EXHIBIT 147

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

REPLY EXPERT REPORT OF DR. BEN ZHAO

February 13, 2026

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

TABLE OF CONTENTS

I. OVERVIEW AND SCOPE OF ENGAGEMENT 1

II. DR. BILMES'S CLAIM THAT LLMS ARE NOT DESIGNED TO MEMORIZE DOES NOT ALIGN WITH THE SCIENTIFIC LITERATURE 2

A. Anthropic's "Goals" for Claude Do Not Alter Fundamental LLM Traits 2

B. Memorization Is Inherent, Not Aberrational 8

C. Dr. Bilmes Overlooks Evidence That Claude Did Memorize and Regurgitate Publishers' Lyrics 10

D. LLM Training Involves the Compression of Data 12

E. Post-Training Guardrails Do Not Sufficiently Address Memorization 14

F. Anthropic's Hugging Face Data Teaches Claude that Humans Prefer Certain Lyrics-Related Responses 15

III. DR. BILMES INCORRECTLY CLAIMS HALLUCINATIONS ARE NOT INHERENT TO LLMS 19

A. Scientific Literature Supports the View that Hallucinations Are Fundamental to LLM Architecture 20

B. Hallucination-like Behavior Can Be Desirable, but there Are Trade-Offs 25

IV. DR. BILMES AGAIN MISTAKENLY ASSERTS ROBOTS.TXT FILES GRANT ANTHROPIC PERMISSION TO CRAWL AND SCRAPE DATA 26

V. DR. BILMES MISREPRESENTS ANTHROPIC'S ABILITY TO IMPLEMENT MEASURES TO AVOID COPYING AND REGURGITATION OF PUBLISHERS' LYRICS 29

VI. DR. BILMES'S GUARDRAIL EXAMINATIONS REMAIN NARROW IN SCOPE AND CRITICALLY FLAWED IN METHODOLOGY 37

C. Dr. Bilmes's Contention that Publishers' Lyrics Are Immaterial for Regurgitation or Performance Misrepresents My Opinions. 43

VII. DR. ALTMAN UNDERSTATES THE HARMS POSED BY LLMS AND THE CHALLENGES IN ADDRESSING THEM 45

A. AI Safety Problems Are Not Adequately Addressed by Simple Changes in Policy or Governance 45

B. Historical Analogs Are Not Probative of the Unprecedented Risks of LLMs 47

I. Overview and Scope of Engagement

1. On December 19, 2025, I issued an initial expert report in the litigation in the caption of this rebuttal that included opinions concerning Anthropic's large language model ("LLM") known as Claude, 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.[1] On January 23, 2026, I issued a rebuttal expert report that included my assessment and opinions of the initial Expert Reports of Dr. Jeffrey Bilmes and Dr. Russ Altman served by Anthropic on December 19, 2025.[2] On the same day that I served my rebuttal report, Anthropic served a series of expert reports, including rebuttal reports from Dr. Jeffrey Bilmes[3] and Dr. Russ Altman.[4]

2. As part of my retention by Publishers, I have been asked to review and comment on sections of the rebuttal expert reports of Dr. Bilmes and Dr. Altman that are relevant to my work and expertise in this matter. In this reply expert report, I provide my assessment of and responses to the Bilmes and Altman Rebuttals.

3. 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.

4. 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,

---

[1] Expert Report of Dr. Ben Zhao (Dec. 19, 2025) (hereinafter "Zhao Report").
[2] Rebuttal Expert Report of Dr. Ben Zhao (Jan. 23, 2026) (hereinafter "Zhao Rebuttal").
[3] Expert Rebuttal Report of Dr. Jeffrey Bilmes, PhD (Jan. 23, 2026) (hereinafter "Bilmes Rebuttal").
[4] Expert Rebuttal Report of Dr. Russ Altman (Jan. 23, 2026) (hereinafter "Altman Rebuttal").

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.

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

6. My credentials, qualifications, and compensation are set forth in my initial expert report. 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. Dr. Bilmes's Claim that LLMs Are Not Designed to Memorize Does Not Align with the Scientific Literature

7. Dr. Bilmes alleges that my initial report relied on "nonstandard" definitions and descriptions concerning LLM training and related concepts. This is incorrect. Moreover, Dr. Bilmes's own reports suffer from such "nonstandard" terms and definitions.[5] In the sections below, I explain why my descriptions of LLM training and related concepts are consistent with scientific and academic literature, and how Dr. Bilmes's characterizations of LLM training "goals," LLM generalization, and LLM architecture deploy nonstandard frameworks to promote a narrative not endorsed by fellow experts in the computer science and artificial intelligence fields.

A. Anthropic's "Goals" for Claude Do Not Alter Fundamental LLM Traits

8. Dr. Bilmes suggests I imply that "memorization is the goal of LLM training"[6] and in turn claims that "generalization, not memorization, is the primary goal of training any machine learning model."[7] This is a distortion and misinterpretation of my opinions.

9. First, my description of how LLMs work at a high level does not conflict with the scientific literature or even Dr. Bilmes's characterization of the "goals" of LLM training. I state in

[5] For example, Dr. Bilmes insists on characterizing Anthropic's pre-training and post-training of Claude as "transformative." Expert Report of Jeffrey Bilmes, Ph.D (Dec. 19, 2025) (hereinafter "Bilmes Report"), Section VI. *See also* Bilmes Rebuttal ¶ 49. As I described in my rebuttal report, the term "transformative" carries no meaning in my field; Dr. Bilmes uses a wholly nonstandard and nontechnical legal framework to characterize pre-training and post-training.

[6] Bilmes Rebuttal ¶ 12.

[7] Bilmes Rebuttal ¶ 12.

my initial report that "[i]n 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."[8] This is entirely consistent with scientific literature, which defines large language models as "systems which are trained on string prediction tasks: that is, predicting the likelihood of a token (character, word or string) given either its preceding context or (in bidirectional and masked LMs) its surrounding context."[9] Other examples of seminal papers on LLMs, including one written by Anthropic's co-founder, define the training objective as predicting the next token in a sequence by modeling the joint probability distribution of text.[10]

10. Although Dr. Bilmes concedes that "LLMs are designed to learn patterns from the training data" and "patterns in language," he claims "this learning is designed to allow LLMs to generalize to new input data and to respond reasonably to new circumstances, not to reproduce exact patterns found in the training data."[11] Dr. Bilmes creates a straw man by inserting the phrase "exact patterns." I make no such claim in my description of the goals of LLM training or elsewhere in my initial report. The fact that Publishers have identified instances where Claude reproduces exact patterns and phrases—their lyrics—from its training data does not mean it is my opinion that Claude and other LLMs are ***only capable*** of regurgitating "exact patterns." It simply confirms that Claude can do so.

11. Further, Dr. Bilmes erroneously narrows the meaning of "generalization" within the context of LLMs. As I explain below, Dr. Bilmes misleadingly suggests generalization solely consists of generating "new content" that does not reproduce texts found in training corpora.[12] This

---

[8] Zhao Report ¶ 26.

[9] Emily M. Bender, et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big*, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (March 2021), 610–623.

[10] Tom Brown et. al, *Language Models are Few-Shot Learners*, arXiv:2005.14165v4 (July 22, 2020), https://arxiv.org/abs/2005.14165.

[11] Bilmes Rebuttal ¶¶ 15–16.

[12] *Id.* ¶ 16.

is overly restrictive and not in line with generally accepted definitions of "generalization" in the LLM field. While my description of LLM training may not qualify as "generalization" under Dr. Bilmes's restrictive definition, it does capture behaviors and outputs that may be characterized as "generalization" under more commonly accepted definitions in the machine learning community.

