EXHIBIT 1

(FILED UNDER SEAL)

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 Judge Eumi K. Lee Magistrate Judge Susan van Keulen

EXPERT REPORT OF DR. BEN ZHAO

December 19, 2025

Contains Information Designated as
Confidential and/or Highly Confidential—Attorneys' Eyes Only

TABLE OF CONTENTS

I. QUALIFICATIONS, EXPERIENCE, AND COMPENSATION 1

II. SCOPE OF ENGAGEMENT 2

III. SUMMARY OF OPINIONS 3

IV. ANALYSIS AND OPINIONS 4

A. Generative AI 4

1. LLM Training 5

2. LLM Memorization 8

B. Anthropic's Claude AI Models 16

1. Overview 16

2. Anthropic's Training of Claude 18

3. Anthropic's Finetuning and Testing of Claude 27

4. Claude Capabilities and Usage 31

5. Anthropic's Evaluation of Claude 35

6. Guardrails 41

a. Constitutional AI 42

b. Prompt Shield 44

c. Content Compare System 47

d. Defects in Anthropic's Guardrails 49

7. Hallucinations 51

V. LLMS AND COPYRIGHTED CONTENT 53

VI. REMEDY 55

VII. RISKS FROM AI 57

I. Qualifications, Experience, And Compensation

1. 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 in numerous courses related to artificial intelligence. Over the years, I have led numerous projects related to artificial intelligence, including projects funded by the National Science Foundation, Department of Defense, and industry research groups. Prior to starting my faculty career in 2004, I received my PhD from UC Berkeley, and my Bachelor of Science from Yale University.

2. 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 USENIX Internet Defense Prize. My early work created the research area now known as distributed hash tables, and those systems are now used in large cloud computing systems today. More recently, I have spent extensive time researching machine learning and deep learning. My more recent projects have involved research related to the use of content by generative AI models. For my body of work, I have been honored as a Fellow of the Association for Computing Machinery (ACM), a TED/AI speaker, MIT Technology Review Young Innovator, and chosen by TIME Magazine as one of the AI/100, the 100 most influential people in AI. My 190 refereed publications have garnered more than 39,000 citations.

3. A detailed curriculum vitae showing my qualifications, experience, publications, awards and patents, as well as a list of cases in which I have testified at trial or by deposition within the last four years, is attached to this report as **Appendix A**.

4. I am being compensated for my time at the hourly rate of $850 per hour. My compensation does not depend on my conclusions or the outcome of this case. I have no other interest in this litigation or the parties thereto.

II. Scope Of Engagement

5. Publishers have asked me to carefully consider Anthropic's large language model ("LLM") known as Claude and to make assess its functionality, training data, and training processes, and, in the context of guardrails to limit or eliminate the harms of reproducing (in whole or in part) Publishers' lyrics, what mechanisms are possible, what mechanisms are currently deployed, and any limitations of those mechanisms.

6. To perform my analysis—and in addition to drawing from my experience as discussed above and in **Appendix A**—I have reviewed and analyzed extensive information and documents produced by Anthropic through discovery in this action; the parties' various pleadings, disclosures, and discovery responses; deposition transcripts; and other documents produced in this case; in addition to various academic literature and articles. The materials that I have considered in reaching my opinions are cited and discussed in this report and/or listed in **Appendix B** to this report.

7. This report is based on information currently available to me. I reserve the right to supplement my analysis in this report in response to any reports prepared on behalf of Anthropic. I also reserve the right to amend or supplement my opinions based on further discovery and information provided in this case, including any additional depositions or other disclosures that may occur after I submit my report.

8. All of the opinions stated herein are based on my own personal knowledge, professional judgment, and analysis of the materials and information that I considered in preparing this report. I reserve the right to create any additional summaries, tutorials, demonstrations, charts, drawings, tables, and/or animations that may be appropriate to supplement and demonstrate my opinions at trial, deposition, or any hearing at which I may be called to testify.

9. I do not offer any legal opinions regarding Anthropic's actual liability or any of its asserted defenses in this case.

III. Summary Of Opinions

10. My full opinions are set forth in the body of this report. I reserve the right to supplement or modify my opinions as appropriate based on any newly learned information. As a result of my expertise and analysis of the information and documents I have considered, which includes the expert reports and analyses conducted by Dr. Robert Leonard, Dr. Shawn Shan, Dr. Michael Smith, and Joshua Dennis, testimony, documents, and statements from Anthropic and its employees, relevant scholarship and articles, and publicly available information about Anthropic's training data, I have formed the following opinions, summarized below:

- Anthropic has copied the lyrics to a huge number of Publishers' musical compositions and used those copies as input to train Claude.
- Anthropic designed and built its Claude models with the knowledge and intent that they would memorize and output Publishers' lyrics.
- Anthropic's use of Publishers' lyrics during the training process is not critical to the general functionality of Claude, and Anthropic could identify and remove those lyrics from training without undermining Claude's operation.
- Anthropic's use of Publishers' lyrics during the training process, however, is critical and sought after by Anthropic for the specific functionality of Claude outputting Publishers' lyrics verbatim, outputting new song lyrics based on Publishers' lyrics, and responding to user requests about song lyrics and music generally.
- Hallucination behavior by Claude is likely fundamental to the transformer architecture and current methods of training on scraped data, such that Claude will continue to hallucinate outputs including those concerning Publishers' lyrics.
- Anthropic has put in place certain guardrails or technological systems to limit Claude from including Publisher's lyrics in output but those guardrails vary in effectiveness and ultimately depend on Anthropic's willingness to maintain them. In the short run, such guardrails might help stem the inclusion of Publishers' lyrics in Claude's outputs. In the long term, however, guardrails are generally inadequate, and the only foolproof

way to ensure that Claude's outputs do not include Publishers' lyrics is for Anthropic to remove copies of those lyrics from its training material for future AI models.

- Anthropic's guardrails and techological systems do not prevent Claude from outputting new song lyrics trained on Publishers' lyrics or outputting song lyrics based on the style of existing works or songwriters.

IV. Analysis and Opinions

A. Generative AI

11. Artificial intelligence ("AI") systems are built to flexibly perform tasks that involve a wide variety of novel situations and that ordinarily would be expected to require human intelligence to accomplish.

12. "Traditional" (sometimes referred to as "discriminative") AI is designed to perform a specific task in a predictable way in response to user-supplied prompts based on pre-determined rules. These systems mainly fall under the broader category of discriminative classification systems because their primary goal is to recognize complex patterns inherent in multi-faceted data and to produce decisions as a result. For example, a drug-discovery AI system might take as input a new hypothetical chemical compound structure, and based on its training, produce a classification result that predicts whether the compound will have a specific prescriptive impact on a particular disease or pathology. In other examples, a trained deep neural network ("DNN") classifier may take in photographic images of skin lesions and predict which samples are likely cancerous, and a different DNN might take audio samples of a human voice and produce a classification of the emotional state of the speaker and if they are telling the truth. Many well-known AI applications are built using traditional AI, such as spam filters in email, AI-generated recommendations on e-commerce platforms, self-driving cars, data mining, original virtual assistants like Siri, and some automated games. These architectures and applications existed years before the current arrival of so-called "generative" AI applications.

13. By contrast, generative AI strives to generate content by combining existing materials together using patterns it derives from training materials. It utilizes complex machine

learning architectures to extract correlation patterns between a textual prompt or request, and specific content samples (which can be text, images, or audio) that match those prompts. These models extract from training data the likelihood of a particular sequence of tokens (response) that follow other sequences of tokens (user prompt), captured and represented in layers and layers of conditional probabilities. Sufficient training on large datasets (including much of all human writing that has ever been written), produces the generative model, the encapsulation of a massive compilation of correlations between text prompts and content, stored as layers and layers of conditional probabilities. These probabilities can then guide the reproduction of content that have the highest chances of matching any textual content, as defined by the text prompt given by the user of the generative model.

1. LLM Training

14. All AI models are based on a particular architecture (the design and structure of the model). LLMs typically utilize a "transformer" architecture, a type of machine learning algorithm.[1] Among other features, a transformer is comprised of an encoder, which reads and processes the input data, and a decoder, which generates predictions. Transformers work on predicting high probability text "tokens" given some contextual information, where the context is often measured in number of tokens of memory. The more powerful and complex the model, the more tokens it can capture in its contextual memory to guide its prediction of meaningful responses to incoming user questions and requests.

15. The creation of an LLM is a multi-stage project.[2] Preliminary steps include defining the object of the model (whether it will be a chatbot, a translation tool, etc.), designing the

[1] *See, e.g.*, Ashish Vaswani et al., *Attention Is All You Need*, arXiv:1706.03762v7 (Aug. 2, 2023), https://arxiv.org/pdf/1706.03762.pdf.
[2] *See generally* U.S. COPYRIGHT OFFICE, COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 3: GENERATIVE AI TRAINING (pre-publication version) 4–21 (2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

architecture, and related tasks. This initial work is followed by data-intensive steps of training the defined model.

16. In preparation for training, the creator of an LLM gathers text material and other content, typically in massive quantity and from a variety of sources, and copies it to a central repository. Accordingly, like other LLM developers, Anthropic must as an initial step in this development process select the data on which it will train its Claude AI models.[3] 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.

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

18. After the data is cleaned, the text is converted into "tokens."[4] 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.

19. After the dataset is cleaned and "tokenized," the LLM is trained. The first step is called pre-training, during which the model is trained to predict the next token/word in a string. For example, it will be given the sentence, "The dog wagged its ___," and will learn to predict the next word as "tail." Or "I like ice ___," where the answer could be "cream," "cubes," or various other choices. Anthropic's Claude models are trained in this way.[5]

---

[3] *See, e.g.*, J. Kaplan Dep. Tr. 135:17–136:14.
[4] *See, e.g.*, *Glossary,* ANTHROPIC, https://docs.anthropic.com/claude/docs/glossary (last visited Dec. 18, 2025).
[5] *See, e.g.*, J. Kaplan Dep. Tr. 13:20–25.

20. Through pre-training, the model will learn which of these is the best choice from context. Pre-training also involves "weighting" the dataset. In this process, the trainer directs the model to use some parts of the dataset more often than others due to differences in perceived quality of the material, with portions engaged multiple times and others not at all. Ultimately, the model will learn the most common patterns in the grammar, syntax, and semantics of a given language. This process requires no supervision (*i.e.*, labeling or metadata describing the contents of the text). Rather, the model recognizes patterns in the training data to infer these rules automatically.

21. 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.[6] This process typically utilizes "reinforcement learning," where computers and humans rate the output supplied by the model in response to a given prompt, adjusting the values of the model's parameters to steer it to prefer certain sorts of responses and avoid others. This enables the model to produce more accurate and tailored responses, consistent with those desired by the LLM developer. The most popular version of this training is commonly referred to as RLHF, or "reinforcement learning with human feedback." Anthropic also finetunes its Claude models in this way.[7]

22. 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.[8] Pre-training also "often requires orders of magnitude more data and computing power than other training."[9]

23. Today's leading generative AI systems can be used to generate text, images, audio visual works, music, and more. Regardless of their specific task or architecture, generative AI systems are trained on very large datasets of content according to the form of content they are

[6] *See, e.g.*, *Glossary,* ANTHROPIC, https://docs.anthropic.com/claude/docs/glossary.
[7] *See, e.g.*, J. Kaplan Dep. Tr. 16:1–20.
[8] *See* U.S. COPYRIGHT OFFICE, *supra* n.2, at 17–18.
[9] *Id.*

intended to generate and input they are designed to accept. The larger the corpus on which a model is trained, the more effective and reliable its results will be as a general matter.[10]

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

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

2. LLM Memorization

26. In general, the goal of an LLM is to accurately reproduce 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 LLMs is the problem of memorization: that is, whether the model is truly capable of generating new content or merely regurgitating portions of its training data. Because the goal of the LLM training process is to capture the most accurate probability or likelihoods of sequences of tokens that follow a particular prompt or context, it means that for any particular piece of writing in its training data, training the LLM means teaching it that specific tokens should occur in a particular sequence as it does in that piece of writing. In other words, training an LLM is, by design,

[10] Anthropic employees have also recognized this principle in published research. *See, e.g.,* Jared Kaplan et al., *Scaling Laws for Neural Language Models*, arXiv:2001.08361 (Jan. 23, 2010), https://arxiv.org/pdf/2001.08361; Yasaman Bahri et al., *Explaining Neural Scaling Laws*, 121 PROC. NAT'L ACAD. SCI. U.S.A. 27 (2024), https://www.pnas.org/doi/10.1073/pnas.2311878121.

teaching it to "overfit" to each piece of data in the training set. The nature of conditional probability training means that an LLM will memorize unique phrases or sequences of phrases that do not commonly occur across numerous training samples, *e.g.*, specific turns of phrasing such as song lyrics or memorable chapters or speeches. In the words of the U.S. Copyright Office, "[w]here the learned pattern is highly specific, 'the pattern is the memorized training data.'"[11]

27. In LLM training, conditional probabilities between memorable or often-seen sequences become very strong, meaning that a specific sequence is highly likely to be generated in that same precise order matching the training data (regurgitation). For example, an LLM that is trained on the Common Crawl snapshot of the internet is likely to have seen a number of copies of lyrics to Don McLean's song "American Pie." After training, an LLM will capture a high conditional probability, that the sequence of tokens "bye-bye, Miss American Pie" should be followed by the sequence "Drove my Chevy to the levee, but the levee was dry." This occurs because it is highly unlikely to have seen "bye-bye, Miss American Pie" followed by any other token sequence.

