common Phrase Exclusion: I utilized a "stop list" to exclude common segments that might appear across the web but are not unique to the lyrics (e.g., common English phrases such as "one two three four", "no no no no no", or "oh oh oh oh"), ensuring that matches were based on unique lyrical content. I built this list of common phrases by finding the top 50 most common phrases from the list of 499 Works in Suit.

iv. At the end, each song lyrics had a unique "fingerprint," which is a set of normalized segments consisting of English words.

30. **Search and Matching Algorithm:** Second, I then ran this script across Anthropic's training data repositories that were made available: the datasets for multiple Claude models, including Claude 1, 1.1, 1.2, 1.3, 2, 2.1, 3 (including Sonnet, Opus, and Haiku), 3.5 (including Sonnet, Opus, and Haiku) were produced for inspection on September 10, 2025, with additional specific Claude Sonnet 3.7 data produced for the second inspection on October 2, 2025. Because the dataset is too large to scan linearly for every song, I used a targeted "triage" approach followed by a deep verification:

i. Domain Filtering: The script first prioritized records where the URL in metadata contains known lyric-hosting domains (*e.g.*, genius.com, azlyrics.com, musixmatch.com, music.apple.com) and any URL containing the word "lyric" at any part of the URL.

ii. Token Intersection Matching: For records that passed through the domain filter (*i.e.*, records that have URL metadata containing known lyric-hosting domains or URL containing the word "lyric"), the script calculated the overlap between the text in the training data and the "fingerprints" of the Works in Suit. To ensure this process is reproducible and precise, I employed a deterministic set-intersection method defined by the following steps. First, I processed both the original lyrics and the training data using the identical normalization function discussed above. This function converted all text to lowercase, removed non-alphabetic characters (such as punctuation), and split the text into individual segments. This standardization ensures that identical lyrical content is mapped to the same canonical form, allowing the algorithm to correctly identify matches despite minor variations such as missing punctuation, differing capitalization, or irregular whitespace. Additionally, to minimize noise, I retained only those segments containing between 2 and 30 words and not appearing on the "stop list" identified above. Then a match was positively identified only if the intersection of the unique sentences in the

training data file and the unique sentences in the Work in Suit met both of the following strict quantitative thresholds:

a. *Matching Threshold:* A file was counted as a match only if it met two clear requirements. First, the file had to contain at least 3 different lyric segments from the relevant Work in Suit, and each of those segments had to match the original lyrics exactly, character-for-character, after applying the normalization process described above. Second, those matching segments had to represent more than 50% of all the unique lyric segments in the original song. This second requirement ensures that the match reflects copying of most of the song's lyrics, rather than a short quotation, a repeated chorus, or an isolated excerpt.

b. *Manual Verification:* To further validate the accuracy of the automated matching process, I performed a manual review of the results. I randomly selected and visually inspected 20 detected matches from the dataset and confirmed that in every instance, the text identified by the script was indeed a copy of the corresponding Works in Suit. I also specifically targeted the 20 matches with the lowest match ratios (*i.e.*, those closest to the 50% threshold). My inspection confirmed that these were also correct matches. The fact that the match ratio was less than 100% in these instances was attributable to formatting differences or textual variants (*e.g.*, ad-libs, intro/outro text differences) that were not fully standardized by the normalization function, rather than being false positives.

C. Claude Training Data Inspection Findings

31. As a result of my review and analysis of Anthropic's "Claude" AI model training datasets, source code, and other information to date, I arrive at the following conclusions.

32. Anthropic has copied every one of Publishers' 499 Works in Suit in connection with training its Claude AI models, each of them many times over.

33. Specifically, my inspection of Anthropic's Claude training data confirmed the presence of full copies of all 499 Works in Suit in that training data. In total, I identified [redacted] matches to the lyrics of the 499 Works in Suit within Anthropic's Claude training data. Even after removing duplicates where the same webpage was scraped multiple times, I identified [redacted] unique webpages across [redacted] domains containing the lyrics to Publishers' 499 Works in Suit. The largest number of matches corresponded to the Common Crawl dataset. The frequency of reproduction was high. Anthropic's training data contained at least [redacted] or more copies of the lyrics for each of Publishers' 499 Works in Suit. For 354 Works, Anthropic's training data contained over [redacted] copies of each. For 19 Works, Anthropic's training data contained over [redacted] copies of each.

34. Given the nature of Anthropic's data and the way in which I designed my searches to avoid false positives (as described above), these figures are certain to be underinclusive. The automated search was limited to domains known to host song lyrics and to URLs containing the term "lyric." This scope limitation was driven by practical constraints, including the sheer size of the dataset, limited available hardware and compute resources, and the inspection timeframe. Expanding the search to the full corpus would almost certainly identify a substantial number of additional matches.