12. Dr. Bilmes's view that post-trained LLMs are designed to "not only generate new content based on language patterns but also to do so in a way that is useful, truthful, and appropriate"[13] also contradicts his prior statement that "learning is designed to allow LLMs to generalize to new input data and to respond reasonably to new circumstances, not to reproduce exact patterns found in the training data."[14] This is an imprecise description of generalization: generalization refers to an LLM's ability to respond to novel prompts and not the generation of "new content" that is not present in its training data.[15]

13. In layman's terms, generalization refers to a model's ability to generate accurate responses to prompts it may not have seen before. But there is no consensus on how or if generalization can be identified in model outputs. For example, in response to a prompt requesting a remixed version of a song in the style of another artist, the model may respond by combining the lyrics of one song with the style of a different songwriter, set to the tune of yet another song. Or in response to a prompt seeking HTML code for a website, the model may generate code with lines that appear in examples of code in its training data. These are not novel outputs in a literal sense. While the output might produce a combination that has never been seen before in that combination, the elements that the output is comprised of are not themselves novel and can more-or-less be traced back to examples in the training data. Put another way, this is a mosaic-like mixture and

[13] *Id.*

[14] *Id.* at ¶ 15.

[15] *See, e.g.*, Chandler Smith et al., *Evaluating Generalization Capabilities of LLM-Based Agents in Mixed-Motive Scenarios Using Concordia*, arXiv:2512.03318v1 (Dec. 3, 2025), https://arxiv.org/pdf/2512.03318v1 (defining generalization as "the capacity to perform well in previously unseen settings"); Chen Wang et al., *Do LLMs Have the Generalization Ability in Conducting Causal Inference?*, arXiv:2410.11385v1 (Oct. 15, 2024), https://arxiv.org/pdf/2410.11385v1 (characterizing generalization as the ability to handle previously unseen questions).

recombination of components from the training data. In my view, these outputs reflect the model's interpolation of memorized patterns from training data and is not indicative of generalization.

14. Another way of defining generalization from model output might involve a measure of creative output—generating something entirely novel and greater than the sum of its parts—where the building blocks of a given output are unidentifiable in the training data. Leading researchers generally agree that LLMs today do not generalize by this definition.

15. For example, Dr. Yoshua Bengio, one of the three Turing award winners for contributions to machine learning, and his co-authors sought to measure the creative capabilities of state-of-the-art LLMs (including Claude 3) and 100,000 human participants, using a validated creativity test, and found that LLMs remain below the mean creativity scores observed among the more creative segment of human participants.[16] In particular, even the best performing LLMs are still largely surpassed by the aggregated top half of human participants randomly recruited from the internet. Dr. Andrew Ng, founder of the AI education forum DeepLearning.AI, came to a similar conclusion, stating "(LLMs) are far less general than human abilities. After pretraining on the entire content of the public web, an LLM still struggles to adapt to write in certain styles that many editors would be able to, or use simple websites reliably."[17] He continued, "[a] typical human, despite having seen vastly less text or practiced far less in computer-use training environments than today's frontier models, nonetheless can generalize to a far wider range of tasks than a frontier model."[18] Dr. Ng further suggests that humans likely achieve such generalization using "many other mechanisms that we do not yet understand."[19]

16. In any case, the alleged "goals" of LLM training and Anthropic's goals for Claude specifically are neither relevant nor probative when observing how LLMs work on a fundamental

[16] Bellemare-Pepin, A., Lespinasse, F., Thölke, P. et al. *Divergent creativity in humans and large language models*, Nature Scientific Report, 16, 1279 (2026). https://doi.org/10.1038/s41598-025-25157-3 https://www.nature.com/articles/s41598-025-25157-3c.

[17] Andrew Ng, *Weekly Issues: Issue* 332, THE BATCH (Dec. 17, 2025), https://www.deeplearning.ai/the-batch/issue-332/.

[18] *Id.*

[19] *Id.*

level. What is more relevant and probative is what the processes and outcomes of LLM training actually are. While LLMs are capable of limited forms of generalization after training, the reality is that Claude and other LLMs ***also*** memorize because of their training and architecture. The fact that Anthropic desires the capability for Claude to generalize does not answer whether memorization behaviors ***actually*** occur, particularly with regard to specific, short, and distinctive texts like Publishers' lyrics. What Anthropic intends for Claude to generalize has no bearing on whether memorization behavior is actually common and endemic to Claude and its architecture.

17. Dr. Bilmes takes his nonstandard descriptions of LLM training further by claiming Claude models are not designed to have their training data "extracted by users" on the basis that "reliable systematic training data extraction . . . is primarily accomplished by adversarial prompting techniques and jailbreaks designed to elude the safeguards and defenses model builders, including Anthropic, put in place."[20] Such a conclusion incorrectly implies that these regurgitation events are rare or difficult to generate.

18. In reality, there is no magic or challenge to generating or creating these jailbreaks, because the available protection mechanisms for LLMs are so weak that average users everywhere are "discovering" these "jailbreaks" accidentally through normal usage. Dr. Bilmes's assertion that LLMs are reliable, because the only way to make them regurgitate is to tell them to regurgitate, is akin to saying that a guard dog is harmless, except when it is instructed to ignore its trainer and go bite its owner.

19. In fact, these "jailbreaks" are so common now, that they are routinely discovered and documented and categorized not by hackers, security researchers, or even tech industry workers, but by ostensibly non-expert, normal internet users on Reddit.[21] A slightly more professional blog entry written a year ago sorts jailbreaks into four large categories, each with five different

[20] Bilmes Rebuttal ¶ 17.

[21] A simple Google search for "llm jailbreak prompts reddit" yields several pages' worth of Reddit threads discussing various jailbreak techniques.

approaches to bypass LLMs.[22] Altogether, this single blog documents more than a hundred different ways to bypass LLM guardrails, and gives simple instructions on how to combine them. Not only that, but there are now tools created by researchers, including multiple Anthropic employees, that package up these hundreds of jailbreaks together into ready-to-use toolkits that anyone can deploy.[23] Such tools provide online source code and a running web service to replicate jailbreaking algorithms and are readily and publicly available.[24] This work (and the associated code provided by it) was in fact used just a few weeks ago (January 2026) to demonstrate that not only was it easy to jailbreak Claude 3.7 Sonnet, but that once jailbroken, "jailbroken Claude 3.7 Sonnet outputs entire books near-verbatim."[25]

20. In addition, Dr. Bilmes's argument is predicated on the successful operation of Anthropic's guardrails and mitigation systems—which are systems ***separate from and in addition to*** the Claude base model—and not the characteristics of Claude's fundamental architecture. These systems have no bearing on whether Claude or any other LLM memorize and regurgitate portions of their training data as a consequence of their fundamental architecture. They only reflect efforts to mitigate an existing process and behavior that results from the training process.

21. Finally, Dr. Bilmes's argument that Anthropic does not desire its Claude models to memorize is belied by the fact that Anthropic employees admit that Anthropic encourages at least some memorization during the training process.[26]

---

[22] Ian Webster, *Jailbreaking LLMs: A Comprehensive Guide (With Examples)*, PROMPTFOO (Jan. 7, 2025), https://www.promptfoo.dev/blog/how-to-jailbreak-llms.

[23] John Hughes et al., *Best-of-N Jailbreaking*, ARXIV:2412.03556 (Dec. 4, 2024), https://arxiv.org/abs/2412.03556.

[24] John Hughes et al., *Best-of-N Jailbreaking*, https://jplhughes.github.io/bon-jailbreaking/ (last visited Feb. 12, 2026).

[25] Ahmed Ahmed et al., *Extracting Books from Production Language Models,* Arxiv:2601.02671v1 (Jan. 6, 2026) https://arxiv.org/pdf/2601.02671.6, 2026.

[26] *See* Drain Dep. Tr. 68:4–19, 103:9–18, 201:4–9; Drain Dep. Ex. 143 at -223; Mann 30(b)(6) Dep. Tr. 178:16–179:13.