28. This storage of "large" or "highly likely" conditional probabilities between trained sequences is fundamental to the transformer architecture, and is the key factor to LLMs memorizing and regugitating content. Simple intuition follows that the conditional probability connecting these lines of lyrics is strongly embedded inside the LLM, and cannot be easily removed, even if the lyrics were decoupled from the song title. Just prompting the LLM with one well trained line of lyric is highly likely to trigger the next line in the song lyric.

29. This is also the reason why input filters are ineffective in trying to block the generation of copyrighted content. A well-trained LLM does not necessarily need the full context (author or name of song) to trigger regurgitation of its lyrics. All it needs is an initial token sequence that triggers the conditional probability to generate a few words from the lyric, at which

[11] U.S. COPYRIGHT OFFICE, *supra* n.2, at 20 (quoting A. Feder Cooper & James Grimmelmann, *The Files are in the Computer: Copyright, Memorization and Generative AI* at 23–24, arXiv:2404.12590 (forthcoming 2025) ("The Files are in the Computer: Copyright, Memorization and Generative AI"), https://arxiv.org/abs/2404.12590.

time the rest of the conditional probabilities will enforce that each following word appears in sequence that matches how they appear in the training data, leading to a regurgitation of the remainder of the lyrics. This is why lyrics to Will Smith's "Fresh Prince of Bel-Air" can be fully regurgitated if an LLM is prompted for "a poem about a young black man moving from Philadelphia to Los Angeles," because the rephrased prompt is close enough to trigger the contextual probability leading to the first few words of the song. Once those first few words are generated as tokens, conditional probability takes over and completes the remainder of the song lyrics.

30. For an input filter to successfully prevent a particular song lyric from being regenerated, it must block *all possible* ways of triggering the initial conditional probability of the first few words of the song, including but not limited to all possible ways to describe the content of the song. Even then, a user prompt that includes words from the beginning of a popular song are highly likely to trigger conditional probabilities stored in the model to start reproducing the sequence of tokens representing words that follow in the lyric, which then leads to the regurgitation of the full song.

31. While these mathematical representations of probability values between words or tokens are seemingly complex and difficult to understand, there are actually existing algorithms in computer science that work very similarly. These algorithms can be used as an accurate analogy to explain the large scale extraction of conditional probabilities from a set of training data. The algorithms I refer to are *compression* algorithms, tools that have been well understood and taught in computer science courses for decades.

32. Next, I will explain the intuition behind compression algorithms, and why LLM training is analogous to computing a large lossy compression representation of the training data given a fixed sized output (the actual trained model itself).

33. Compression algorithms are deterministic algorithms that take an input and replace the original content with shorter, smaller, and more efficient bit representations of that content. *Lossless* compression algorithms ideally perform this process without the loss of any data, so that

the process can be deterministically reversed, *i.e.*, uncompressed or decompressed, to recover the original content. For example, the Lempel–Ziv–Welch ("LZW") family of compression algorithms works by iteratively building a table of sequences of characters and replacing them with short identifiers. When content is run through these algorithms, sequences in the content are replaced with short identifiers as specified in the tables. This compression is deterministic (each run of the algorithm produces an identical result) and lossless, in that the content can always be uncompressed to recover the exact original content.

34. Compression algorithms generally have the property that the more times some piece of data appears in the input, the easier those pieces of data can be compressed. The more unique a piece of data is, *i.e.*, the higher its "entropy," the harder that data will be to compress losslessly. This is in part why AI developers generally continue to push for bigger and bigger models and correspondingly larger datasets. Models with bigger capacities are able to encode layers of conditional probabilities that effectively represent a compressed version of the training data.

35. In a *lossy* compression algorithm, by contrast, the compression algorithm sacrifices some loss of information in order to compress the data further than a lossless algorithm could achieve. Unlike lossless compression algorithms, compressions produced by lossy compression algorithms may not be fully reversed, *i.e.*, used to fully recover the entire and exact original input before it was compressed. However, it is generally possible to recover a close approximation of the original, simply by replacing the lost components with random or deterministically generated data.

36. As an example of a lossy compression algorithm, consider a simple function that takes a real number and rounds it off to produce an integer. This trivial algorithm would take the number "13.6" as input and produce "13" as output. "13" will take fewer bits to store and thus compression has occurred. But decompressing this lossy stored value presents a problem of recovering the digit lost during compression. One solution would be to simply take "13" and append a random decimal component. If we know that the inputs have at most 1 decimal value,

then the recovery algorithm has roughly a 1 in 10 chance of fully recovering the original input ("13.6") simply by guessing that final digit.

37. As a general matter during the training process, a portion of training data is sent as input into the network and the network—following the strength of its weights—produces a predicted output. That output is compared to the "correct" output, *i.e.*, the value contained in training data. For example, the input might be "the quick brown fox jumps over the lazy," the "correct" output distribution (per the training data) might be "dog" (100%), but the model's predicted possible outputs distribution might be "boy" (50%), "dog" (12.5%), and a long tail of less-likely potential answers taking up the remaining 12.5%. The meat of the training process—and the equivalence of the LLM training process and compression—comes from how the algorithms measure and correct the neural network in response to this incorrect guess.

38. The central feature of this process is the "loss function." In LLM training, the loss function measures how much error there is between the predicted output made by current model parameters and the true output in the training data. The loss function calculates a "loss" which represents how far the current model's output is to the ideal, and is then used to compute updates to the model parameters. Loss functions simply represent the error in how well the current model parameters match a compressed version of the training data. This is because any given loss function is generally calculated by testing the model's ability to complete text based on partial inputs. That is equivalent to testing how completely a compressed file corresponds to the original data by decompressing it and measuring the difference between the decompressed version and the original data. As such, the lower the measured loss, the closer the compressed version stored in the model is to an ideal compression of the training data.

39. An example of a loss function is cross-entropy loss. As with other loss functions, cross-entropy loss measures the difference between the ideal or true output and the model's predicted output.[1] A lower cross-entropy loss indicates that the predicted probabilities are closer to the true or ideal output. An example of a cross-entropy loss calculation looks like:

$$L = -\sum_{t=1}^{T} \log P_\theta(y_t \mid y_{<t})$$

40. In this formula the loss equals the negative sum of the log probabilities of the model's prediction given the previous sequence of tokens. Note that in cross-entropy loss generally the calculation is performed relative to "distributions," but in LLM training, where output comes discretely in token-form, a comparison between distributions can effectively operate as a comparisons of outputs. It is my understanding that LLMs generally use this simple type of cross-entropy loss computed by comparing a predicted output to the true output present in the training data.

41. During the LLM training process, the model's parameters are adjusted continuously by measuring how closely the parameters reproduce the training data as shown by loss functions and optimizing the model to reduce the calculated loss. That process of compression, decompression, comparison, and adjustment is repeated over and over until the model approximates a compression of the training data as measured by the loss function. This training process is thus functionally equivalent to a lossy compression of the training data.

42. To be sure, LLMs have a fixed size or capacity, which is defined by the architecture and the model parameters. That size will not change regardless of the training data used, whether it is 5 sentences or 5 million files downloaded from the web. A large LLM that is trained on a small dataset may have a large amount of redundancy in its model weights, and more strongly encode training data more precisely with little loss. That same LLM trained on much more training data will still store large portions of the training dataset, but may discard the rest. The discarded portions amount to the lost data in the lossy compression process. In other words, training an LLM which has a predefined capacity based on its architecture, is functionally equivalent to computing a lossy compression of the training data where the size of the lossy output is fixed. Further, the more times a piece of data occurs in training data, the higher it will be prioritized in the lossy compression, and the more likely it will be memorized in the compressed representation of the training data that is the model.

43. In short, the goal of the training process is to compress the data sufficiently to give the model as much information as possible and enable the model to reproduce that data accurately when requested. Accordingly, there is no effective way to avoid the problem of memorization since the model is trained to do precisely that.[12]

44. The LLM training process is not analogous to the process of human learning. Humans learn from identifying, understanding, and applying top-down rules through generalizable examples. LLMs, in contrast, extract answers for a particular question only by looking at millions of examples and tabulating probabilities. For example, children learn as a rule that they should be careful and look both ways when crossing a street. They do not need to be shown millions of images showing cars either hitting or missing pedestrians, on every possible kind of street or road. A child can learn that any area with moving vehicles can become a road, but an LLM would only recognize a particular type of road once it has seen enough copies of images labeled as a road. In other words, a child learns the underlying "rules" that produce outcomes in the real world, whereas LLMs only estimate the probability of a particular outcome based on what it has seen and nothing else.

45. In another example, in the now-classic example of an LLM regurgitating nonsense, an LLM was shown to answer the question "how do I stop pizza toppings from sliding off a pizza" with "add 1/8 ounce of non-toxic glue to the toppings."[13] It returned this answer because it has a single Reddit post in its training data that suggested (jokingly) that glue should be used to fix toppings to slices of pizza. While any child knows the rule that "glue" is not edible and therefore should not be placed on food, an LLM would only be able to learn such a rule if presented with

[12] *See* Ofir David, et al., *On statistical learning via the lens of compression*, PROC. OF THE 30TH CONF. ON NEURAL INFO. PROC. SYS. (NIPS 2016) at 2, https://dl.acm.org/doi/pdf/10.5555/3157382.3157410; G. Deletang et al., *Language Modeling is Compression*, 12TH INT'L. CONF. ON LEARNING REP. (2024), https://openreview.net/pdf?id=jznbgiynus.

[13] Katie Notopoulos, *Google AI Said to Put Glue in Pizza—So I Made a Pizza With Glue and Ate It*, BUSINESS INSIDER (May 24, 2024), https://www.businessinsider.com/google-ai-glue-pizza-i-tried-it-2024-5.

sufficient copies of training data that explicitly states the dangers of putting glue of any kind on food.

46. This exposes the critical problem with learning by examples: you cannot generate all possible positive *and negative* examples in training data. Therefore, an LLM will not have any inclination, positive or negative, on a topic that is completely uncovered in its training data (*e.g.*, one should not snort gasoline into one's nostril). If an LLM came across a piece of training data suggesting such an action, *e.g.*, a Reddit post suggesting it clears your sinuses, those conditional probabilities trained from the Reddit post would likely lead to an LLM making that suggestion when asked about how to clear clogged sinuses. In contrast, a child would have learned from a parent that gasoline is toxic to the human body and should not be ingested in any form. This fundamental training sparsity problem and a failure to generalize is what fundamentally limits the ability of LLMs to answer questions it has not seen in its training data. This generalization failure is well known in the research community,[14] and is the main reason why LLMs often struggle with complex mathematics, because it is difficult to generate or gather training samples that capture every possible mathematical operation involving every number.[15]

47. This failure to generalize is also not a surprise, because it is fundamental to machine learning in general, and is well known in the previous generation of machine-learning models (neural networks) as out of distribution (OOD) data. It is widely accepted that neural networks fail when they are trained on one type of data distributions, and then given queries on data from a different distribution, *e.g,.* a DNN trained to recognize a wolf using images of wolves in the snow

---

[14] *See* Parshin Shojaee et al., *The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity*, arXiv:2506.06941 (Nov. 20, 2025), https://arxiv.org/pdf/2506.06941; Lukas Berglund et al., *The Reversal Curse: LLMs Trained on "A is B" Fail to Learn "B is A,"* arXiv:2309.12288 (May 26, 2024), https://arxiv.org/pdf/2309.12288; Tomer Ullman, *Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks*, arXiv:2302.08399 (Mar. 14, 2023), https://arxiv.org/pdf/2302.08399.

[15] Erxin Yu et al., *Self-Error-Instruct: Generalizing from Errors for LLMs Mathematical Reasoning,* PROC. OF THE 63RD ANN. MTG. OF THE ASS'N FOR COMPUT. LINGUISTICS (VOL. 1: LONG PAPERS), 8504–19 (2025), https://aclanthology.org/2025.acl-long.417.pdf.

fails to recognize a wolf given images of wolves in a person's backyard. A human or other biological system learns both through internal computation and physical interaction with the external world—AI models cannot interact with and learn through the physical world.[16] Thus, AI model training lacks a significant portion of what we consider human "learning." Further, AI researchers still don't fully understand how LLMs work, making a comparison to human learning processes very difficult.[17]

B. Anthropic's Claude AI Models

1. Overview

48. Since its inception, Anthropic has branded itself as an AI company that claims to prioritize AI safety and minimize harm. Anthropic claims it conducts research to "investigate the safety, inner workings, and societal impacts of AI models" so that "artificial intelligence has a positive impact as it becomes increasingly capable."[18] Its approach to AI "safety" includes a number of techniques to inculcate its models with desired traits and values, including "Constitutional AI" and "character training."[19] These are discussed in further detail below.

49. Anthropic offers its main product, "Claude," a series of commercial LLMs, with an interactive interface ("chatbot") available through its claude.ai website, as a mobile application for iPhone and Android users, as a desktop application, and as an application programming interface ("API") for integration by businesses and other customers into their products and applications.

50. Among other things, Anthropic offers a "Claude for Enterprise" plan for business customers to collaborate with a 500K context window, increased usage capacity, and additional security protocols like single-sign on and role-based permissions. The Enterprise plan permits organizations to share and reuse information shared with Claude by employees.[20]

---

[16] *See, e.g.*, D. Ganguli 30(b)(1) Dep. Tr. 27:2–7.
[17] *See id.* at 23:18–21, 102:7–14.
[18] *See Research*, ANTHROPIC, https://www.anthropic.com/research (last visited Dec. 17, 2025).
[19] *See* B. Mann 30(b)(6) Dep. Tr. 148:3–16.
[20] *See* Claude for Enterprise, ANTHROPIC (Sept. 4, 2024), https://www.anthropic.com/news/claude-for-enterprise.