35. To determine the primary sources of Publishers' Works in Suit within Anthropic's training data, I performed a quantitative analysis of the source URLs associated with the matched content. I utilized a programmatic approach to extract, parse, and aggregate the domain metadata for every confirmed match identified in my previous inspection. This process followed three specific steps:

a. URL Extraction: First, for every instance where a Work in Suit was positively identified in the training dataset, I extracted the corresponding raw URL from the data's metadata.

b. Domain Parsing: To ensure that the counts reflect unique websites rather than specific sub-pages, I parsed each raw URL to isolate its network location. I utilized a standard URL parsing algorithm (specifically the urllib.parse library) to strip away protocol

prefixes (e.g., "http://") and specific file paths, leaving only the root domain name (*e.g.*, converting http://www.lyricsfreak.com/h/hello.html to lyricsfreak.com).

c. Aggregation: I then aggregated the data to calculate the total frequency of each unique domain. I tallied the number of times each domain appeared as the source for a Work in Suit. I filtered out invalid or empty domain records and ranked the remaining domains by frequency to identify the top ten most prolific sources of the Plaintiffs' lyrics.

36. My analysis revealed that a large number of the Publishers' Works in Suit found in Anthropic's training data were copied from a relatively small number of lyric websites. The top 10 lyric websites where Publishers' Works in Suit included in Anthropic's training data appeared, and their respective match counts, are listed below:

Lyric Website	**Number of copies of lyrics to Works in Suit in Anthropic's Claude training data**
Lyricsfreak.com	[redacted]
Lyrics.com	
Lyricsmode.com	
Stlyrics.com	
Lyricstranslate.com	
Sing365.com	
Elyrics.net	
Lyrics-video-music.blogspot.com	
Metrolyrics.com	
Lyrics.net	

37. Certain of these websites (such as pages from Lyricsfreak.com, Lyrics.com, and Lyricsmode.com displaying a specific Work in Suit, for example) indicate on their face that they are sublicensees of LyricFind.com,[10] which I understand that Publishers have licensed to display

[10] *See, e.g.*, *Lynyrd Skynyrd – Sweet Home Alabama Lyrics*, LYRICSFREAK, https://www.lyricsfreak.com/l/lynyrd+skynyrd/sweet+home+alabama_20086081.html (bearing LyricFind logo and Publishers' copyright management information ("CMI"); *Sweet Home Alabama Lynyrd Skynyrd*, LYRICS.COM, https://www.lyrics.com/lyric/8478697/Lynyrd+Skynyrd/Sweet+Home+Alabama (bearing

their lyrics online (including the right to sublicense). Content copied from other of Publishers' licensees, such as Genius.com [redacted] of copies of the Works in Suit found) and Musixmatch.com ([redacted] of copies of the Works in Suit found), also appears in Anthropic's training data.

38. Based on my inspection and understanding of Anthropic's data acquisition practices, it is clear to me that Anthropic acquired copies of Publishers' lyrics by downloading these lyrics via crawling and scraping webpages or by downloading third-party datasets, which are largely comprised of content crawled and scraped from the internet as well.

39. My inspection of Anthropic's Claude training data also confirmed the presence of various third-party datasets in the training data.

40. I understand that Publishers allege and that discovery produced by Anthropic during this case has shown that Anthropic trains its Claude AI models using certain third-party datasets—including Common Crawl, OpenWebText2, The Pile, YouTube Subtitles, and Reddit data—that contain Publishers' lyrics.[11] The findings of my inspection, as described below, support that conclusion.

41. During my inspection of Anthropic's training data repositories, I identified several distinct categories of data used to train Anthropic's Claude models. I navigated the file directory structure of Anthropic's training data repositories and observed that the folder names in those repositories identified in general terms the source and nature of the data contained within. To verify the reliability of these labels, I opened and examined ten random samples from folder with names that indicates hosting possible lyrics. In every instance, the content of the files was consistent with the dataset label on the folder. During the inspection I also examined the data mapping code, which

---

Publishers' CMI and "Lyrics Licensed & Provided by LyricFind); *Lynyrd Skynyrd – Sweet Home Alabama lyrics*, LYRICSMODE, https://www.lyricsmode.com/lyrics/l/lynyrd_skynyrd/sweet_home_alabama.html (bearing LyricFind logo and Publishers' CMI).