B. Memorization Is Inherent, Not Aberrational

22. While Dr. Bilmes quibbles with the scope and magnitude of memorization, he does not deny that LLMs memorize and regurgitate large portions of their training data. Indeed, as I previously noted in my rebuttal report, recent papers published in 2025 and 2026 found that LLMs, including Claude 3.7, can regurgitate verbatim or near-verbatim copies of entire books.[27]

23. In reality, memorization is an inherent structural property of LLMs like Claude due to their architectural design. The mere fact that Anthropic actively seeks to mitigate memorization by its deduplication and filtering actions does not change this reality. This is because LLMs train on data by extracting and storing observed patterns as conditional probabilities between different windows of contextual and output token sequences.

24. LLMs learn the likelihood of outputting any particular sequence of words, given any preceding sequence of words. ***The LLM cannot generalize on patterns it never observes***. For example, if an LLM only sees (in its training) that a particular sequence of words is *always* followed by "they stowed their wands beneath their cloaks and started walking briskly in the same direction," the conditional probability of the LLM outputting this phrase after being prompted by that particular preceding sequence of words is very close to or equal to 100%. As the length of the preceding context window gets longer and longer, the probability of the LLM returning the next sentence is again close to or equal to 100%. In fact, researchers discovered that one of the easiest and most reliable ways to get LLMs to reproduce training content is to simply give it a starting sentence or two of the desired content. This is a well-known weakness of autoregressive LLMs sometimes referred to as "prefix-forcing" or "response-prefilling" because forcing the LLM output to begin with a particular prefix can send it down a sequence of actions that produces the desired

[27] *See* Zhao Rebuttal ¶ 12. *See also* A. Feder Cooper, et al., *Extracting Memorized Pieces of (Copyrighted) Books from Open- Weight Language Models,* ARXIV:2505.12546v4 (Nov. 30, 2025), https://arxiv.org/pdf/2505.12546; Ahmed Ahmed et al., *Extracting Books from Production Language Models,* Arxiv:2601.02671v1 (Jan. 6, 2026) https://arxiv.org/pdf/2601.02671.6, 2026.

outcome.[28] This can be used to make the LLM "agree" to the user's request. It can also be used to prefill the earlier parts of the LLM's response, after which the LLM must continue to generate the content, because that follows from the conditional probabilities extracted from the training data.

25. The persistent quality of memorization behaviors in LLMs also supports the argument that memorization is fundamental to LLM architecture. For example, while many (including those at Google DeepMind) initially considered memorization an easy-to-fix bug (primarily through deduplication),[29] it is now widely accepted that memorization is actually a fundamental part of LLM models, and not a bug that can be addressed by deduplication techniques.[30] In fact, recent experimental work from Stanford University has demonstrated that the four leading LLM models (Claude 3.7 Sonnet, Gemini 2.5 Pro, Grok 3, and GPT-4.1) memorize so much content that entire copyrighted books can be fully extracted from them.[31] Another recent study from the Stanford AI Lab shows that memorization emerges in later-stage training and is interwoven with general language capabilities, making it difficult to remove without collateral damage to the model.[32]

26. While model developers can take steps to mitigate memorization, it is widely recognized that memorization has not yet been eliminated, and to date I am not aware of any model that avoids non-trivial amounts of memorization.

27. Dr. Bilmes asserts that my alleged acknowledgment of "the fact that frequent repetition of training data poses a greater risk of regurgitation" contradicts my statement that

---

[28] Yakai Li et al., *Prefill-level Jailbreak: A Black-Box Risk Analysis of Large Language Models*, arXiv:2504.21038 (Apr. 28, 2025), https://doi.org/10.48550/arXiv.2504.21038.

[29] Katherine Lee et al., *Deduplicating Training Data Makes Language Models Better*, ARXIV:2107.06499 (Mar. 24, 2022), https://arxiv.org/pdf/2107.06499; Nikhil Kandpal et al., *Deduplicating Training Data Mitigates Privacy Risks in Language Models*, ARXIV:2202.06539 (Feb. 12, 2022), https://doi.org/10.48550/arXiv.2202.06539.

[30] Igor Shilov et al., *The Mosaic Memory of Large Language Models*, NATURE COMM. (2026), https://doi.org/10.1038/s41467-026-68603-0.

[31] Ahmed Ahmed et al., *Extracting Books from Production Language Models,* Arxiv:2601.02671v1 (Jan. 6, 2026), https://arxiv.org/pdf/2601.02671.

[32] Jing Huang et al., *Demystifying Verbatim Memorization in Large Language Models*. STANFORD AI LAB BLOG (Apr. 28, 2025), https://ai.stanford.edu/blog/verbatim-memorization/.

"conditional probability leads to memorization of unique phrases or sequences of phrases."[33] This is incorrect; both statements are true, and the truth of one does not negate the truth of the other. Ubiquity in training data is the *primary* driver of memorization, but LLMs are more likely to memorize texts that are both ubiquitous to their training data *and* unique[34]—such as song lyrics. Generalization and memorization are not mutually exclusive concepts in machine learning. Stochastic behavior (*i.e.*, behavior involving a degree of randomness or uncertainty) does not eliminate regurgitation. Dr. Bilmes misses the point that Claude can both generalize *and* memorize.

C. Dr. Bilmes Overlooks Evidence That Claude Did Memorize and Regurgitate Publishers' Lyrics

28. Dr. Bilmes's insistence that Claude does not memorize and regurgitate copyrighted content to any meaningful degree[35] is plainly belied by the fact that Claude did just that for hundreds of Publishers' musical compositions.

29. Publishers have provided evidence of hundreds of unique instances of regurgitation of their lyrics by Claude models[36] and identified many more in the 5-million record sample cited in Mr. Joshua Dennis' expert report.[37] Dr. Bilmes does not meaningfully engage with this evidence, nor does he explain how or why Claude reproduced verbatim or near-verbatim copies of Publishers' lyrics if Claude did not memorize those same lyrics during the training process. In my view, Claude regurgitated Publishers' lyrics because it memorized the words to Publishers' lyrics and outputted them in response to user prompts.

---

[33] Bilmes Rebuttal ¶ 21.

[34] Zhao Rebuttal, ¶ 18 & n.20.

[35] Bilmes Rebuttal ¶ 28.

[36] *See, e.g.,* Plaintiffs_000000000–2677; Plaintiffs_0000004584–5062; Plaintiffs_0000005198–6967; Plaintiffs_0000009389–9525.

[37] Expert Report of Joshua W. Dennis (Dec. 19, 2025), ¶ 16. I understand that due to methodological changes, the Expert Reply Report of Joshua W. Dennis (Feb. 13, 2026) reflects revised figures relating to his analysis of the 5-million sample.

30. Dr. Bilmes suggests Claude ***may*** have learned about Publishers' lyrics by training on sources ***other*** than the actual lyrics.[38] Such a conclusion is speculation and unsupported by any empirical evidence. Moreover, I understand that Anthropic does not contest that Publishers' lyrics were copied into Claude's training data. Given that Publishers' lyrics are in Claude's training data, it is ***much more likely*** that Claude learned about Publishers' lyrics by training on the lyrics themselves, resulting in Claude's ability to regurgitate those same lyrics as output. Dr. Bilmes offers no concrete evidence that Claude was able to do this by learning from sources like musical critiques and reviews or Wikipedia pages. For the same reason, it is also highly likely that Claude also learned about syntax, poetic structure, etc. by training on Publishers' lyrics.

31. Again, it would be impossible or near-impossible for Claude to output copies of Publishers' lyrics unless copies of those same lyrics existed in Claude's training data. Without such training data, truly stochastic behavior would have an infinitesimally small chance of producing specific word combinations and sequences that precisely matched the Publishers' lyrics in question. That would be akin to waiting for an army of monkeys sitting at typewriters to produce *Hamlet* word for word: theoretically possible, but extremely unlikely.