51. Anthropic also sells or licenses first-party API-based access to Claude models to commercial customers, incorporating Claude into these customers' software, products, and systems. Anthropic partners with commercial customers like Quora and DuckDuckGo, to allow users to interact with Claude models through those customers' websites or services.[21] Anthropic also offers API access to its Claude models through third party AI platforms including Amazon Bedrock and Google Vertex.

52. Anthropic first released Claude to certain business customers in late 2022, and it first launched Claude to the general public in March 2023. It has subsequently released a series of newer and more-advanced versions of Claude, both prior to this lawsuit and during the pendency of this litigation.[22] The currently available versions of Claude include Claude Opus 4.5 (released November 24, 2025),[23] Claude Haiku 4.5 (released October 15, 2025),[24] Claude Sonnet 4.5 (released September 29, 2025).[25]

53. The Claude.ai web version is available through multiple tiers of paid subscriptions with different levels of service and a more limited free version.[26] The free version of Claude can be accessed by anyone who creates an account verified with a text message sent to a mobile phone number.

54. The current and earlier versions of Claude are also available to commercial customers to be incorporated into their own software, products, and systems through an API. Claude models are currently already integrated into a variety of products, including group

---

[21] *Duck.ai*, DUCKDUCKGO, https://duckduckgo.com/duckduckgo-help-pages/duckai; *How does Duck.ai protect my privacy?*, DUCKDUCKGO, https://duckduckgo.com/duckduckgo-help-pages/duckai/ai-chat-privacy.
[22] Anthropic's Supplemental Response to Plaintiffs' Interrogatory No. 14, dated Oct. 9, 2025, at 4.
[23] *Introducing Claude Opus 4.5*, ANTHROPIC (Nov. 24, 2025), https://www.anthropic.com/news/claude-opus-4-5.
[24] *Introducing Claude Haiku 4.5*, ANTHROPIC (Oct. 15, 2025), https://www.anthropic.com/news/claude-haiku-4-5.
[25] *Introducing Claude Sonnet 4.5*, ANTHROPIC (Sept. 29, 2025), https://www.anthropic.com/news/claude-sonnet-4-5.
[26] *Individual Pricing*, Anthropic, https://claude.com/pricing (last visited Dec. 14, 2025).

messaging applications (like Slack), Internet search engines (such as DuckDuckGo), note-taking applications (like Notion), and AI-aggregator sites (such as Poe). To interact with the Claude chatbot in these various platforms, users typically enter a prompt into the customer software. Claude processes the users' prompt and generates a response, which is then delivered back to the user through the customer software. This process is invisible to the ultimate end user. I understand Anthropic currently serves about 300,000 businesses through its Claude products.[27]

2. Anthropic's Training of Claude

55. Claude's training appears to be consistent with the general description above.[28] Anthropic first selects the specific content, or training data, it will use to train its Claude AI models. In general, Anthropic prioritizes large and diverse datasets of preexisting human-written content when selecting training data.[29] 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.[30]

56. Notably, however, Anthropic does not publicly disclose the specific training datasets it used to train its Claude AI models. Anthropic has also expressed the view that its selection and identification of training data can afford it a [redacted]" over other AI companies,[31] which is consistent with my understanding and which is likely to influence Anthropic's decisionmaking with respect to its selection of training data.

57. One of the methods by which Anthropic obtains much of its training data is by scraping content from websites on the internet using bots, a practice it began in 2023.[32] Anthropic retains copies of this scraped content in cloud blob storage.[33]

---

[27] Dan. Amodei Dep. Tr. 68:8–9.
[28] *Supra*, Sections IV. A. 2–3.
[29] *See, e.g.*, Anthropic_0000218479; Kaplan Decl. (Jan. 16, 2024), ECF No. 67-1 ¶ 19; Kaplan Decl. (Nov. 22, 2023), ECF No. 55-1 ¶ 15.
[30] Anthropic_0000218479. *See also* Anthropic_0000014110 ("we should prioritize books 'a little' more highly).
[31] Anthropic_0000218479.
[32] B. Mann 30(b)(6) Dep. Tr. 49:1–4.
[33] *Id.* at 55:8–13.

58. "Scraping" is a process by which a bot "crawls" webpages (*i.e.*, visiting a webpage) and creates a copy of the webpage's contents.[34] A crawler bot can review a webpage's robots.txt files before conducting a scrape.[35] If the owner of a webpage does not want a crawler to be able to scrape the webpage's content, the owner can insert "do-not-crawl" instructions in the webpage's robots.txt files.[36] Some crawlers will respect those instructions and accordingly refrain from scraping that webpage's content.[37]

59. 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 and scraping data from their websites in violation of their policies and robots.txt instructions.[38] One Anthropic co-founder also admitted that Anthropic has not always honored robots.txt instructions.[39] Additionally, during a period of time in [redacted], at least one of Anthropic's scraping tools did not even have the [redacted].[40]

60. It is worth noting that AI crawlers written by Anthropic and others now account for the overwhelming majority of network traffic on the Internet, dominating traffic that is actually used by human users. Numerous reports demonstrate that the relentless and aggressive nature of AI crawlers are placing extremely high loads on websites of all types, in many cases overloading

---

[34] *See generally* Columbia University Mailman School of Public Health, *Web Scraping*, https://www.publichealth.columbia.edu/research/population-health-methods/web-scraping. *See also* B. Mann 30(b)(6) Dep. Tr. 47:7–10.
[35] *See* Columbia University Mailman School of Public Health, *Web Scraping*, https://www.publichealth.columbia.edu/research/population-health-methods/web-scraping.
[36] *See id.*
[37] *See id.*
[38] *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-scraping rules 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.*
[39] B. Mann 30(b)(6) Dep. Tr. 56:18–57:1.
[40] P. Sangani Dep. Ex. 128 (Anthropic_0000346083).

them, causing failures, and in other cases dramatically increasing network costs for larger and smaller web services alike.[41] In specific instances, web operators have equated Anthropic's AI crawler as performing "denial of service" attacks on their site, using large numbers of cloud servers to scrape data so aggressively as to overload and bring down servers.[42] Given the fact that online content is now becoming increasingly restricted by publishers and websites that have taken steps to prevent bots from crawling their websites,[43] it is my view that LLM developers like Anthropic are correspondingly incentivized to ignore robots.txt instructions to harvest the data needed to train bigger and newer models.

61. Anthropic also utilized extractor tools, such as the "Newspaper" extractor tool, during the scraping process to extract text from websites without what it referred to as "useless junk" contained in webpage headers and footers.[44] I understand footers and footnotes on many webpages typically contain copyright information, including copyright owner names and © notices. Based on my review of internal Anthropic discussions, it is my view that Anthropic utilized such extractor tools because they removed "undesirable" information and repetitive information—such as licenses and other so-called "garbage" [45] —from scraped content, including copyright management information.[46] Practices such as this are common among LLM developers, since the

[41] *See* Michael Kan, *Wikipedia Faces Flood of AI Bots That Are Eating Bandwidth, Raising Costs*, PC MAG (Apr. 2, 2025), https://www.pcmag.com/news/wikipedia-faces-flood-of-ai-bots-that-are-eating-bandwidth-raising-costs; Chris Stokel-Walker, *AI Crawlers Are Hammering Sites and Nearly Taking Them Offline*, FAST COMPANY (Sep. 26, 2024), https://www.fastcompany.com/91197457/ai-crawlers-are-hammering-sites-and-nearly-taking-them-offline.
[42] davepattern, *DDoS from Anthropic AI*, AKAMI CLOUD (2024), https://www.linode.com/community/questions/24842/ddos-from-anthropic-ai.
[43] 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. Some researchers have also suggested that LLM training will deplete the stock of publicly available human text data between 2026 and 2032. Pablo Villabos et al., *Will we run out of data? Limits of LLM scaling based on human-generated data* at 1, arXiv:2211.04325 (June 4, 2024), https://arxiv.org/pdf/2211.04325.
[44] *See* J. Kaplan Dep. Ex. 234 (Anthropic_0000014096–106).
[45] *See* Anthropic_0000500641; Anthropic_0000216882; D. Drain Tr. 118:13–121:5.
[46] *See id. See also* First Amended Complaint ("FAC") ¶¶ 70–74, ECF No. 337.

repetitive inclusion of certain such data in training corpora during the pre-training process may reinforce regurgitation of texts that developers such as Anthropic do not desire.[47]

62. Anthropic has also disclosed that it relies heavily on datasets other third parties have scraped from the internet (among other methods of collection), including the Common Crawl, The Pile, and OpenWebText2 datasets that include large amounts of copyrighted and unlicensed material, including Publishers' lyrics.[48] In particular, the Common Crawl includes unlicensed copies of Publishers' lyrics scraped from licensed lyrics websites like LyricFind.com and Musixmatch.com.[49] Meanwhile, The Pile is known to include the "YouTube Subtitles" sub-dataset, which contains various unlicensed copies of Publishers' lyrics compiled from the subtitles of licensed YouTube videos for those works.[50] Anthropic has also trained on data specifically taken from Reddit,[51] among various other datasets.

63. After obtaining its datasets, Anthropic trains Claude to its specifications, selecting the level of compression and data loss, the weights it will give different forms of data, and other criteria. The resulting "pre-trained" 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."[52] This pre-trained model may be suited to various tasks but memorizes large amounts of its datasets; Anthropic's

---

[47] Katherine Lee et al., *Deduplicating Training Data Makes Language Models Better*, arXiv:2107.06499 (Mar. 24, 2022), https://arxiv.org/pdf/2107.06499.
[48] *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; Anthropic's Amended and Supplemental Response to Plaintiffs' Interrogatory No. 6, dated Oct. 9, 2025, at 3.
[49] *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/.
[50] *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/.
[51] Anthropic's Amended and Supplemental Response to Plaintiffs' Interrogatory No. 6, dated Oct. 9, 2025, at 3.
[52] *Id.* at 93:24–94:7.

internal documents concede that "Large LMs memorize a LOT."[53] Indeed, LLMs are probability models and are, fundamentally, memorizing how often certain words or phrases appear next to each other (encoded as conditional probabilities in the model). This leads to literal memorization and regurgitation of texts, as I have explained above.[54] In 2023, one Anthropic employee discovered that Anthropic's models were "memorizing in this harmful, superficial way . . . large amounts of text" and that "[t]he effect was getting worse at scale. It seemed like the bigger models were memorizing more[.]"[55] This occurs when a model is "trained on a piece of text many times."[56] Note that this is entirely consistent with the concept of LLM training as lossy compression, as I have also explained above.[57] A larger capacity model is more capable of compressing training data in its entirety, allowing more accurate memorization. Similarly, training on the same text multiple times teaches the lossy compression process to prioritize this text above others, increasing the likelihood of it being memorized by the model. It also helps the training process to find better compressed representations of the training data, which manifests itself in training as lower loss values in the optimization process. This is a feature and not a bug of LLMs and their training process, and something that developers might choose to mitigate due to degradation of LLM performance but cannot eliminate.[58] As a result, even if Anthropic were to apply aggressive data cleaning or input filtering, Claude would nonetheless be prone to memorization and regurgitation of texts that appear frequently in its training corpus.

---

[53] *See* D. Drain Dep. Ex. 140 (Anthropic_0000507796); *see also, e.g.*, Anthropic_0000458427 (July 2021 memo discussing whether "we can reduce data recitation / solve the copyright problem internal to the model via RL").
[54] *Supra,* Section IV. A. 2.
[55] D. Drain Dep. Tr. 90:2–24. *See also* Anthropic_0000500599-500647.
[56] D. Drain Dep. Tr. 71:4-7.
[57] *Supra*, Section IV.A.3.
[58] *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.

64. But LLMs also memorize in the sense that that they are trained to generate specific texts. As the same Anthropic employee 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[.]"[59] The employee further testified that it is "reasonable for the model to memorize" certain content,[60] 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[.]"[61] Such texts include texts that are "very important to our culture" such as the U.S. Constitution, the Gettysburg address, or Martin Luther King's "I Have a Dream" speech.[62] 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. Unfortunately, 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 indiscriminately scraping the Internet for training data produces massive quantities of data, which is then used for model training wholesale without distinguishing between culturally important 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.[63]

65. Anthropic's training data includes copies of Publishers' lyrics, including in particular the 499 Works in Suit.[64] Anthropic does not appear to dispute that the training data for its Claude models, including Claude 1, 2, 3, and 4, all the way through current releases of Claude

[59] D. Drain Dep. Tr. 68:8–19.
[60] D. Drain Dep. Tr. 76:3–15.
[61] D. Drain Dep. Tr. 182:6–12.
[62] *Id.* 75:3–76:15.
[63] *See, e.g.*, Anthropic_0000434712.
[64] FAC Ex. B, ECF No. 337-2.

4.5, include copies of Publishers' lyrics.[65] I also understand that Publishers' expert Dr. Shan confirmed the presence of copies of Publishers' lyrics, on a large scale, in Anthropic's training data during his inspection of that data.[66] Dr. Shan's finding that Publishers' lyrics are included in Anthropic's training data is consistent with my expectations, given the fact that Anthropic employs massive amounts of training data scraped from the internet, as well as in light of the output Anthropic's AI models have generated containing Publishers' lyrics, as discussed in more detail below, among other reasons.