[11] *See, e.g.,* FAC, ¶¶ 67, 69, 72; Ex.B, Anthropic's Amended and Supplemental Response to Publishers' Interrogatory No. 6, dated Oct. 9, 2025 (identifying Common Crawl, OpenWebText2, YouTube Subtitles, and Reddit data as sources of training data used by Anthropic); Ex. C, Anthropic_0000016066 (identifying "cc" and "reddit" as training datasets); Ex. D, Anthropic 0000352701 (identifying multiple "cc" datasets, "pile" datasets, "bitbox.openwebtext2," and "reddit.processed" as training data).

indicates which Claude models were trained on which datasets. Because Anthropic objected to producing a copy of the mapping code, I do not describe with specificity herein what datasets are associated with particular Claude models. However, based on my inspection and analysis of Anthropic's training data, as well as the other discovery produced by Anthropic during this case regarding its training datasets, and because all Claude models rely on significant amounts of training data, I believe every Claude model relevant to this case for which Anthropic made training data available for inspection was trained on all or most of the following datasets:[12]

a. Common Crawl: [redacted] of the training data consists of "Common Crawl" data. Common Crawl is a collection of website crawls from 2008 to the present, including raw web pages as well as corresponding metadata and text extractions. My inspection revealed that Anthropic has collected Common Crawl data [redacted] spanning data from the [redacted]. Common Crawl collected these data from 2014 to 2024.

b. The Pile: I observed folders labeled "pile," which refers to "The Pile," a large-scale, open-source dataset commonly used for LLM training.[13] According to the creators of The Pile, it is "composed of 22 diverse and high-quality datasets, including both established natural language processing datasets and several newly introduced ones."[14] The datasets contained in The Pile include, among other datasets, "Pile-CC" (a "Common Crawl-based dataset"), "Books3" ("a dataset of books derived from a copy of the contents of the Bibliotik private tracker"), OpenWebText2 ("a generalized web scrape dataset . . . includ[ing] more recent content from Reddit submissions up until 2020"), and "YouTube

[12] This conclusion is also confirmed by information obtained in discovery including Anthropic's response to Publishers' Interrogatory No. 6 (in which it identified Common Crawl, YouTube subtitles, and Reddit data as the most prevalent sources of lyrics in the training data) and Anthropic's Response to Publishers' Requests for Admission No. 30 (in which it admitted that at least one commercially released Claude model was trained on data from versions of the Pile). *See* Ex. B, Anthropic's Amended and Supplemental Response to Plaintiffs' Interrogatory No. 6, dated Oct. 9, 2025 at 2; Chung Decl. Ex. 21, Anthropic's Response to Publishers' Requests for Admission No. 30.

[13] Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* 2, ARXIV (Dec. 31, 2020), https://arxiv.org/pdf/2101.00027, at 1.

[14] *Id.*

Subtitles" (a "corpus of text gathered from human generated closed captions on YouTube").[15]

c. YouTube Data: I found separate data labeled as "YouTube data." Upon inspecting the sample files, I confirmed that this data consists of text transcriptions of videos.

d. OpenWebText2: I identified folders containing "OpenWebText2," another well-known dataset used for AI training derived from URLs scraped from Reddit submissions.

e. Reddit Data: I identified separate folders containing data scraped from the social media platform Reddit.

42. My findings regarding Anthropic's use of the datasets above to train its Claude AI models are consistent with various Anthropic deposition testimony, documents produced as part of this litigation, and other discovery in this case.[16]

43. Based on my review and analysis of Anthropic's Claude training data, the scale of Anthropic's use of the Works in Suit for AI training is expanding rather than contracting. Based on my automated search script detailed above, the latest model subject to my inspection, Claude 3.7 Sonnet, was overwhelmingly likely to have been trained on larger datasets with more copies of Works in Suit than earlier models.

44. Anthropic explicitly prioritizes and highly values English text—including lyrics—over other data types during training. Anthropic's system configuration assigns specific "weights" to different data categories to influence the model's learning. Based on my inspection, it appears that English data—which encompasses the lyrics found in the training set—[redacted]

[15] *Id.* at 3–5.

[16] *See* Ex. C, Anthropic_0000016066 (identifying "cc" and "reddit" as training datasets); Ex. D, Anthropic 0000352701 (identifying multiple "cc" datasets, "pile" datasets, "bitbox.openwebtext2," and "reddit.processed" as training data); Ex. E, Anthropic_0000240831 (Anthropic employees discussing completed processing of Common Crawl training set); Ex. F, Anthropic_0000342792 (Internal Anthropic email confirming extraction of YouTube captions); Chung Decl. Ex. 5, B. Mann 30(b)(1) Dep. Tr. 39:15–20, 41:6–21, 43:2–12, 47:1–19, 52:10–20; Chung Decl. Ex. 6, J. Kaplan 30(b)(1) Dep. Tr. 143:9–14; Ex. B, Anthropic's Amended and Supplemental Response to Plaintiffs' Interrogatory No. 6, dated Oct. 9, 2025.

[redacted]

[redacted].

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

Shawn Shan