32. In support of Dr. Bilmes's argument that an LLM can produce lyrics it has not seen in its training data, one might argue that stochasticity guided by generalized patterns may lead to such outputs (matching Publisher's lyrics) much faster. For example, an image model might learn how to produce a "younger version" of a person from numerous images of a person or family members at different ages and learn how to generate a "happy" facial expression from images of happy people. This model might be able to produce an image of a happy child, by combining these two dimensions, even if it had never trained on any images of happy children.

33. While such output might be possible under different assumptions and contexts, it would not apply to generation of lyrics by LLMs, specifically because most song lyrics are not easily predicted by generalizable patterns. The "feature space" that describes all possible song

[38] Bilmes Rebuttal ¶ 33.

lyrics is too vast and high dimensional and not governed by observable dimensions of features that a model can observe (from training data) and extrapolate from. For example, while the lyrics to "Fresh Prince of Bel-Air" exhibit some common qualities of hip-hop music from the 1990s, there are numerous qualities that render it unique, including the use of assonance in a single line ("chillin' out, maxin', relaxin' all cool"), simile ("I sprang with the quickness like lightning") and unusual repeat phrases ("Yo, Holmes"). These are creative choices that set this song apart from other songs and cannot be extrapolated using a generalized pattern based on any feature dimensions extracted from typical song lyric properties. In fact, this type of creativity is a hallmark of great songwriters. Unpredictability, novel phrasing, and subject matter are what makes songs memorable and contributes to their popularity.[39]

D. LLM Training Involves the Compression of Data

34. Dr. Bilmes distinguishes LLMs like Claude from compression algorithms and claims I "improperly equate[]" the two on the basis that LLM training and compression algorithms do not share the same "purpose, context, meaning, and use[s]."[40] I disagree.

35. LLM training can be considered a form of lossy compression because in mathematic terms, LLM training and compression seek to optimize the same loss function.[41] This does not mean that the only thing LLMs do is compression, or that all properties of compression are extant in LLM architecture or training. In any event, as much of the LLM research community and peers

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

[40] Bilmes Rebuttal ¶ 41.

[41] G. Deletang et al., *Language Modeling is Compression*, 12TH INT'L. CONF. ON LEARNING REP. (2024), https://openreview.net/pdf?id=jznbgiynus.

in my field argue, LLM training is aptly viewed as a form of lossy compression.[42] Indeed, even Dr. Bilmes agrees that compression algorithms and LLM training have similarities.[43]

36. Dr. Bilmes also argues that because the alleged "purpose" of LLMs is to generalize, LLM training is distinct from compression algorithms.[44] But the idea that a primary "purpose" separates LLM training from compression is not scientific, nor does a "purpose" framework have any meaning from a technical perspective.

37. Generalization is only possible because LLMs train on massive amounts of data in a process analogous to the lossy compression process. In other words, LLM training does not happen without compression. Even if the goal of LLM developers like Anthropic is for the LLM to "generalize," the process by which it does so is by lossy compression, which involves compressing the data sufficiently so the model can reproduce that data upon request—with some stochastically introduced variation, of course.

38. The process of "compressing" data by running it through an algorithm, such that it can be reconstructed (with varying degrees of fidelity, depending on how lossy the algorithm is) by running a decompression algorithm, versus the process of "compressing" data by converting it into conditional probabilities, such that that data can be reconstructed in response to prompts (with varying degrees of fidelity, depending on how the model develops the conditional probabilities and the model's stochasticity), equate to essentially the same processes.

39. Again, Dr. Bilmes's arguments are erroneously based on LLMs being evaluated by metrics different from that of compression. A 2024 paper authored by members of Google's DeepMind compares the optimization objective of a compressor to the optimization objective (the

---

[42] *See, e.g.,* Giorgio Franceschelli et al., *Training Foundation Models as Data Compression: On Information, Model Weights, and Copyright Law*, ARXIV:2407.13493v3 (Oct. 7, 2024), https://arxiv.org/pdf/2407.13493v3; 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.

[43] Bilmes Rebuttal ¶ 41.

[44] *Id.* ¶ 44.

loss function) used to train LLMs.[45] The authors concluded that "minimizing the log-loss is equivalent to minimizing the compression rate of that model used as a lossless compressor with arithmetic coding, i.e., current language model training protocols use a maximum-compression objective."[46] Indeed, even Dr. Bilmes recognizes that compression seeks to minimize the compression ratio (rate),[47] which supports the above comparison between LLM training and compression.

E. Post-Training Guardrails Do Not Sufficiently Address Memorization

40. Dr. Bilmes also argues that "[r]eliable systematic training data extraction," *i.e.*, regurgitation of memorized training data, is "primarily accomplished by adversarial prompting techniques and jailbreaks (i.e., atypical and/or abusive prompting)" designed to elude guardrails.[48] But this misses the point: guardrails do nothing to address the memorization itself inherent to the LLM architecture.

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

42. At bottom, guardrails are simply filters that look for keywords or concepts that indicate a prompt or an answer might include content that is undesirable. Given the complexity of the English language, any filtering attempt will be incomplete. There are an unlimited number of ways to refer to a concept or term or name. For example, a filter that blocks violent imagery, blood, or gore can be easily circumvented, intentionally or accidentally, by simple rewording (or

---

[45] G. Deletang et al., *Language Modeling is Compression*, 12TH INT'L. CONF. ON LEARNING REP. (2024), https://openreview.net/pdf?id=jznbgiynus.
[46] *Id*.
[47] Bilmes Rebuttal ¶ 48.
[48] *Id.* ¶ 17.

euphemistic substitution) of keywords like blood as "red liquid."[49] In general, LLMs will be unable to pick up on prompts for sensitive topics using a wide-range of verbiage, many of which will be difficult to anticipate because they come from very diverse sources in the training data, such as different ethnic communities using different terms to describe the same concept.[50] This highlights the fundamental mismatch between word-based guardrails and the in-context representations inside LLMs the guardrails seek to block. Recent research identifies this gap as a fundamental mismatch that explains why external guardrails will never be fully effective.[51] But even where guardrails are partially successful at blocking issues such as regurgitation, the problem remains: partial regurgitation is still regurgitation.

F. Anthropic's Hugging Face Data Teaches Claude that Humans Prefer Certain Lyrics-Related Responses

43. Dr. Bilmes claims I incorrectly characterize Anthropic's use and purpose of the human feedback dataset (the "HF Dataset") created in connection with their reinforcement learning with human feedback ("RLHF") training pipeline,[52] stating that I conflate "task-oriented and behavior-oriented tuning" of the HF Dataset as solely "task-oriented tuning."[53] He alleges the dataset does not reflect Anthropic's decision to design Claude to perform lyrics-related tasks,[54] and that the "helpful data" has "only ever been used for model behavior tuning, i.e., encouraging Claude to

---

[49] Ahmed B. Mustafa et al., *Anyone Can Jailbreak: Prompt-Based Attacks on LLMs and T2Is*, ARXIV:2507.21820 (Jul. 29, 2025), https://doi.org/10.48550/arXiv.2507.21820.

[50] Emilia Napolano, *Sora Inappropriate and Harmful Content Creation Easily Bypassed Through Simple Prompt Engineering*, ZENODO (Apr. 28, 2025), https://zenodo.org/records/15295087; Yihao Huang et al., *Perception-guided Jailbreak against Text-to-Image Models*, ARXIV:2408.10848 (Aug. 20, 2024), https://doi.org/10.48550/arXiv.2408.10848.

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

[52] Bilmes Rebuttal ¶ 10e.

[53] *Id.* ¶¶ 99–100.

[54] *Id.* ¶ 99.

become more helpful, honest and harmless, and not for any task-oriented tuning."[55] Dr. Bilmes is incorrect; he understates the use, scope, and impact of Anthropic's HF Dataset on Claude.