66. Since Publishers filed their Complaint in this case on October 18, 2023, Anthropic has steadily released new versions of Claude. In particular, it released Claude 2.1 in November 2023, the Claude 3 model family (Haiku, Sonnet, and Opus) in 2024, and the Claude 4 family in 2025. I understand the most recent version of Claude for which training data was made available to Publishers for inspection in this case was Claude 3.7 Sonnet. Each one of these models for which training data was made available for inspection was trained on a dataset containing Publishers' works.[67]

67. Anthropic has also acknowledged that the training data for its Claude models may contain multiple copies of the same lyrics.[68] Indeed, I understand that Dr. Shan confirmed during his inspection of Anthropic's training data that each of the 499 works in suit were copied multiple times, as many as [redacted] times. The most common websites from which these lyrics were copied include: www.lyricsfreak.com, www.lyricsmode.com, www.lyrics.com, www.stlyrics.com, lyricstranslate.com, www.elyrics.net, lyrics-video-music.blogspot.com, www.sing365.com, www.metrolyrics.com, and www.lyricsdepot.com, among other websites.[69]

68. The fact that Claude's training data includes large numbers of copies of specific lyrics increases the likelihood that Claude memorizes and regurgitates specific lyrics. As I

[65] B. Mann 30(b)(6) Dep. Tr. 96:10–15.
[66] *See generally* Expert Report of Dr. Shawn Shan (Dec. 19, 2025) ("Shawn Expert Report").
[67] *Id.* at Section VII.B.
[68] Decl. of Jared Kaplan, ECF No. 67-1 ¶ 26.
[69] Shan Expert Report Section VII.

described earlier, training on multiple copies of a piece of data allows the training process to prioritize that piece of data for memorization in the model.

69. Apart from Dr. Shan's inspection, it is separately evident from the Claude prompts and output records produced by Anthropic (as well as the evidence of Claude prompts and outputs Publishers collected and produced in this case)[70] that Anthropic has trained Claude on Publishers' lyrics. Claude would obviously not generate exact or near-exact copies of song lyrics in response to a query unless the system had copied and stored those lyrics. The probability of 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.

70. In addition, while LLMs such as Claude are designed to mimic fluent speech and text composition, they cannot naturally perfectly or near-perfectly generate song lyrics for specific songs without having been trained on those lyrics. Song lyrics often use non-standard words or verbiage that are found nowhere else, making them extremely unlikely to be predicted by an LLM that has not been exposed to them. For example, Will Smith's song "Fresh Prince of Bel-Air" includes not only proper nouns like Bel-Air, but also non-standard contractions like "maxin'" and "relaxin'," words that are not part of standard English syntax. The presence of these unique words in an LLM's output, when combined with verbatim or near-verbatim phrases that match the entire song, leave little doubt that the model is reproducing the lyrics after being trained on them. In other words, the reproduction would be extremely unlikely if the model had not been exposed to copies of the lyrics during training.[71]

71. Publishers' copyrighted lyrics have evident value to Anthropic. First, as discussed below, once an LLM is trained on a given piece of content, it can then use that content for any purpose.[72] Anthropic's inclusion of Publishers' lyrics in its training data of course enables Claude to deliver full or partial copies of those lyrics in output. But Anthropic's inclusion of Publishers'

[70] *See, e.g.,* Decl. of Dan Seymour, Exs. B, C, D, ECF Nos. 49-2, 49-3, 49-4, 49-5, 49-6, 49-7.
[71] *See* U.S. COPYRIGHT OFFICE, *supra* n.2, at 12-13.
[72] *Infra,* ¶ 9191.

lyrics in its training data also enables Claude to perform various other functions and use cases, including those relating to songs and lyrics specifically, such as identifying songs based on lyrics; translating or modifying the lyrics to existing songs; generating "new" or "original" song lyrics in the style of a specific songwriter, artist, or song, on a given topic, or otherwise in response to a user prompt; or generating other output that includes parts of the original work blended with different text. Such output might comprise distinct musical compositions, poems, short stories, marketing pitches, or any other form of written work—but particularly those requiring the unique structure of song lyrics.

72. 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. Song lyrics employ language in unique and unusual ways, and contribute to the overall diversity of training data, among other important contributions. In the words of Anthropic employees, song lyrics have value to the Claude system even if they are not being delivered as verbatim copies of the original work, including in the following ways:

- "[S]ong lyrics [are] useful to include in [Claude's] training corpus because they add to the diversity of the training data;"[73]
- "[S]ong lyrics deploy human language differently than do news articles, which deploy human language differently than novels, which deploy human language differently than speeches, and so on;"[74]
- "[S]ong lyrics often incorporate use of figurative language, such as metaphor, simile, or personification, that a model needs to understand in order to interpret the full range of human input;"[75]

[73] Kaplan Decl. (Aug. 22, 2024), ECF No. 209 ¶ 35;
[74] *Id.*
[75] *Id.*

- "In the finetuning process to make Claude very good at writing code, presumably song lyrics are not important for that part. However, you benefit from the pre-training model having seen a wide a variety of data."[76]

73. Anthropic has also previously claimed that it does not "store" copies of Publishers' works.[77] I interpret this to mean that Anthropic does not preserve a copy of the original training data as readable text. But at a minimum, Anthropic must effectively store Publishers' lyrics in tokenized format in the Claude model, along with conditional probabilities that instruct the model to re-integrate those tokens into full or partial copies of the lyrics on command in response to user queries. That is how Claude is able to output complete or near-complete copies of Publishers' lyrics.

74. Anthropic co-founder and Chief Science Officer Jared Kaplan previously stated in a declaration in this case that "Anthropic has no interest in extracting or retaining the actual expressive content of any particular text used to train Claude. Claude does not retain specific quotes or chunks of texts from the corpus for the purpose of storing information it can regurgitate in response to a user prompt."[78] But even if Claude does not store what we would traditionally consider to be a "copy" of a given work, Claude will still produce of a copy of the work if the model's parameters are weighted such that it produces a copy of the work in response to a prompt—Claude doesn't *need* to store a copy of the work in its memory in order to output a copy.

3. Anthropic's Finetuning and Testing of Claude

75. After a Claude model is pre-trained, it is then finetuned. As described above, the purpose of finetuning is to help the LLM give more precise and tailored responses, consistent with the developer's expected and intended use of the model. 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.[79]

---

[76] D. Ganguli 30(b)(1) Dep. Tr. 207:10–14.
[77] Anthropic's Opposition to Motion for Preliminary Injunction, at 6, 15–16, ECF No. 67.
[78] Kaplan Decl. ¶ 36, ECF No. 209.
[79] *See* Dan. Amodei Dep. Ex. 172 (Anthropic_0000479021).

76. 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."[80]

77. Finetuning is typically done by individuals who either are the intended users of the model, or are trained to develop queries in the style of the intended users.

78. In a prior declaration submitted in this litigation, Dr. Kaplan cites to Claude's Model Card,[81] attached as Exhibit A to the declaration, as an authoritative statement of Claude 2's training inputs.[82] 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.[83]

79. That Anthropic dataset contains numerous instances where Anthropic used song lyrics to "finetune" versions of Claude.[84] Some of Anthropic's finetuning prompts that demonstrate that Anthropic used lyrics-related prompts to finetune the model include the following:

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

---

[80] *See* D. Ganguli 30(b)(6) Dep. Tr. 94:9–19.
[81] A model card is a short statement describing the purpose of the model.
[82] Kaplan Decl. ¶ 14, ECF No. 209.
[83] 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* B. Mann 30(b)(6) Dep. Tr. 199:7–204:22.
[84] *Dataset Card for HH-RLHF*, HUGGING FACE, https://huggingface.co/datasets/Anthropic/hh-rlhf (last visited Dec. 18, 2025).

80. These are just a handful of examples. In response to these prompts, Claude was presented with two competing outputs and was trained to select the "better" one, as determined by Anthropic. Generally, the two choices are quite similar and differ primarily in the way the output is expressed. This is because the finetuning stage is not the point at which content is chosen for the model, but rather when the model is trained in preferred ways of presenting the content in its response for various contexts. In these examples, the context is that of song lyrics. As described in one Publishers' filings in this case,[85] among the songs Anthropic included in its prompts, reflected in the Hugging Face data, are:

- "All Along the Watchtower," written and performed by Bob Dylan, and to which Universal owns the copyright;[86]
- "Brown Sugar," written by Michael Philip Jagger/Keith Richards, performed by the Rolling Stones, and to which ABKCO owns the copyright;[87]
- "Can't Get You Out of My Head," written by Cathy Dennis and Rob Davis, performed by Kylie Minogue, and to which Universal owns the copyright;[88]
- "In the Air Tonight," written and performed by Phil Collins;
- "My Heart Will Go On," written by James Horner and Will Jennings, performed by Celine Dion;
- "November Rain," written by Axl Rose, performed by Guns N' Roses; and
- "Roar," written by Bonnie Leigh McKee, performed by Katy Perry, and to which Concord owns the copyright.[89]

81. Based on the limited data set available, it is my opinion that these finetuning prompts indicate that Anthropic deliberately designed Claude to have the functionality to provide lyrics to well-known songs and create new lyrics in response to user requests. If Anthropic did not

[85] Chung Decl., Ex. J, ECF No. 180-10.
[86] FAC, Ex. A, ECF No. 337-1.
[87] *Id.*
[88] *Id.*
[89] *Id.*

intend for Claude to be able to respond to song- or lyric-related requests, Anthropic would not have prompted Claude using such requests during the finetuning process; the purpose of finetuning is largely to train Claude to perform specific tasks or reinforce certain behaviors.

82. Anthropic's finetuning data (*e.g.*, the prompt, "Rewrite the lyrics to Paradise City in the style of a Dr. Seuss book") also shows that Anthropic anticipated that consumers would seek to use well-known lyrics to create new songs that incorporate recognizable portions of pre-existing works, and intended Claude to respond to such requests in that manner.

83. Finally, the finetuning data indicates that the types of prompts directed to Claude by Publishers and Claude users[90] are not outliers; rather, these types of prompts are exactly the sort that Anthropic expected ordinary users of Claude would construct.

84. While Anthropic claims that Claude is "not designed to output copyrighted material," that "[n]ormal people" would not seek lyrics from Claude, and that Claude's ability to output copyrighted material is a "'bug,' not a 'feature' of the product,"[91] the record of Claude's finetuning activities and internal testing, described above, contradicts these assertions. Based on my understanding of LLM development, if Anthropic did not intend for Claude to be able to respond to song- or lyric-related requests, Anthropic would not have prompted Claude using such requests during the finetuning process; the purpose of finetuning is largely to train Claude to perform specific tasks or reinforce certain behaviors. The finetuning process incurs a significant portion of the overall model training costs, and finetuning tasks are chosen intentionally. Anthropic also acknowledges that users can use Claude to summarize song lyrics, search for song lyrics, and create song lyrics, and that Claude is capable of delivering on those requests.[92]

85. Anthropic's assertion is further contradicted by the hundreds of prompts and outputs for lyrics discovered in Claude interactions supplied by Anthropic in discovery, as

[90] *See generally* FAC, Ex. B, ECF No. 337-2; Expert Report of Joshua Dennis (Dec. 19, 2025) ("Dennis Expert Report").
[91] Anthropic's Opp. to Mot. for Prelim. Inj., at 10, ECF No. 67.
[92] *See* B. Mann 30(b)(6) Dep. Tr. 23:13–26:24.

discussed in the reports of Publishers' experts Dr. Robert Leonard and Joshua Dennis, and as described further below.

86. 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 2022 and 2023 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."[93] 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,[94] 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.[95] 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.[96] Anthropic separately undertook efforts to "benchmark" Claude's ability to generate various output—including in particular output containing lyrics—against competitor AI models.[97] Like the finetuning prompts above, these Anthropic records further support my opinion that Anthropic intentionally designed Claude to have the functionality to provide lyrics to existing songs and to respond to other song- and lyric-related requests.

4. Claude Capabilities and Usage

[93] Anthropic_0000478271 at -302.
[94] *See, e.g.*, Anthropic_0000525947 at -948–949; Anthropic_0000459162 at -168; Anthropic_0000554366 at -384–85, -387–89; Anthropic_0000554319 at -322; Anthropic_0000554980 at -986–7; Anthropic_0000527903 at -915.
[95] Anthropic_0000554769 at -792, -769, -845–7; J. Kaplan Tr. 317:21–324:14.
[96] *See, e.g.*, Anthropic_0000016458; Anthropic_0000016025.
[97] Anthropic_0000458491.

87. Anthropic promotes Claude as a safe, accurate, high intelligence model that is capable of manipulating a large amount of data.[98] Anthropic also promotes Claude's capabilities in showing its chain of thought or reasoning when doing deep research, as well as its ability to code.[99]

88. Prior to Claude's launch, one internal Anthropic planning document suggests that the company intended for its AI models' use cases and capabilities to include "essay writing" and "answering questions about a book"[100]—use cases similar to the song- and lyric-related uses at issue in this case. At the time Claude was first released to the public in March 2023, Anthropic touted its capability for "a wide variety of conversational and text processing tasks," including "summarization, search . . . Q&A, coding, and more."[101] I understand that Anthropic also specifically promoted Claude's capabilities for "creative and collaborative writing."[102] This capability includes "draft[ing] song lyrics."[103] After Claude was released to the public, Anthropic continued, at least initially, to describe its use cases in similar ways.

89. Anthropic also touted Claude's abilities to retrieve information upon request. For instance, Anthropic noted that "Claude skills" include "information retrieval"[104] and "knowledge retrieval,"[105] among other uses, and that the model can "write at superhuman speed."[106] The U.S. Copyright Office later noted Claude's retrieval-augmented generation ("RAG") capabilities, through which Claude can search outside of its training data to respond to requests, allowing Claude to respond to the question "What show won the Outstanding Drama award at the 2024

---

[98] *See* J. Leibrock Dep. Tr. at 26:6–27:2.
[99] *See id.*
[100] Anthropic_0000444967.
[101] Anthropic_0000349548; *see also* Anthropic_0000013680; Anthropic_0000012356 at -378; Anthropic_0000347289 at -305.
[102] *Id. See also* J. Leibrock Dep. Tr. 48:7–17.
[103] *See* D. Ganguli 30(b)(1) Dep. Tr. 138:22–139:1; *see also* B. Mann 30(b)(6) Dep. Tr. 26:8–24; Anthropic_0000013680.
[104] Anthropic_0000518571.
[105] Anthropic_0000441766; *see also* Anthropic_0000479018.
[106] Anthropic_0000227058.