44. Accepting as true Dr. Bilmes's assertion that the "HF Dataset was designed to reflect human judgment of helpfulness or harmlessness" and not to "train any models to perform any specific tasks (such as lyric retrieval)," the lyrics-related prompt-output pairs in the HF Dataset nonetheless would have influenced the "aggregated human preference judgments" that were subsequently used to align models.[56] Even just marginally, use of these prompt-output pairs would teach Claude models trained using that dataset—[redacted][7]—to learn humans ***prefer*** certain types of responses in connection with lyrics-related requests as helpful—such as the reproduction of copyrighted lyrics.

45. Dr. Bilmes claims that Anthropic introduced a separate [redacted] subsequently and separately to the HF Dataset suggests Anthropic knew earlier behaviors introduced by the HF Dataset needed correction through counter-training.[59]

46. Dr. Bilmes also critiques my citation of seven non-exhaustive examples of lyric-related prompts-output pairs created and ultimately selected by Anthropic employees as proof of Anthropic's awareness of how users would use Claude and Claude's capabilities to respond.[60] Here too, Dr. Bilmes misinterprets my opinion, claiming my argument "fails to consider the dataset's broader composition and methodological purpose."[61]

---

[55] *Id.* ¶ 107.
[56] *Id.* ¶¶ 101–102.
[57] *Id.* ¶ 107.
[58] *Id.* ¶ 108.
[59] *Id.*
[60] *Id.* ¶¶ 100, 109.
[61] *Id.* ¶ 100.

47. As I have previously stated,[62] the lyric-related records found in the HF Dataset reveal to Anthropic that (1) users would prompt Claude for lyrics-related requests; and (2) Claude would likely respond to such requests, responses which may include regurgitated lyrics. In my view, these are crucial points of data that should have signaled to Anthropic how users would use Claude and how models subsequently trained on this dataset would likely respond. Further, in addition to the HF Dataset, I identified other additional internal records generated by Anthropic employees showing that Anthropic prompted and tested its early models with lyrics-related prompts and requests.[63] Dr. Bilmes fails to respond to this additional evidence.

48. Dr. Bilmes also critiques the selected seven instances of lyric-related prompts from the HF Dataset, arguing that responses to those prompts reflect varying levels of regurgitation of lyrics.[64] After his own analysis of the HF Dataset, Dr. Bilmes alleges that "only three entries [in the Hugging Face Dataset] contained [a] meaningful portion of the lyrics belonging to the requested songs, and of those, only two entries . . . included these lyrics in the response chosen by the crowdworkers."[65] This alone, he concludes, "rebuts [my] argument that this data set was used to train a model to regurgitate lyrics."[66]

49. Again, Dr. Bilmes misses the point: I do not cite the HF Dataset as evidence that Claude regurgitated lyrics in response to user prompts or that the data set was used to train a model to regurgitate lyrics. Rather, it is evidence of Anthropic's design choices for Claude and biases Anthropic introduced during the Claude training process that would lead to the regurgitation of Publishers' lyrics.

50. Moreover, his conclusion suffers from several flaws. First, the metrics of "[a] meaningful portion" or "meaningfully regurgitated" lyrics are arbitrary and undefined, if not

---

[62] Zhao Report ¶¶ 81–85.
[63] *Id.* ¶ 86.
[64] Bilmes Rebuttal ¶¶ 100, 111.
[65] *Id.* ¶ 114.
[66] *Id.*

wholly subjective.[67] What Don McLean considers a "meaningful portion" of his song lyrics may be different from what Rihanna considers to be a "meaningful portion" of her lyrics. In any event, they are not metrics used by machine learning experts and researchers.

51. In addition, Dr. Bilmes's emphasis on the apparent partial nature of the lyrics regurgitated during this finetuning[68] ignores the fact that partial regurgitation is still regurgitation, and such regurgitation demonstrates that Claude's training process included (1) prompting the model for lyrics and (2) the model regurgitating lyrics, regardless of the amount regurgitated.

52. Dr. Bilmes also notes that the regurgitated material in certain of the Hugging Face Dataset examples did not contain "actual copyrighted text."[69] For purposes of determining whether Anthropic's finetuning process included prompting the model for song lyrics and receiving output that contained song lyrics, the appearance of "actual copyrighted text" is irrelevant. The fact remains that during finetuning, Claude was prompted to complete lyrics-related tasks.[70] Moreover, knowledge of whether something is "actual copyrighted text" is not a concept within the scope of machine learning research and development.

53. In addition, Dr. Bilmes's exclusion of two prompts from his analysis on the basis that he believes them to be irrelevant to this litigation because they requested song names or artist names is puzzling.[71] Such prompts are plainly relevant because Claude's accurate response to such prompts requires Claude to ***have been trained on the related lyrics for the song or artist***. Such uses are at issue here, where Publishers allege that Anthropic's training on their lyrics without permission constitutes a violation of their copyrights.[72]

54. Dr. Bilmes also declines to engage with my opinion that Anthropic designed Claude to generate new song lyrics, on the basis that Anthropic's counsel asked him to "assume that lyric-

---

[67] *Id.* ¶¶ 109, 114.
[68] *Id.* ¶ 114.
[69] *Id.* ¶ 116.
[70] Zhao Report ¶¶ 79–80.
[71] Bilmes Rebuttal ¶ 110.
[72] *See, e.g.*, First Amended Complaint ("FAC") ¶¶ 16, 60–62, ECF No. 337.

related tasks such as creating original lyrics may be considered fair uses of copyrighted material in training data."[73] Such an omission renders his analysis of the HF Dataset, and the seven prompts I selected in my initial report, incomplete.

55. Moreover, Dr. Bilmes's analysis of the remaining prompt-output sets incorporates the assumption, based on instructions from Anthropic's counsel, that the lyrics contained in the output are "non-infringing, fair use."[74] As I explained in my rebuttal report, such characterizations drawing from legal concepts have no meaning in scientific machine learning research,[75] and they should be disregarded.

III. Dr. Bilmes Incorrectly Claims Hallucinations Are Not Inherent to LLMs

56. LLM hallucination is a well-documented phenomenon. Again, while Dr. Bilmes quibbles with the scope and magnitude of hallucination, he does not deny that it occurs. Ultimately, LLMs hallucinate because of fundamental characteristics in their architectural design: an LLM will hallucinate where there are gaps or patterns in its training data.[76] This will likely always be the case given the obvious difficulties in adding new datasets to LLM training corpora as new facts and knowledge constantly come to light.

---

[73] Bilmes Rebuttal ¶ 99.

[74] *Id.* ¶ 114.

[75] Zhao Rebuttal ¶¶ 61–62.

[76] *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-aremathematically-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; Ziwei Xu, *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, ARXIV:2401.11817 (Jan. 22, 2024), https://doi.org/10.48550/arXiv.2401.11817.

A. Scientific Literature Supports the View that Hallucinations Are Fundamental to LLM Architecture

57. It is clear, more than five years after the GPT-3 paper,[77] that LLMs are not consistently accomplishing the goals that AI companies envisioned in 2022. For example, despite early promises of hallucinations being quickly-fixed bugs,[78] today's researchers (including those from OpenAI) are consistent in their belief that hallucinations are in fact fundamental to the architecture.[79]

58. While hallucination behaviors might be less common in newer models than older models as LLM developers employ newer methods to mitigate, as far as I am aware, no LLM developer has been able to reduce hallucination to a trivial level. As Dr. Bilmes concedes, "[c]urrent LLMs do still produce hallucinations."[80]

59. Dr. Bilmes takes issue with my interpretation of the research paper *Why Language Models Hallucinate*, which he claims does not support my argument that hallucination is fundamental to LLMs and the transformer architecture.[81] Dr. Bilmes argues that the paper's "main takeaway" is that "LLMs will hallucinate ***less*** if the key LLM evaluations and benchmarks use grading schemes that penalize wrong answers and penalize guessing (on average) more than 'I

---

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

[78] Catherine Thorbecke, *AI Tools Make Things Up a Lot, and That's a Huge Problem*, CNN (Aug. 29, 2023), https://www.cnn.com/2023/08/29/tech/ai-chatbot-hallucinations.