Emmys?" by sending the query to a third-party search engine and answering the question with the retrieved text.[107]

90. In Claude's own words, it is "an AI without human limitations," and thus has the capabilities for "perfect recall" to precisely identify things like lyrics and musical sections.[108] This functionality is only possible because there is a compressed version of the content in question as tokens inside the trained model. The very term of "perfect recall" implies there is an original copy of the content that is being recalled when the model is generating its response to the user.

91. Indeed, as mentioned above,[109] 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 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.

92. 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.[110] 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

---

[107] U.S. COPYRIGHT OFFICE, *supra* n.2, at 22–23. *See also Retrieval Augmented Generation (RAG) for Projects,* CLAUDE SUPPORT, https://support.claude.com/en/articles/11473015-retrieval-augmented-generation-rag-for-projects (last visited Dec. 18, 2025).
[108] *See* D. Ganguli 30(b)(1) Dep. Ex. 21 (ANTH_SAMPLE000000246).
[109] *Supra*, ¶ 7171.
[110] *See supra,* Sections IV. A. 2–3.

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.[111] An Anthropic employee similarly commented, "I think [the Books3 dataset] includes magazines, textbooks, and niche text people on the internet would want. People on the internet like programming textbooks, so that's my guess at what makes [B]ooks3 good at programming."[112] 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.

93. 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 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)."[113] In response to this query, Claude generated 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."[114] 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.

[111] Anthropic's Opp. to Prelim. Inj. Mot. (Aug. 22, 2024), ECF No. 207.
[112] Anthropic_0000344064.
[113] Anthropic_0000434712.
[114] *Id.*

94. 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,[115] as discussed in more detail in the reports of Publishers' experts Dr. Robert Leonard and Joshua Dennis.

95. 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;[116] generate new song lyrics;[117] and answer questions more generally about specific works or arists.[118]

96. 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.

5. Anthropic's Evaluation of Claude

97. Anthropic's own analysis of Claude usage further corroborates these findings. I understand that monitoring Claude usage and understanding how users actually interact with Claude is important to Anthropic, as it permits Anthropic to measure Claude's impact on users and ensure its models are safe and are being used in accordance with its Acceptable Use Policy.[119] This is a fairly common practice among user facing services, which often tweak or adjust their products to better meet their users' needs and conform to applicable laws or policies. This strongly suggests

[115] See, e.g., Anthropic_0000003118, Anthropic_0000024473, Anthropic_0000027596, Anthropic_0000197273; D. Ganguli 30(b)(6) Dep. Tr. at 94:1–100:3.
[116] *See, e.g.,* Anthropic_0000000395, Anthropic_0000430258, Anthropic_0000430283, Anthropic_0000430272, Anthropic_0000430297.
[117] *See, e.g.,* Anthropic_0000430291, Anthropic_0000431526, Dan. Amodei Dep. Ex. 174 (Anthropic_0000570711), Dan. Amodei Dep. Ex. 175 (Anthropic_0000570714).
[118] *See, e.g.,* Anthropic_0000431532, Anthropic_0000431895, Anthropic_0000430273, Anthropic_0000570712, Dan. Amodei Dep. Ex. 176 (Anthropic_0000570712)
[119] *See* D. Ganguli 30(b)(6) Dep. Tr. 48:6–24; 50:5–14.

that Anthropic is aware of large-scale behavioral trends in its users and how they interact with Claude.

98. Anthropic monitors Claude usage through three main methods: top-down evaluations, red-teaming, and bottom-up evaluations.[120] Anthropic's safeguards team also conducts certain evaluations of Claude usage.[121] In Anthropic's top-down evaluations, researchers will design prompts to look like simple conversations in order to test the models for safety characteristics.[122] Red-teaming is a method by which an LLM model is adversarially probed with prompts designed to "'make the AI behave badly,'" in order to observe potential harms and update the model to eliminate such harms.[123] Anthropic's bottom-up evaluation of Claude usage is done through its "Claude Insights and Observations" ("Clio") platform.[124]

99. Clio reflects aggregated data from Claude conversations with millions of users.[125] Clio data is comprised of "high-level conversations" (*i.e.*, Claude user prompts and completed Claude outputs) which are then reviewed and summarized by Claude itself.[126] Anthropic will then group those conversations together into "clusters" and use Claude to analyze the "high-level gist of what those conversations actually are" about.[127] The clusters do not reflect either a specific prompt or the completed output within a conversation.[128]

100. Anthropic's Clio data and reporting reflects a wide variety of real-world use cases. While Anthropic has claimed that the main use case of Claude products (including Claude.ai and

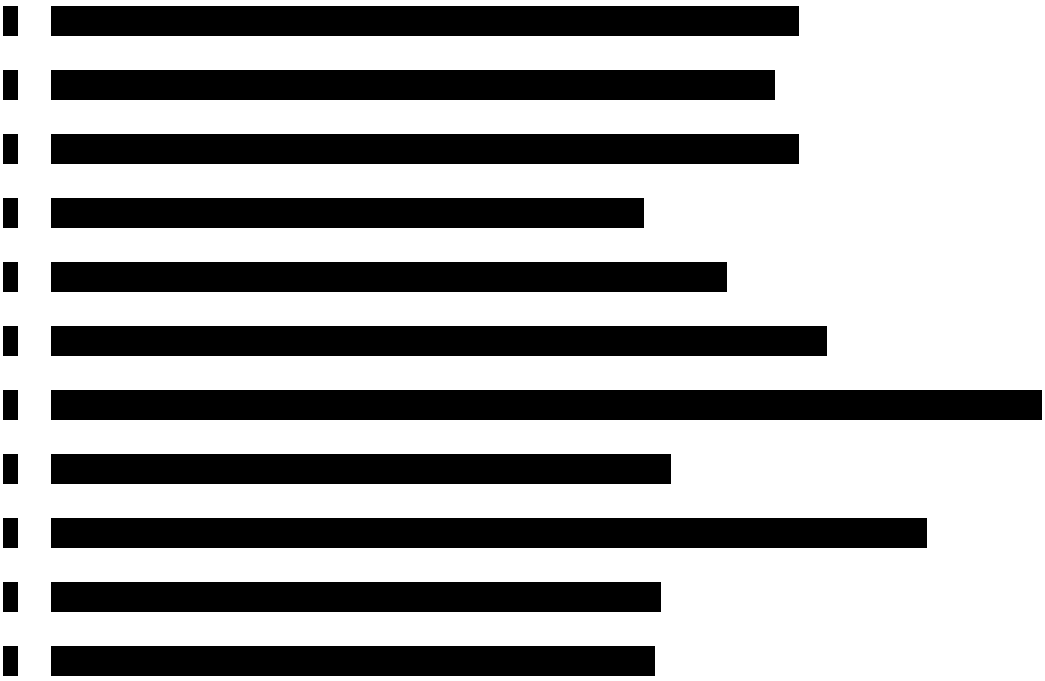
---

[120] *See* D. Ganguli 30(b)(6) Dep. Tr. 47:12–48:1.
[121] *See id.*
[122] *See id.* at 29:1–25.
[123] Deep Ganguli et al., *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, arXiv: 2209.07858v2 (Nov. 22, 2022), https://arxiv.org/pdf/2209.07858.
[124] A. Tamkin, et al., *Clio: Privacy-Preserving Insights into Real-World AI Use*, arXiv: 2412.13678v1 (Dec. 18, 2024), https://arxiv.org/pdf/2412.13678; Anthropic_0000012987.
[125] *See* Anthropic's Supplemental Responses and Objections to Publishers' Third Set of Rogs (Rog No. 18), dated Nov. 5, 2025. *See also* Ganguli 30(b)(6) Dep. Ex. 5 (Anthropic_0000012987).
[126] *See* D. Ganguli 30(b)(6) Dep. Tr. 61:20–62:2.
[127] *Id.* at 62:2–7.
[128] *See id.* at 63:3–5.

Claude API) is writing code, Claude is a general-purpose technology that is used for tasks including content creation, research, business strategy development, translation, and writing.[129]

101. Anthropic's Clio data reflect that users use Claude for a number of music- and lyric-related use cases, including basic song identification and lyric retrieval and original song generation. A number of use cases identified in Anthropic's Clio data relate to songs, music, and lyrics in particular.[130] Based on my review of this data, common song-related uses of Claude include requesting lyrics from Claude,[131] lyric translation,[132] lyric interpretation or analysis,[133] songwriting in a particular style,[134] and writing wholly new songs.[135]

102. By way of further example, a June 2025 spreadsheet reflecting Clio data for a single week reflects a variety of other music- and lyric-related clusters of tasks users have used Claude for, including:



[129] *See id.* at 67:10–22; Ganguli 30(b)(6) Dep. Ex. 5 (Anthropic_0000012987).
[130] *See* D. Ganguli 30(b)(6) Dep. Tr. 68:23–69:4, 105:23–106:3.
[131] *See id.* at 98:5–8.
[132] *See id.* at 100:4–11.
[133] *See id.* at 101:7–9.
[134] *See id.* at 101:10–15, 104:1–17.
[135] *See id.* at 103:21–25.

[redacted]

103. Clio output from several other weeks in May and June 2025 reflect similar clusters; *i.e.*, they reflect a variety of music- and lyric-related prompts submitted by Claude users, including [redacted].[142]

104. I understand that Anthropic does not utilize Clio to analyze whether Claude is being used to request copyrighted materials, nor the extent to which users request song lyrics specifically.[143]

105. Clio suffers from a number of limitations in analyzing Claude usage, as Anthropic acknowledges. Among other things, Clio reflects user patterns as analyzed by Claude itself, not humans.[144] By design, Clio obfuscates the raw conversations between users and Claude as a privacy-preserving measure.[145] Anthropic's human employees do not actually review the conversations between users and Claude that Clio analyzes, nor do they have access to the raw conversations being analyzed.[146] Moreover, Clio is a tool for summarization and generalization; the results of a Clio analysis do not show the actual prompts or output from Claude

[136] *See* Ganguli 30(b)(6) Dep. Ex. 6 (Anthropic_0000429489). *See also* D. Ganguli 30(b)(6) Dep. Tr. 59:1–4; D. Ganguli 30(b)(1) Dep. Tr. 275:7–276:17, 282:6–287:13.
[137] *See* Anthropic_0000429482; Anthropic_0000429483; Anthropic_0000429484; Anthropic_0000429485; Anthropic_0000429486; Anthropic_0000429487; Anthropic_0000429488; Anthropic_0000429490.
[138] *Id.*
[139] *Id.*
[140] *Id.*
[141] *Id.*
[142] *Id.*
[143] *See* D. Ganguli 30(b)(6) Dep. Tr. 52:10–19, 55:9–56:17, 58:8–59:1, 64:6–16.
[144] *See* Anthropic_0000012987 at 2. *See also* D. Ganguli 30(b)(6) Dep. Tr. 31:7–12.
[145] *See* Anthropic_0000012987 at 2–3. *See also* D. Ganguli 30(b)(6) Dep. Tr. 110:16–19.
[146] *See* D. Ganguli 30(b)(6) Dep. Tr. 71:20–72:13, 109:11–15, 110:16–19.

conversations.[147] Clio also has difficulty generating a correct summarization or analysis of a conversation where a single conversation includes multiple different topics or requests.[148] Clio's analysis does not indicate whether Claude's output guardrails are functioning correctly.[149] Clio also cannot definitively exclude specific use cases; *i.e.*, the absence of a particular use case in Clio's output does not confirm that particular use case does not exist.[150] Indeed, in its published paper regarding Clio, Anthropic stated that, "Clio cannot conclusively prove the absence of certain patterns or behaviors," and that, "Clio will only identify patterns across many different conversations, which means it is not useful for identifying rare (but possibly very consequential) patterns, such as a single example of extreme misuse," and which "underscores the importance of complementing Clio's analysis with other methods."[151]

106. In short, Anthropic's Clio data is helpful to the extent it corroborates what I have observed regarding Claude usage based on Claude prompt-output records produced by Anthropic, including regarding song- and lyric-related uses of Claude. However, this data also has significant limitations as noted above, and cannot be used to definitively exclude use cases or identify specific prompts and outputs, such as those seeking or generating specific lyrics.

107. Notably, Anthropic has stated publicly that "[k]nowing how people actually use language models is important for safety reasons."[152] I agree that it is critical for AI companies to understand how their models are being used after deployment. However, prior to Anthroipc's creation of Clio, the Anthropic employee who created Clio, Alex Tamkin, wrote in an internal presentation: "Problem. We have (close to) no idea how Claude is being used in the world."[153]

---

[147] *See supra,* n.25 at 2–3. *See also* D. Ganguli 30(b)(6) Dep. Tr. 62:1–63:5, 112:7–16; D. Ganguli 30(b)(1) Dep. Tr. 186:1–6.
[148] *See* D. Ganguli 30(b)(1) Dep. Tr. 270:20–271:4.
[149] *See* D. Ganguli 30(b)(6) Dep. Tr. 60:6–10.
[150] *See supra,* Anthropic_0000012987 at 13. *See also* D. Ganguli 30(b)(6) Dep. Tr. 127:13–16.
[151] Anthropic_0000012987.
[152] *Id*.
[153] Anthropic_0000524002.