[79] *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-aremathematically-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; Ziwei Xu, *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, ARXIV:2401.11817 (Jan. 22, 2024), https://doi.org/10.48550/arXiv.2401.11817.

[80] Bilmes Rebuttal ¶ 36.

[81] *Id.* ¶ 37.

don't know' answers."[82] But it is telling that Dr. Bilmes does not argue that hallucinations can be eliminated, only that LLM developers can take measures to reduce hallucinations.

60. My reading of *Why Language Models Hallucinate* is consistent with that of its writers, who explicitly conclude that "[h]allucinations are inevitable" for base models.[83] What's more, papers from other leading researchers demonstrate that hallucinations are not only inherent to LLMs, but also incredibly difficult, if not impossible, to eradicate. Notably, Xu et al. (2024) argue that "hallucination is inevitable for computable LLMs if the ground truth function is any computable function" and that "it is impossible to eliminate hallucination in the real world LLMs."[84] Huang et al. (2023) also describe hallucinations as a "compelling and ongoing concern" that are caused by "multifaceted origins, spanning the ***entire spectrum*** of LLMs' capability acquisition process," including in LLM data, training, and inference processes.[85] I concur with both papers' findings.

61. Measures to mitigate hallucinations and increase accuracy have not eliminated hallucinations from LLMs either. Retrieval Augmented Generation ("RAG"), which Dr. Bilmes claims helps with model output accuracy,[86] does not wholly eliminate hallucinations including because RAG introduces additional failure points.

62. RAG enables LLMs to retrieve and incorporate new information from external data sources (either a curated database or data store, or even the internet) before generating responses

---

[82] *Id.* (emphasis added).

[83] Adam Tauman Kalai et al., *Why Language Models Hallucinate*, ARXIV:2509.04664 (Sep. 4, 2025), https://arxiv.org/pdf/2509.04664.

[84] Ziwei Xu et al., *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, ARXIV:2401.11817 (Jan. 22, 2024), https://arxiv.org/pdf/2401.11817.

[85] Lei Huang et al., *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*, ARXIV:2311.05232 (Nov. 9, 2023), https://arxiv.org/pdf/2311.05232.

[86] Bilmes Rebuttal ¶ 36 n.152.

to a query.[87] These external data sources serve as supplements to the training data used to train the subject LLM. A RAG system includes an external knowledge base or data repository (or the internet), and a retriever that finds relevant data from this repository and passes it to the LLM as relevant context information, supplementing the user's prompt.[88] The LLM then uses this information to help generate a response to the current query.

63. Multiple studies have found that, even by using accurate and relevant data, RAG systems can still lead to statements that are either unsupported or contradict the retrieved information, a phenomenon called "RAG Hallucination." Hallucinations can occur during generation in a RAG-enabled model when external information retrieved by the RAG function conflicts with the model's training,[89] or because the LLM prioritizes the use of parametric knowledge (*i.e.*, "knowledge" that is the result of LLM training) over external retrieved knowledge.[90] A recent Stanford University study documented RAG-induced hallucinations in LLMs and LLM-based proprietary legal AI tools, finding that two major legal AI tools (Lexis+ AI and Westlaw AI- Assisted Research and Ask Practical Law AI) each hallucinate between 17% and 33% of the time.[91]

---

[87] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, PROC. OF THE 34TH INT'L CONF. ON NEURAL INF. PROC. SYS. (NEURIPS) 9459 (Dec. 6, 2020), https://doi.org/10.48550/arXiv.2005.11401.

[88] *See Introducing Contextual Retrieval*, ANTHROPIC (Sep. 19, 2024), https://www.anthropic.com/engineering/contextual-retrieval ("RAG is a method that retrieves relevant information from a knowledge base and appends it to the user's prompt, significantly enhancing the model's response").

[89] Wan Zhang & Jing Zhang, *Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review*, 13 MATHEMATICS 2025, 856 (Mar. 4, 2025), https://www.mdpi.com/2227-7390/13/5/856.

[90] Zhongxiang Sun et al., *ReDeEP: Detecting Hallucination in Retrieval Augmented Generation via Mechanistic Interpretability*, PROC. OF THE 13TH INT'L CONF. ON LEARNING REP. (ICLR) (Jan. 21, 2025), https://arxiv.org/pdf/2410.11414;

Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, 22 J. EMP. LEGAL STUD. 216 (Mar. 14, 2025), https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf

64. Moreover, during the retrieval of external knowledge, RAG can cause the model to retrieve low-quality or inaccurate information, as RAG also often relies on the public internet (via internet search) and sites like Reddit, where false, sarcastic or incorrect information can be extracted and fed into the LLM as context.

65. Nor does Constitutional AI provide the answer to hallucinations. Constitutional AI refers to models trained to self-detect harmful outputs, which can include hallucinations.[92] Dr. Bilmes notes that Constitutional AI is one of the techniques used by Anthropic to reduce hallucinations.[93]

66. However, the primary issue with the use of Constitutional AI to prevent hallucinations is that the knowledge a model consults to assess whether it is hallucinating (or not) is the same knowledge base (or lack thereof) that ***causes the model to hallucinate in the first place***. This is effectively asking the current LLM to review its own generated answer for potential hallucinations against a set of rules. Thus, the detection of hallucinations still relies on the knowledge base of the original LLM, and an LLM relying on Constitutional AI (such as Claude) can still hallucinate.

67. This is akin to asking a middle school math student to solve math problems for a concept they have not learned and then asking them to check their own work. If the system in question (here an LLM) cannot accurately produce answers to a question, it is unlikely to do so when asked to check its own answers. A student who makes a mistake on a particular type of problem will tend to make that same mistake repeatedly. At best, one can hope that there is a chance that the LLM is less likely to repeat the same mistake, but there is no evidence that hallucinations cannot be repeated by an LLM.

68. In addition to the results of the hallucination studies discussed in my rebuttal report,[94] another survey published in January 2026 used 60 questions designed to rigorously test the ability

---

[92] Yuntao Bai, et al., *Constitutional AI: Harmlessness from AI Feedback*, ARXIV:2212.08073v1 (Dec. 15, 2022), https://arxiv.org/pdf/2212.08073.

[93] Bilmes Rebuttal ¶ 39.

[94] Zhao Rebuttal ¶¶ 60, 131.

of 37 different LLMs, including 7 Anthropic models, to retrieve accurate factual information.[95] The researchers found that the hallucination rate for Anthropic models ranges from 17% to 27%.[96] Finally, another recent paper authored by OpenAI employees studied LLM hallucinations using learning theory demonstrated that LLMs continue to hallucinate after training due to both the statistical method used to train the model and the benchmarks used to evaluate the model. [97]

69. While some researchers have theorized that it might be possible to cure hallucinations, such cures would only be available in controlled settings. For example, Kalai et al. argue that, in theory, it might be possible to eradicate hallucinations in models created using a question-answer database to answer ***only*** a fixed set of questions.[98] But Claude models are not trained or used to answer only a fixed set of questions, and thus such theoretical hallucination mitigation would not be applicable to Claude. In fact, the utility of a general-purpose LLM like Claude would be greatly limited and no longer fit for "general purposes" if such a framework were applied to Claude's architecture.