108. Anthropic also employs red-teaming as a way to prevent unwanted output or behavior by Claude. This is a method by which an LLM can be tested with adversarial prompts to probe for harmful outputs and then subsequently updated to avoid such outputs.[154] Anthropic's early red-teaming involved team members having open-ended conversations with an AI model with the goal of making the model "behave badly" and get it to say "obnoxious, offensive, and harmful things."[155] Anthropic's early red-team attacks focused on harms such as soliciting advice on violence, making and smuggling drugs, and racist language.[156] By probing a pre-deployment model for these types of harms, Anthropic can use the data to mitigate the harms prior to release.[157]

109. Based on discovery in this case, Anthropic appears to have not considered copyright infringement a "harm" that needed to be addressed through red-teaming before releasing Claude. An internal October 2023 document explaining Anthropic's red-teaming work up to that point categorized the company's efforts around copyright (*i.e.*, preventing the model from regurgitating copyrighted content) as a category for which "[Anthropic has] only done some work on it, but Claude probably still has many failures modes that aren't hard to find."[158] Red-teaming Claude to recognize copyright as a harm came "with the risk of making the model less creative at writing, or less willing to write creatively."[159]

110. I understand that the lower priority of red-teaming for copyright violations is consistent with Anthropic's interpretation of what a "harmful" result is—to Anthropic, "harmful" usually refers to "things like hate speech, discrimination, harm to people's lives, health, emotional or physical safety, or potential for harm to other people's lives, health, or . . . safety,"[160] not copyright protections.

---

[154] D. Ganguli et al., *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, arXiv: 2209.07858v2 (Nov. 22, 2022), https://arxiv.org/pdf/2209.07858.
[155] *See id.* at 4.
[156] *See id.* at 3.
[157] *See* D. Ganguli 30(B)(6) Dep. Tr. 32:23–33:7.
[158] *See* J. Kaplan Dep. Ex. 239 (Anthropic_0000226164).
[159] *See id.*
[160] *See* B. Mann 30(b)(6) Dep. Tr. 132:4–11.

6. Guardrails

111. In an internal memo from November 2023, Anthropic acknowledges the "axiom" that "AI models should never be shipped without a [Trust & Safety]-like system."[161] I agree that it is critical for AI companies such as Anthropic to develop safeguards to prevent their models from being misused, and that such systems be tested and implemented prior to releasing the models to the public.

112. Prior to the filing of this lawsuit, however, Anthropic lacked sufficient safeguards to prevent the output of copyrighted material or regurgitation of copyrighted material. This is likely due in part to inherent weaknesses in guardrails, as discussed further below. In this regard, a recent study demonstrated Claude 2's vulnerability, in particular.[162] These guardrails did not prevent Claude from reproducing verbatim or near-verbatim versions of Publishers' lyrics to the 499 Works in Suit.[163]

113. Anthropic claims the "guardrails that Anthropic has implemented prevent Claude from including copyrighted material, like song lyrics, in outputs."[164] 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.[165]

114. 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

---

[161] Anthropic_0000477437.
[162] Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models*, arXiv: 2307.15043 (July 27, 2023), https://arxiv.org/pdf/2307.15043.pdf.
[163] *See* FAC, Ex. B.
[164] Anthropic's Amended and Supplemental Responses and Objections to Plaintiffs' First Set of Interrogatories (Nos. 9, 11, 12), dated Nov. 12, 2025, ("Nov. 12, 2025 Interrogatory Responses") at 14.
[165] *See* Anthropic_0000477560 [redacted]; Anthropic_0000217166 [redacted]

significant problem or obstacle to launch.[166] Instead, Anthropic researchers identified reproduction of copyrighted material as [redacted][167] Even after launching Claude, Anthropic's researchers did not prioritize guardrails for the prevention of reproducing copyrighted material, noting in 2024 that they did not specifically train Claude *not* to reproduce copyrighted text.[168]

115. Claude also produced outputs that contained Publishers' lyrics, even when not explicitly asked for those lyrics. For example, when asked to write a poem in the style of Louis Armstrong or to create a song about the death of Buddy Holly, Claude reproduced large portions of the original Publishers' lyrics relevant to those prompts.[169] Regurgitation of lyrics in such instances further reflects the inadequacy of Anthropic's then-existing guardrails.

116. 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."[170] As discussed below, these measures—largely adopted after Claude's public launch and/or in response to Publishers' lawsuit—are still insufficient to effectively prevent output copying Publishers' lyrics and other copyrighted content.

a. Constitutional AI

117. First, 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

[166] *See* J. Kaplan Dep. Ex. 238; J. Kaplan Dep. Tr. 255:4–25; 259:25–260:9.
[167] *See* J. Kaplan Dep. Ex. 238; J. Kaplan Dep. Tr. 246:1–247:16.
[168] *See* J. Kaplan Dep. Ex. 240; Anthropic_0000240665; Anthropic_0000240669.
[169] FAC ¶¶ 97, 102.
[170] Nov. 12, 2025 Interrogatory Responses at 11.

as principles generated by Anthropic itself.[171] 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").[172] Anthropic claims this process ensures that when Claude is ultimately run, it will do its best to obey the constitutional principles.[173] In comparison, OpenAI's ChatGPT models rely more heavily on "Reinforcement Learning with Human Feedback" ("RLHF").[174]

118. Relatedly, I understand Anthropic employs "character training" to further inculcate desired values in Claude. Anthropic's character training involves a [redacted][177]

119. 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 behaviorial guidelines" into the model's "training objectives."[178] 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

[171] B. Mann 30(b)(6) Dep. Tr. 136:21–137:24, 142:20–25, 148:3–24; V. Rao 30(b)(6) Dep. Ex. 84 (Anthropic_0000000485).
[172] Yuntao Bai et al., *Constitutional AI: Harmlessness from AI Feedback* (Dec. 15, 2022), arXiv: 2212.08703, https://arxiv.org/pdf/2212.08073.pdf.
[173] B. Mann 30(b)(6) Dep. Tr. 136:20–137:24.
[174] 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.
[175] *See* B. Mann 30(b)(6) Dep. Tr. 149:13–149:18.
[176] *See id.* at 149:18–21.
[177] *See id.* at 149:22–150:8.
[178] Nov. 12, 2025 Interrogatory Response at 12.

copyright."[179] 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."[180] However, Anthropic claims it only started using this copyright "principle" on [redacted] models to the public.[181] In the absence of such a principle, "Constitutional AI" prior to [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 [redacted] Indeed, Publishers were able to obtain the lyrics to all the Works in Suit after [redacted],[182] demonstrating the ineffectiveness of the Constititutional AI training as a functioning guardrail.

b. Prompt Shield

120. 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." *Id.* at 12. This system is allegedly designed to—when it "detects a prompt that may seek to elicit copyrighted content"—"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 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][183] It also did not implement Prompt Shield on its third-party API product through

[179] *Id.*
[180] *Id.*
[181] *Id.* at 11–12. *See also* Anthropic_000000485.
[182] 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).
[183] Nov. 12, 2025 Interrogatory Responses at 13.

Amazon's Bedrock until [redacted] and its third-party API product through Google's Vertex until [redacted].[184]

121. Anthropic's Prompt Shield also had systematic difficulties addressing so-called [redacted] but *nonetheless* led to Claude models producing copyrighted texts.[185] In other words, Prompt Shield failed to prevent Claude from outputting copyrighted content even when the prompts themselves did not ask for copyrighted content.

122. Recall my earlier explanation for how the parameter weights of a trained LLM encapsulate multi-layers of conditional probabilities, encoding the likelihood of particular sequences of words or tokens following other sequences of words or tokens. 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. 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.[186] 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.[187] Even when

[184] *Id.*
[185] Anthropic_0000227016.
[186] 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.
[187] 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.

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.[188]

123. 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 "slightly at risk" because of a "couple open questions" regarding the Prompt Shield and Copyfairy systems.[189] In particular, initial trials of "Prompt Shield" were "generating a lot of false positives," and Anthropic employees were of the view that Prompt Shield was "not launchable in [its] current state."[190]

124. This is concerning, because for real world deployments of filtering or detection systems, there is always a tradeoff between false positives and false negatives. Given Anthropic (or any LLM provider)'s focus on optimizing the user experience, false positives 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 reduces false positives, and in turn increasing the likelihood of false negatives, *i.e.,* prompts that should be ideally blocked but are missed by the filter. Nevertheless, Anthropic launched Claude 2 to the public in July 2023[191] without first addressing those issues with Prompt Shield.

125. As noted above, Anthropic did not implement Prompt Shield for Claude.ai and its first-party API [redacted][192] Moreover, the issues described above seemed to have affected Prompt Shield globally, and so the

[188] *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.
[189] Anthropic_0000010906.
[190] *Id.*
[191] Kaplan Decl. (Nov. 22, 2023), ECF No. 55-1 ¶ 13.
[192] Nov. 12, 2025 Interrogatory Responses at 13. *See also* V. Rao 30(b)(6) Dep. Tr. 122:12–22.

copyright classifier included in Prompt Shield, if it had been deployed in [redacted], would have been plagued with the same issues.

c. Content Compare System

126. Third, 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").[193] 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."[194]

127. For example, the first iteration of CCS (an updated version of a previous system called "Copyfairy")[195] implemented by Anthropic on or around [redacted][196] I understand that outputs that contained less than the exact threshold of *continous words* would not be flagged as potentially infringing content by CCS.[197] In other words, an output containing only [redacted] continuous words from the lyrics of one of Publishers' works (or any other rightsholders' works) would not have been flagged by this version of CCS.

128. I also understand from Anthropic's testimony that there is no specific reason why [redacted][198] In my opinion, there are no compelling technical reasons why the match threshold had to be set at that specific number. Anthropic could have set the threshold at [redacted] if it wanted to. In fact, Anthropic did decrease the CCS match threshold for [redacted][199]

[193] Nov. 12, 2025 Interrogatory Responses at 13.
[194] *Id.*
[195] V. Rao 30(b)(6) Dep. Tr. 72:8–20.
[196] Nov. 12, 2025 Interrogatory Responses at 13.
[197] *See* V. Rao 30(b)(6) Dep. Tr. at 83:3–12.
[198] V. Rao 30(b)(6) Dep. Tr. at 81:13–16.
[199] Nov. 12, 2025 Interrogatory Responses at 14.

129. 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.

130. Around June 28, 2023, Anthropic determined that "[n]o changes to [CCS] [were] needed" and that "adding copyright reproduction protections into [its] constitutional training seems to be adequate for the moment."[200] However, it is clear to me that Anthropic's pre-suit guardrails were not sufficient at that time, as demonstrated by the fact that Anthropic implemented additional guardrails *after* the suit was filed to prevent Claude from outputting specific copyrighted lyrics. After Publishers filed their initial Complaint, Anthropic only *then* procured and added the lyrics of so-called [redacted] and added them to the hash corpus for CCS.[201] 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."[202] Anthropic initially represented that on [redacted] Anthropic added

[200] Anthropic_0000012438.
[201] Anthropic_0000347810; V. Rao 30(b)(6) Dep. Tr. 96:25–97:3; Nov. 12, 2025 Interrogatory Responses at 13.
[202] Nov. 12, 2025 Interrogatory Responses at 13.

an additional [redacted] to the CCS hash corpus,[203] though it later had to correct that this number was actually [redacted][204]

131. I am not aware of any evidence suggesting Anthropic added lyrics for songs not contained in those datasets to the CCS hash corpus. Based on my technical understanding of CCS, any lyrics that were omitted from the [redacted] datasets 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.

132. My opinions regarding the flaws in Anthropic 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."[205]

d. Defects in Anthropic's Guardrails

133. It is my opinion that Anthropic's mitigation systems 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,[206] Publishers were able to obtain verbatim copies of the 499 Works in Suit after the copyright principle was implemented, using plain language requests in their prompts for those works.[207] Then in February 2024, Publishers identified copies of many works in suit in output collected through Claude Instant using the Poe application[208] —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

[203] *Id.*
[204] Anthropic's Amended Supplemental Responses and Objections to Plaintiffs' First Set of Interrogatories (No. 12), dated Dec. 11, 2025, at 6.
[205] Anthropic_0000242173.
[206] *Supra,* ¶ 119.
[207] *See* FAC at ¶¶ 8–9, 12, 79, 84; FAC Ex. B; *supra* n. 182.
[208] Reply Decl. of Michael Candore, Ex. A, ECF No. 96-1.

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.[209]

134. Further, based on my review of third-party Claude prompts and output, Claude still does not prevent users from requesting putatively original lyrics or even requests for songs written in the style of certain artists or songwriters.[210]

135. 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] songs against which to compare potential outputs. Given the millions of commercially produced songs released each year,[211] 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.

136. 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 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.

---

[209] *See, e.g.,* Reply Decl.of Timothy Chung, Ex. B, ECF No. 93-2; Plaintiffs_0000009379 –9500; Plaintiffs_0000009921–10345; Expert Report of Dr. Leonard (Dec. 19, 2025) ("Leonard Expert Report").
[210] *See, e.g.,* Plaintiffs_0000009921–10345; Leonard Expert Report; Dennis Expert Report).
[211] *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/.

137. Case in point, I understand that Anthropic very recently discovered on December 4, 2025 that the lyrics for one of the Works in Suit, "Joy to the World," performed by Three Dog Night, were not actually in the CCS hash corpus, and that the lyrics to ***another*** song titled "Joy to the World," were instead added to the CCS hash corpus.[212] Anthropic represents the correct "Joy to the World" has now been added to the hash corpus. Days later on December 11, 2025, Anthropic also disclosed the fact that on [redacted], Anthropic did not, in fact, add [redacted] to CCS' hash corpus, but instead added only [redacted], which is thousands fewer than previously claimed.[213]

138. 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 adding the correct lyrics in a timely fashion and keeping track of which lyrics are actually in the hash corpus—something Anthropic appears to be struggling with even today.