70. Anthropic's claims of prioritizing building models that do not hallucinate does not mean it is successful in doing so. While Dr. Bilmes cites Anthropic's high-level efforts to reduce hallucination and improve honesty,[99] there is ample evidence that Claude continues to hallucinate with meaningful frequency. For example, as I pointed out in my rebuttal report, the HHEM study currently shows that Claude Haiku 4.5 hallucinates at a 9.8% rate, Claude Opus 4.5 (Anthropic's second most-recent model) hallucinates at a 10.9% rate, and Claude Opus 4.6 (Anthropic's most-recent model) hallucinates at a 12.2% rate.[100] Additionally, in January 2026, AIMultiple released

---

[95] Cem Dilmegani & Aleyna Daldal, *AI Hallucination: Compare Top LLMs like GPT-5.2*, AIMULTIPLE (Jan. 23, 2026), https://research.aimultiple.com/ai-hallucination/.

[96] *Id.*

[97] Adam Tauman Kalai et al., *Why Language Models Hallucinate,* ARXIV:2509.04664 (Sep. 4, 2025), https://arxiv.org/pdf/2509.04664.

[98] *Id*.

[99] Bilmes Rebuttal ¶ 39.

[100] *Hallucination Leaderboard*, https://github.com/vectara/hallucination-leaderboard (last visited Feb. 12, 2026).

a report that sought to measure the hallucination rate of 37 different LLMs, including seven Claude models. Based on the model's responses to questions that sought to measure their respective hallucination rates, AIMultiple found that the seven Claude models hallucinated at rates between 17% and 27%.[101] Even in this litigation, I understand that counsel for Anthropic submitted a court filing containing a hallucinated academic source.[102] Today, there is now a public database tracking publicized instances of LLMs hallucinating in legal case filings alone.[103] As of the date of this report, the database now includes 916 cases identified, spread across 30 international jurisdictions.

B. Hallucination-like Behavior Can Be Desirable, but there Are Trade-Offs

71. Dr. Bilmes also misinterprets my report when he claims that I argue hallucination is equivalent to a model's ability to perform creative tasks or produce novel outputs.[104] My position is that hallucination-like behavior is likely desirable in many contexts and that known methods for reducing hallucinations may cause models to become less flexible and less useful for creative purposes.

72. For example, one potential way to reduce hallucinations is to reduce a model's "temperature" parameter, which reduces the degree of randomness in the model's output, minimizes the stochasticity of the model, and makes the model more deterministic. However, reducing a model's temperature also reduces the randomness or semantic diversity of the model's output by making the model less likely to generate unusual or less-expected outputs.[105]

---

[101] *See supra* note 96.

[102] ECF No. 371 ¶¶ 3–6.

[103] Damien Charlotin, *AI Hallucination Cases Database*, https://www.damiencharlotin.com/hallucinations/ (last visited Feb. 12, 2026).

[104] Bilmes Rebuttal ¶ 38.

[105] *See* Bilmes Report ¶ 98 ("While models can be configured to behave deterministically, such as by using a zero-temperature setting that always selects the highest-probability next token, many LLM deployments implement some degree of stochasticity to improve diversity, creativity, and robustness in responses.").

73. Therefore, LLM developers may prefer increasing the model temperature[106] so that their models are more capable of assisting with more creative tasks. But hallucinations also often occur in less creative tasks, such as when a user requests factual information—for example, when a user asks who wrote a certain song but the model incorrectly credits the wrong songwriter in its response. In such cases, hallucinations may be considered "lying in a factual context," to quote one Anthropic employee.[107] The difficulty lies in seeking to reduce hallucination-like behaviors for outputting factual information but encouraging more random, diverse outputs for more creative tasks.

74. For example, a recent paper studied code generation by 16 currently-available LLMs and found that all 16 models exhibited a clear increase in hallucination rate as temperature value increases, where the "hallucination" was LLM-generated code containing references to coding packages that do not actually exist.[108]

IV. Dr. Bilmes Again Mistakenly Asserts Robots.Txt Files Grant Anthropic Permission to Crawl and Scrape Data

75. Dr. Bilmes disagrees with my discussion of robots.txt and Anthropic's crawling practices as unsupported by evidence, and undertakes his own empirical review to demonstrate that (1) Anthropic crawled and scraped publicly-available webpages in accordance with their robots.txt instructions and (2) my assessment that, at certain points in time, Anthropic [redacted], is incorrect.[109] His disagreement is unfounded.

[106] Antoine Bellemare-Pepin et al., *Divergent Creativity in Humans and Large Language Models*, 16 SCIENTIFIC REPORTS 1279 (Jan. 21, 2026), https://doi.org/10.1038/s41598-025-25157-3.

[107] Sara Sams, *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/.

[108] Joseph Spracklen et al. *We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs*, 34TH USENIX SEC. SYMP. (Aug. 2025), https://www.usenix.org/conference/usenixsecurity25/presentation/spracklen.

[109] Bilmes Rebuttal ¶¶ 55, 57–59.

76. I note that Dr. Bilmes not only limits his analysis of Anthropic's crawling behavior to lyric-specific crawled domains, but also concludes that "Anthropic's crawler overwhelmingly accessed lyric-containing webpages in accordance with the robots.txt rules in effect at the time of crawling."[110] Assuming Dr. Bilmes's analysis is sound, such a conclusion is not at odds with the sworn testimony and contemporaneous documentary evidence that [redacted][111] "Overwhelmingly" adhering to robots.txt instructions is not the same as "always" adhering to them.

77. But Dr. Bilmes's empirical analysis of Anthropic's crawling behavior is not sound. As explained in Dr. Shawn Shan's rebuttal report,[112] Dr. Bilmes's methodology relies on an arbitrary definition of "relevant" URLs (capturing webpages containing unrelated noise or common idioms, thereby artificially inflating the count of "relevant" URLs) and failed to include manual verification to avoid such artificial inflation of alleged permissive access.[113] Moreover, Dr. Bilmes's reliance on the Wayback Machine assumes a static web and ignores the potential for web permissions to have changed; a snapshot from the Wayback Machine "does not prove authorization existed at the actual moment of Anthropic's crawl," and there is no method to verify that the archived file matches what was actually served to Anthropic's user agent.[114] Dr. Shan points out numerous additional flaws with Dr. Bilmes's methodology that indicate his conclusion that 98.12% of crawled, "relevant" URLs actually crawled is a "statistical guess built on incomplete and potentially outdated historical data rather than actual compliance facts."[115]

---

[110] *Id.* ¶ 59.

[111] B. Mann 30(b)(6) Dep. Tr. 56:18–57:1; P. Sangani Dep. Ex. 128 (Anthropic_0000346083); P. Sangani Dep. Ex. 129 (Anthropic_0000343228).

[112] Rebuttal Expert Report of Dr. Shawn Shan (Jan. 23, 2026) (hereinafter "Shan Rebuttal").

[113] Shan Rebuttal ¶ 26.

[114] *Id.* ¶¶ 26–28.

[115] *Id.* ¶ 28.

78. Most importantly, as I explained in my own rebuttal, Dr. Bilmes's analysis of Anthropic's crawling behavior is fundamentally flawed because he infers permission to crawl was given even where he identified a lack of robots.txt files on webpages at the time Anthropic's crawler accessed them.[116]

79. Dr. Bilmes continues to erroneously conflate robots.txt instructions as "permitting" crawling.[117] As I have also explained, robots.txt files (or lack thereof) do not grant a permissive license.[118] Robots.txt merely request that crawlers do not crawl.[119] The instructions alone do not carry a technical means of enforcement; a crawler may choose to ignore them.[120] Just because a site may lack robots.txt does not mean the door is wide open for crawlers. The metaphorical door, assuming other access-control measures like paywalls are not in place, is always unlocked. The difference is whether there is a sign at the front asking visitors not to crawl and scrape.