139. For the reasons described above, I do not believe Anthropic's other technological systems described above can ever fully prevent the output of copyrighted content. 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.[214]

7. Hallucinations

140. Another issue that is common to Claude and other LLMs is hallucinated outputs.[215] Hallucination refer to the phenomenon where an LLM generates outputs that are factually incorrect or nonsensical but appear coherent. For example, Anthropic's counsel used Claude to generate an erroneous citation in one of its filings in this litigation.[216] Although that citation appeared coherent,

[212] Defendant's Notice re Guardrails Update, ECF No. 525.
[213] Anthropic's Am. Suppl. Resp. and Obj. to Plfs. First Set of Rogs (No. 12), at 6.
[214] *See* 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.").
[215] *See, e.g.*, D. Ganguli 30(b)(1) Dep. Tr. 218:2–5.
[216] *See* Decl. of Ivana Dukanovic Related to ECF No. 365, ECF No. 371.

the information contained in the citation was fabricated.[217] Other examples of hallucinations may also include instances where Claude incorrectly attributes the author of a work, or erroneously combines lyrics from multiple works while representing they are part of the same work.

141. Hallucinations can be attributed to the inherent design and functionality of LLMs, including Claude: Claude is trained to predict the next word based on language patterns it has learned from texts in its training corpus. Claude, like all LLMs, does not rely on an inherent understanding of truth or facts in the same way humans do, but rather generates output based on learning statistical associations between tokenized text.

142. In the case of Claude, which Anthropic has promoted as an LLM that excels at creative tasks, hallucination-like behavior is desirable and is likely encouraged by Anthropic. That is because if Anthropic trains its models to be cautious and avoid hallucinations, the models may become less useful for creative tasks and open-ended requests. Reducing hallucinations generally means reducing flexibility and fluency. Despite years of claims by AI researchers that improvements in model training processes and data will remove hallucinations, the community is slowly recognizing that hallucinations are quite likely fundamental to the transformer architecture and current methods of training on scraped data. This was in fact the final conclusion of a recent research publication by researchers at OpenAI.[218] This is an inherent quality of LLMs and it is my opinion that hallucinations are still a significant hurdle that Anthropic has not yet cleared. As one Anthropic employee testified, although "older versions [of LLMs] hallucinated more than the current versions . . . language models can still hallucinate."[219] In an earlier interview, that same Anthropic employee put it more bluntly: "Model developers call these 'hallucinations,' which is

---

[217] *Id.*

[218] *Gyana Swain*, OpenAI Admits AI Hallucinations Are Mathematically Inevitable, Not Just Engineering Flaws, COMPUTER WORLD (Sep. 18, 2025), https://www.computerworld.com/article/4059383/openai-admits-ai-hallucinations-are-mathematically-inevitable-not-just-engineering-flaws.html; Adam Tauman Kalai et al., *Why Language Models Hallucinate*, arXiv:2509.04664 (Sep. 4, 2025), https://arxiv.org/pdf/2509.04664.

[219] Ganguli 30(b)(1) Dep. Tr. 218:2–5.

an interesting term for me. But you know, another term for this, is that the model is just lying in a factual context."[220]

V. LLMs and Copyrighted Content

143. Respecting copyright while building an LLM model is reasonable and feasible. Of course, an AI developer can seek licenses from copyright holders for their content. But even without taking licenses, the developer can (1) avoid training AI models on unlicensed copyrighted materials (including both "pirate" content not legally obtained by the website/database and material that may have been licensed from the copyright owner by the site operator but was not licensed by the AI developer); and (2) implement effective guardrails.

144. There are numerous available means by which LLM developers can avoid training their AI models on unlicensed copyrighted materials. For instance, model developers can avoid scraping from sources known to contain unlicensed copyrighted material, such as Books3, which is a massive database of nearly 200,000 books downloaded from pirate sources.[221] Model developers can also look out for copyright identifiers on the materials they use in training, such as the copyright symbol ("©"), authors' or publishers' names, copyright dates, and other related indicia of copyright protection, and avoid including such material in training. Some of that material may be licensed for public use through Creative Commons licenses, but that should be clearly indicated. Developers can also utilize databases of copyrighted works as a "blacklist" to exclude from the training data any instance of those works found in that database. Additionally, developers can seek and secure licenses to include copyrighted content in their training data, either by licensing from content owners directly or by using training datasets of pre-licensed and cleared data. By avoiding training their AI models on unlicensed copyrighted materials, LLM developers

---

[220] *In the Dark Room: Deep Ganguli & Alan Michael Parker*, Identity Theory (Aug. 21, 2023), https://www.identitytheory.com/in-the-dark-room-deep-ganguli-alan-michael-parker/; D. Ganguli 30(b)(1) Dep. Tr. 131:10–23.

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

can avoid copying those materials in the output of their models. All of these options are and have been available to Anthropic.

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

146. First, content-based blockers are known to be fragile, and can be bypassed by a user with sufficient knowledge of LLMs. Recent research published last year at the top machine learning conference, the International Conference on Machine Learning ("ICML"), demonstrates fundamental flaws in the way transformer language models are built, such that there are deterministic ways to compute non-sensical characters, which, when added to a prompt, will bypass prompt blockers, and allow the query to be answer "truthfully" by the LLM.[222] This work also demonstrates that these vulnerabilities are fundamental to the transformer architecture and are applicable to all commercial LLM models. It is also not a property or weakness that can be easily addressed. Transformer architecture is inherently vulnerable to these attacks and simply cannot be "debugged" to prevent infringing outputs in all cases.[223] Accordingly, while guardrails have their place in a well-designed LLM, they are inherently limited and research suggests that they can always be bypassed.[224]

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

---

[222] Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models* (July 27, 2023), arXiv: 2307.15043, https://arxiv.org/pdf/2307.15043.pdf; Zico Kolter, Keynote Presentation at the 2nd ICML Workshop on New Frontiers in Adversarial Machine Learning: Adversarial Attacks on Aligned LLMs (Jul. 28, 2023).
[223] *See supra,* n.218218.
[224] *See* Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models*, arXiv:2307.15043 (Dec. 20, 2023), https://arxiv.org/pdf/2307.15043.

from the training set. Output-based guardrails also do not prevent users from requesting, and Claude producing, outputs containing putatively original lyrics or lyrics written in the style of a certain artist or songwriter.[225]

148. 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.

VI. Remedy

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

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

151. As discussed above, Publisher's lyrics cannot be removed from the current versions of Claude, given those models have already been trained on the lyrics. But that data can be removed from Claude's dataset the next time Anthropic trains and launches a new version of the model—if Anthropic chooses to do so. This is perfectly reasonable. Commercial LLM models such as Claude are periodically re-trained and re-launched from scratch. For example, in recent years, when updating Claude 1 to Claude 2, and Claude 2 to Claude 3, Claude 3 to 4, and so on, Anthropic had to create and train its model on new datasets each time.

[225] *See, e.g.,* Plaintiffs_0000009921–10345.

152. I understand Anthropic claims that removing specific works, such as Publishers' lyrics, from its training datasets, is technically infeasible due to the large scale of data in question, the amount of processing needed, and the "complexity of what the definition of a given piece of content is."[226] This is incorrect.

153. Unlike most other types of data, song lyrics are almost always simple text of a limited length. There are a wide range of techniques that can be applied to detect song lyrics in a large corpora of data.[227] The most simple, trivial and fast method is to simply apply the Grep tool using a regex (regular expression) filter that captures key substrings or keywords in large datasets. This is a common unix tool that is introduced to first year computer science majors. A simple Google query for "using grep to find text in files" will return hundreds of tutorial pages, Howto pages, and YouTube videos showing the use of Grep and regular expressions to quickly and efficiently scan through large datasets for specific substrings. The baseline benchmark results for Grep suggest that it processes 1-2GB per second per machine. Today, the largest and most complete snapshot of the entire Internet (Common Crawl November 2025)[228] includes 2.29 Billion pages, roughly 378 TB in size. Rough estimates (assuming 1.5GB per second speed), would suggest that a single server can search through the entire snapshot of the Internet in roughly 71 hours. Put another way, you can parallelize the same search task across a 1000-node CPU cluster, and complete the same search in 4.3 minutes. Note that with large-scale data centers, Anthropic (or any AI company) can easily reduce this query cost to a matter of a few seconds.

154. While this may seem like a significant cost to compute for each song, it is worth noting that this only needs to be done ***once*** (*i.e.,* older webpages do not need to be scanned repeatedly for the same lyrics). Once a search for lyrics has been completed for a single snapshot of the training dataset, only new additions to the training data need to be searched. Finally, it is worth noting that this assumes the simplest, non-optimized search. In practice, a search for song

[226] *See* B. Mann 30(b)(6) Dep. Tr. 97:11–22.
[227] *See, e.g.,* Shan Expert Report, Section VI.
[228] Malte Ostendorff, *November 2025 Crawl Archive Now Available*, COMMON CRAWL (Nov. 23, 2025), https://commoncrawl.org/blog/november-2025-crawl-archive-now-available.

lyrics can be much more efficient, using optimizations that combined similar lyrics in a single scan, or utilized an efficient implementation using GPUs, or using fast hashing techniques like fuzzy fingerprint hashes.[229]

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

VII. Risks from AI

156. AI development poses significant risks to a number of different stakeholders, ranging from the disruptive to the catastrophic. Anthropic and its employees have opined extensively on these risks, and I have conducted significant research in this area over the course of the last three years, publishing a number of peer-reviewed articles in premier publication venues in security and machine learning,[230] and receiving multiple awards in the process.[231] My work has largely focused on how to provide solutions that quantify and mitigate such harms, particularly to human creatives such as artists. In addition, I have conducted multiple meetings with different

---

[229] *See* Feng Zhou et al., *Approximate object Location and spam filtering on Peer-to-peer systems,* PROC. OF ACM/IFIP/USENIX INTL. MIDDLEWARE CONF. (2003), https://disco.ethz.ch/courses/ws0506/seminar/papers/peer_spam_filtering.pdf; *see also* Udi Manber, *Finding Similar Files in a Large File System*, Winter USENIX Technical Conference (1994), https://www.cs.princeton.edu/courses/archive/spr05/cos598E/bib/manber94finding.pdf.

[230] *See, e.g.*, Ben Y. Zhao, et al., *Somesite I Used to Crawl: Awareness, Agency and Efficacy in Protecting Content Creators from AI Crawlers*, ACM Internet Measurement Conference 2025 (Oct. 2025), https://people.cs.uchicago.edu/~ravenben/publications/pdf/crawlers-imc25.pdf; Ben Y. Zhao, et al., *Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?* PROC. OF THE ACM CONF. ON COMPUTER AND COMM'N SEC. 2024 (Oct. 2024), https://people.cs.uchicago.edu/~ravenben/publications/pdf/organic-ccs24.pdf; Ben Y. Zhao, et al., *Nightshade: Prompt-Specific Poisoning Attacks on Text-to-Image Generation Models*, PROCEEDINGS OF THE 45TH IEEE SYMPOSIUM ON SECURITY AND PRIVACY (May 2024), https://arxiv.org/pdf/2310.13828; etc.

[231] *See, e.g. The 100 Most Influential People in AI 2024*, TIME Magazine, https://time.com/7012857/ben-zhao/ (awarded to Ben Zhao, inventor of Nightshade); *CCS 2024 Distinguished Paper Awards*, Association for Computing Machinery Special Interest Group on Security, Audit and Control, https://www.sigsac.org/ccs/CCS2024/program/awards.html (awarded to Ben Zhao, and others, for the paper "Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?"); etc.

stakeholders to help them understand and address some of these harms produced from the usage of LLMs, including federal regulatory agencies (U.S. Copyright Office, Federal Trade Commission, Dept. of Defense, Dept. of State), multiple universities, multiple labor organizations representing artists, actors, writers and musicians, and state legislatures in Illinois and California.

157. Since their arrival, LLMs have demonstrated through real events that unregulated and uncontrolled AI can inflict a variety of harms on broad subsets of the population.[232] These harms are real and measurable today, even though many of Anthropic's justifications for its products and potential benefits for humanity remain unrealized years after its first release of Claude. All this stands in stark contrast to Anthropic's stated mission to help mankind through AI, and its goal of differentiating itself from its competitors by more ethical design and practice.

158. AI models are capable of having more net knowledge than any human, and that knowledge can be exploited by humans to create extremely dangerous things, such as bioweapons or nuclear bombs.[233] In light of all these present-day harms, many are openly questioning the value of LLMs and their societal harms, particularly their cumulative, long-term risks to science, education and shared social truth in democratic societies.[234]

159. One of the major concrete harms of LLMs is the massive proliferation of what is now commonly called "AI slop," *i.e.*, automated content that has little meaning other than being low qualify mimics of human created content. In the early days of LLM usage, one of the first organizations to disclose harms from LLMs was writers. Specifically, writer communities who are being inundated with a massive volume of AI-generated text. For example, in February 2023, Clark's World, the sponsor of the well-regarded competition for fictional writing, announced that

---

[232] *See, e.g.*, J. Kaplan Dep. Tr. 85:10–15.
[233] *See* D. Ganguli 30(b)(1) Dep. Tr. 80:22–81:16.
[234] *Large Language Models Pose a Risk to Society and Need Tighter Regulation, Say Oxford Researchers*, OXFORD UNIVERSITY (Aug. 7, 2024), https://www.ox.ac.uk/news/2024-08-07-large-language-models-pose-risk-society-and-need-tighter-regulation-say-oxford; Sandra Wachter et al., *Do Large Language Models Have a Legal Duty to Tell the Truth?*, 11 (8) R. SOC. OPEN. SCI. (2024), https://royalsocietypublishing.org/rsos/article/11/8/240197/92624/Do-large-language-models-have-a-legal-duty-to-tell.

they would discontinue their competition because they received many times greater submissions than they could possibly evaluate, and the overwhelming majority of them were quickly generated AI-text samples.[235]

160. Similar trends were quickly found on more commercial platforms like Amazon, where the vibrant book market was flooded with unlabeled, low quality AI-books that were turning away customers who could not immediately identify which were human written books without first paying for the purchase.[236] Beyond just outputting an overwhelming volume of AI slop, it is clear that LLMs were overwhelmingly used by bad actors to target and take away profits from individual authors by replicating, imitating, or summarizing human written books and trying to displace them on the marketplace.[237] Unfortunately, without significant changes by AI model trainers, this wave of AI slop is unlikely to stop any time soon.[238]

161. When we consider the cost of this large volume of AI slop, it is critical to understand that they do not come for free. Anthropic (and other AI companies) burn large quantities of energy in order to generate the wave of AI slop now overwhelming the Internet.[239] There are massive costs involved in both training and operating LLM models like Claude. Research studies estimate that training a model such as GPT-4 used about 44 GWh, which is

---

[235] Vincent Acovino et al., *A Sci-Fi Magazine Has Cut Off Submissions After a Flood of AI-Generated Stories*, NPR (Feb. 24, 2023), https://www.npr.org/2023/02/24/1159286436/ai-chatbot-chatgpt-magazine-clarkesworld-artificial-intelligence.
[236] Victor Tangermann, *Amazon is Being Flooded With Books Entirely Written by AI*, FUTURISM (May 6, 2023), https://futurism.com/the-byte/amazon-flooded-books-written-by-ai.
[237] *See, e.g.,* Kate Knibbs, *Scammy AI-Generated Book Rewrites Are Flooding Amazon*, WIRED (Jan. 20, 2024), https://www.wired.com/story/scammy-ai-generated-books-flooding-amazon/; Will Oremus, *He Wrote a Book On a Rare Subject. Then a ChatGPT Replica Appeared on Amazon*, WASH. POST (May 5, 2023), https://www.washingtonpost.com/technology/2023/05/05/ai-spam-websites-books-chatgpt/;
[238] *See* Max Read, The Internet's AI Slop Problem Is Only Going to Get Worse, N.Y. MAG. (Sept. 25, 2024), https://nymag.com/intelligencer/article/ai-generated-content-internet-online-slop-spam.html.
[239] Nidhal Jegham et al., *How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference*, arXiv:2505.09598 (Nov. 24, 2025), https://arxiv.org/pdf/2505.09598.

enough to power San Francisco for days, or a small city for a month.[240] Current estimates put a single LLM query at around 0.3 Wh, roughly equivalent to running a desktop computer for 36 minutes. Another research study finds that each Claude query produces about 3.5 grams of CO2, roughly equivalent to watching 35 minutes of YouTube videos.[241]

162. Looking at the broad impact on society, Claude and LLMs are also disrupting education at every level. LLMs such as Claude provide an easy way for students of all ages to quickly generate text for homework for a variety of subjects, particularly research and creative writing assignments, where Claude can quickly generate entire essays in seconds with minimal prompting from a student.[242] This is already having a dramatic impact on K-12 schools, where AI cheating runs rampant, particularly for school districts with limited resources to monitor and detect such behavior.[243]

163. The widespread and unregulated proliferation of LLMs is also making waves among higher education as well. For example, a recent study shows that many or most college students have used LLMs in some way, with many admitting that LLMs have replaced the learning process in numerous classes.[244] High profile cases of student plagiarism at universities are also becoming increasingly common, including among the most selective and prestigious schools.[245] Educators across the world are now grappling with the challenge of adapting to teaching in a world

---

[240] James O'Donnell and Casey Crownhart, *We Did the Math on AI's Energy Footprint. Here's the Story You Haven't Heard*, MIT TECH. REV. (May 20, 2025), https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/.

[241] Christine Hall, *What's Your Chatbot's Carbon Footprint?*, FOSSFORCE (Apr. 3, 2025), https://fossforce.com/2025/04/whats-your-chatbots-carbon-footprint/.

[242] *See, e.g.*, D. Ganguli 30(b)(1) Dep. Tr. 193:9–195:17, 196:23–202:6.

[243] *Confronting AI Cheating in K-12 While Embracing Responsible Innovation*, EDCIRCUIT (Sep. 3, 2025), https://edcircuit.com/confronting-ai-cheating-in-k-12-while-embracing-responsible-innovation/; *see also, e.g.*, D. Ganguli 30(b)(1) Dep. Tr. 193:9–195:17, 196:23–.197:22

[244] James D. Walsh, *Everyone is Cheating Their Way Through College,* N.Y. MAG. (May 7, 2025), https://nymag.com/intelligencer/article/openai-chatgpt-ai-cheating-education-college-students-school.html.

[245] *See, e.g.,* Angela Yang, *Kicked Out of Columbia, This Student Doesn't Plan to Stop Trolling Big Tech With AI*, NBC NEWS (Mar. 27, 2025), https://www.nbcnews.com/tech/tech-news/columbia-university-student-trolls-big-tech-ai-tool-job-applications-rcna198454.

where cheating via AI tools is common.[246] Attempts to detect AI-based plagiarism is extremely challenging, and schools are struggling with the need to balance academic integrity with the potential of harming innocent students with plagiarism charges stemming from false positives in AI-detection tools.[247]

164. Beyond the classroom, the use of Claude and other LLMs poses risks of physical harm to young people by offering tools and advice for self-harm. Because LLMs scrape and train on content indiscriminately, they absorb facts about dangerous substances, as well as mechanisms for self-harm. While Claude tries to block some of these subject matters from user prompts, research has found that these guardrails are trivial to overcome. For example, a recent study from Northeastern University found Claude readily provided specific information about self-harm.[248] In a more formal study published in Psychiatric Services, a journal of the American Psychiatric Association, researchers found that Claude answered questions about ropes, firearms, or poisons and their success rate in self-harm.[249]

165. LLMs are also effective tools in generating misinformation campaigns online. LLMs are found to be extremely convincing in producing misinformation campaigns (*e.g.*, politically-motivated misinformation attacks, conspiracy theories), and at the same time, have very

---

[246] Alice Evans, *How Should Academics Respond to AI?*, GGD (May 19, 2024), https://www.ggd.world/p/how-should-academics-respond-to-ai.
[247] *See* Will Coldwell, '*I Received a First But It Felt Tainted and Undeserved': Inside the University AI Cheating Crisis*, THE GUARDIAN (Dec. 15, 2024), https://www.theguardian.com/technology/2024/dec/15/i-received-a-first-but-it-felt-tainted-and-undeserved-inside-the-university-ai-cheating-crisis; *AI Detectors Don't Work. Here's What to Do Instead*, MIT MGMT STS TEACHING & LEARNING TECH., https://mitsloanedtech.mit.edu/ai/teach/ai-detectors-dont-work/.
[248] Cody Mello-Klein, *New Northeastern Research Raises Concerns Over AI's Handling of Suicide-Related Questions*, NORTHEASTERN GLOBAL NEWS (July 31, 2025), https://news.northeastern.edu/2025/07/31/chatgpt-suicide-research/.
[249] Matt O'Brien and Barbara Ortutay, *Study Says AI Chatbots Need to Fix Suicide Response, As Family Sues Over ChatGPT Role in Boy's Death*, AP NEWS (Aug. 26, 2025), https://apnews.com/article/ai-chatbots-selfharm-chatgpt-claude-gemini-da00880b1e1577ac332ab1752e41225b.

limited effectiveness when used as a detector for misinformation.[250] Indeed, native halluncinations generated by LLMs are difficult to detect.[251] In a recent paper, Anthropic disclosed that during a stress-test of a number of AI models, including its own Claude, the models resorted to "malicious insider behaviors when that was the only way to avoid replacement or achieve their goals—including blackmailing officials and leaking sensitive information to competitors."[252]

166. Finally, it is worth noting that LLMs pose significant risks to entire populations and nation states, by enabling AI agents that can simplify and automate the generation of cyber attacks. Again, these are not hypothetical, but real harms happening today. In the research realm, a number of groups have tested and evaluated LLMs' vulnerability to be taken over and weaponized by attackers. The most recent study conducted by Carnegie Mellon on realistic enterprise network environments showed that LLMs are easily weaponized to autonomously plan and execute complex, multi-step network attacks without human instruction.[253] Another study tested AI agents built on existing LLMs, including Claude 4, and showed that such agents could be compromised using a range of attacks (*e.g.*, prompt-injection, attack via RAG systems and inter-agent interactions). These agents would accept and execute malicious commands such as installing and executing malware on victim machines.[254]

---

[250] Yulong Ma et al., *Linguistic Features of AI Mis/Disinformation and the Detection Limits of LLMs*, NATURE COMMC'NS. (DEC. 11, 2025), https://www.nature.com/articles/s41467-025-67145-1_reference.pdf;

[251] *See* Sebastian Farquhar, et al., *Detecting hallucinations in large language models using semantic entropy*, NATURE 630 (2024), 625, https://www.nature.com/articles/s41586-024-07421-0.pdf.

[252] *Agentic Misalignment: How LLMs could be insider threats,* ANTHROPIC (June 20, 2025), https://www.anthropic.com/research/agentic-misalignment.

[253] Michael Cunningham, *When LLMs Autonomously Attack*, CARNEGIE MELLON UNIVERSITY COLLEGE OF ENGINEERING (Jul. 24, 2025), https://engineering.cmu.edu/news-events/news/2025/07/24-when-llms-autonomously-attack.html. *See also* Brian Singer et al., *Incalmo: An Autonomous LLM-Assisted System for Red Teaming Multi-Hose Networks*, arXiv:2501.16466 (Nov. 22, 2025), https://arxiv.org/pdf/2501.16466.

[254] *See* Matteo Lupinacci, et al, *The Dark Side of LLMs Agent-based Attacks for Complete Computer Takeover*, arXiv:2507.06850v3 (Jul. 11, 2025), https://arxiv.org/pdf/2507.06850v3.

167. In another recent report, Anthropic disclosed that in September 2025, a Chinese state-sponsored group utilized Claude to conduct a "highly sophisticated cyber espionage operation."[255] The espionage group used Claude Code to support reconnaissance, vulnerability discovery, credential harvesting, data analysis, and exfiltration operations.[256] Notably, the "AI autonomously discovered vulnerabilities in targets selected by human operators and successfully exploited them in live operations, then performed a wide range of post-exploitation activities from analysis, lateral movement, privilege escalation, data access, to data exfiltration," without the need for significant human intervention.[257] In each case, Claude code was used across multiple phases of the attacks, including reconnaissance (mapping target networks and identifying high-value systems), vulnerability and exploitation (generating and testing exploit code for discovered vulnerabilities), credential harvesting (systematically extracting and testing credentials), data exfiltration (querying databases, extracting sensitive data including personal records, financial information, and government credentials, and categorizing it by intelligence value), and documentation (generating detailed attack documentation to enable seamless handoff and future campaign resumption).[258] A major saving grace in this case was that Claude also suffered from hallucinations, which limited the level of autonomy in these attacks and required human input in some cases.

168. Finally, there is the significant risk that AI will pose to large swathes of the global economy and job market. I understand that Anthropic's CEO has opined extensively on the likelihood that AI will replace a majority of entry level jobs and lead to massive unemployment

[255] Anthropic, *Disrupting the first reported AI-orchestrated cyber espionage campaign* (Nov. 2025), https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf.
[256] *Id.*
[257] *Id.*
[258] Mary Cunningham, *Anthropic Says Chinese Hackers Used Its Claude AI Chatbot in Cyberattacks*, CBS NEWS (Nov. 14, 2025), https://www.cbsnews.com/news/anthropic-chinese-cyberattack-artificial-intelligence/.

within the next five years, calling it a potential "bloodbath"[259] and warning that "this is already happening."[260] I understand that other Anthropic employees have expressed similar views on the high likelihood that such "knowledge work" will be automated through AI within the next five years.[261] I agree that the possibility of these things occurring is high, including for the reasons discussed above. But regardless of whether LLMs become sophisticated enough to replace skilled workers, it is true that they are already disrupting and replacing entry-level positions in a number of industries. In doing so, they are breaking the skill training pipeline for human workers, making it much more challenging for any entrants to the workforce from learning the skills necessary to become an expert.[262]

[259] Jim VandeHei and Mike Allen, *Behind the Curtain: A white-collar bloodbath*, AXIOS (May 28, 2025), https://www.axios.com/2025/05/28/ai-jobs-white-collar-unemployment-anthropic. *See also* Anderson Cooper and Nichole Marks, *Why Anthropic CEO Dario Amodei spends so much time warning of AI's potential dangers*, CBS NEWS (Nov. 16, 2025), https://www.cbsnews.com/news/anthropic-ceo-dario-amodei-warning-of-ai-potential-dangers-60-minutes-transcript/.
[260] Samantha Delouya, *AI CEO says technology 'moving very quickly,' could soon replace more jobs*, CNN (Sept. 17, 2025), https://www.cnn.com/2025/09/17/business/anthropic-warns-ai-could-soon-replace-jobs.
[261] *See* J. Kaplan Dep. Ex. 245 (Anthropic_0000428963).
[262] Walther, Cornelia C., *Is AI Pushing Us to Break the Talent Pipeline?*, KNOWLEDGE AT WHARTON (Aug. 12, 2025), https://knowledge.wharton.upenn.edu/article/is-ai-pushing-us-to-break-the-talent-pipeline/.

Dated: December 19, 2025

Ben Zhao