80. Dr. Bilmes argues that Publishers' lyrics are on a wide assortment of websites. Dr. Bilmes fails to consider the likely possibility that much of the content on those websites it actually unlicensed themselves. Indeed, a cursory review of the top ten domains Dr. Bilmes studied reveals that at least two lack discernible copyright information indicating that they were licensed for display.[121]

81. Further, Dr. Bilmes's emphasis on the fact that for each of the 499 Works in Suit, "Anthropic permissively crawled at least one unique webpage that lacked CMI" to argue that Anthropic did not remove CMI from Publishers' lyrics, is neither probative nor compelling.[122]

---

[116] Zhao Rebuttal ¶ 73.
[117] Bilmes Rebuttal, Section IV.
[118] Zhao Rebuttal ¶¶ 74–77.
[119] *Id.* ¶ 74.
[120] *Id.* ¶¶ 74, 76.
[121] *See 7 Rings – Ariana Grande*, ST LYRICS, https://stlyrics.org/songs/a/arianagrande71622/7rings3172807.html (last visited Feb. 12, 2026); *Imagine Dragons – Radioactive (Remix) Lyrics and Video (Ft. Kendrick Lamar)*, LYRICS-VIDEO-MUSIC, https://lyrics-video-music.blogspot.com/2014/01/imagine-dragons-radioactive-remix.html (last visited Feb. 12, 2026).
[122] Bilmes Rebuttal ¶¶ 10c, 76.

Identifying one or even a few instances where CMI was absent does not prove that Anthropic did not remove CMI in other instances.

82. In addition, web crawlers are not all built the same nor treated by robots.txt instructions the same. Dr. Bilmes again makes the assertion that "web crawlers are standard tools used throughout the technology industry to collect publicly available webpages,"[123] suggesting Anthropic's crawlers are also "standard tools" and Anthropic's practices are similarly "standard." This is starkly misleading.

83. LLM crawlers, like Anthropic's "Crabby" bot, differ from traditional search crawlers. Traditional search crawlers do not extract the content but instead index pages for the purpose of ranking them as part of their search engine indexing process. LLM crawlers, on the other hand extract and download content for training LLMs.

84. Consequently, robots.txt files today are not well-equipped to deal with LLM crawlers. For example, website owners may edit their robots.txt instructions to allow limited crawling of certain paths (but not others) that might be useful for search-indexing purposes. That is, it may be desirable for website owners to allow limited forms of crawling to occur on their website for increased visibility in search results. The same is not necessarily true for LLM crawlers, which do not offer the same benefits to website operators as search crawlers.

85. On one hand, a robots.txt file may contain instructions on a crawler-by-crawler basis. However, website operators that set crawler-agnostic rules or are unaware of how to identify specific crawlers, such as Anthropic's "Crabby," would need to adapt an "all or nothing" approach; either block all crawling activities or allow all crawling activities.

86. Accordingly, reliance on the absence of robots.txt instructions as evidence that Anthropic was permitted to crawl, scrape, and download Publishers' lyrics from various websites is highly unreliable and misleading.

V. Dr. Bilmes Misrepresents Anthropic's Ability to Implement Measures to Avoid Copying and Regurgitation of Publishers' Lyrics

[123] *Id.* ¶ 54.

87. Dr. Bilmes dedicates multiple paragraphs to his discussion of why Anthropic cannot curate its training datasets to avoid the ingestion of copyrighted content and implies that I advocate for "a pre-training dataset comprised only of public domain and open-licensed data."[124] But I have not argued that Anthropic must do so. While that might be a good practice for many reasons, that is not within the scope of my reports or this litigation. Again, Dr. Bilmes attributes to me a position I have not expressed. Instead, based on my knowledge, experience, and review of the record, it is my opinion that it is *possible* for Anthropic to identify and remove Publishers' lyrics, including the 499 Works in Suit, from Claude's training data.

88. Dr. Bilmes again argues that because the Common Pile dataset—which was allegedly curated to prioritize the ingestion of public domain or licensed data—nonetheless contains copyrighted content, it is "complex and non-trivial" to "[c]ompletely remov[e] copyrighted works from large-scale Internet-derived training data."[125] In essence, Dr. Bilmes argues that, because third-party data curators scraped Publishers' lyrics, Anthropic should be absolved of its responsibility to further curate and filter the training datasets it feeds into Claude. This is nonsensical. Anthropic alone decides whether to use the Common Pile dataset and has every ability to further filter its contents before use in Claude models. Further, as I have previously explained, the Common Pile is merely a research dataset, not a legal benchmark, curated by researchers who do not appear to have been attempting to remove all instances of copyrighted materials from the datasets.[126]

89. With regard to the actual techniques Anthropic might use to remove copyrighted material, as a threshold matter, Dr. Bilmes misrepresents the simple baseline search method, and fails to understand the algorithms and properties of fuzzy fingerprint search methods. Dr. Bilmes argues that my descriptions of the simple grep method as a way to search for lyric fragments is inconsistent with the realities of large-scale systems and datasets, and that I misunderstand modern

[124] Bilmes Rebuttal ¶ 77.
[125] *Id.* ¶ 81.
[126] Zhao Rebuttal ¶ 88.

engineering constraints.[127] First, I offered the grep method as one example of, in my words, "the most simple, trivial and fast method."[128] "Large-scale systems and networking" is in fact my original area of research in my PhD work. I have built and tested systems on large computing clusters and worked with systems running on high-performance computing cluster systems running on fast speed networking backplanes such as Mellanox Infiniband networks, before they were acquired by NVidia. As a Computer Science professor, I have taught Operating Systems as a class on and off for over 14 years, where a large component of the class covers I/O performance in file systems. Finally, with my co-authors, I have published multiple award-winning papers at the top storage and architecture conferences (including our OceanStore paper which received the test-of-time award from ASPLOS,[129] the top systems architecture conference, and our Pond paper received best student paper award at USENIX FAST,[130] the top file storage conference attended and partly managed by professionals from networking and storage companies). I am therefore qualified to assess the engineering constraints on LLMs in performing these tasks.

90. I offered the grep method as "the most simple, trivial and fast method" to detect song lyrics in a large corpora of data.[131] But grep is just one example; there are numerous methods with which anyone could search for song lyrics in a large training dataset, including simple unix tools. Yes, moving data across a networked cluster does take time and incurs costs, but that is a one-time pre-compute cost. The point remains that it is absolutely feasible for Anthropic to carefully load and partition its [redacted] data across a large cluster and pre-load the data into local disk storage. How specifically Anthropic prefers to store its training data is up to its engineers, but data gathered from online sources (via scraping) or licensed partners (via shared files) tend to come in simple

[127] Bilmes Rebuttal ¶¶ 82–88.
[128] Zhao Report ¶ 153.
[129] John Kubiatowicz et al., *OceanStore: An Architecture for Global-scale Persistent Storage*, 35 ACM SIGPLAN NOTICES 190, https://dl.acm.org/doi/10.1145/356989.357007.
[130] Sean Rhea et al., *Pond: The OceanStore Prototype*, PROC. OF THE 2ND USENIX CONF. ON FILE AND STORAGE TECHNOLOGIES (FAST '03) (Mar. 2003), https://www.usenix.org/conference/fast-03/pond-oceanstore-prototype.
[131] Zhao Report ¶ 153.

formats, often large groups of flat files and archives. They do not come in object stores and distributed databases. It is not difficult to collect these flat files and perform a search for specific pieces of data (like Publishers' lyrics), prior to integrating the files into Anthropic's own proprietary or commercial data storage solutions for its training needs.

91. Latency across networked file systems, can be entirely eliminated using local I/O on flash disks, as is the norm on today's data center compute servers. Contrary to what Dr. Bilmes seems to imply, there are no performance degradation effects from "cache misses"[132] when reading a stream of data into memory for processing, because all of the data is new to the cache, and cache performance is irrelevant. Caches, whether they be on-chip L1, or in main memory, or even disk caches, are only beneficial when repeated access to the data is necessary. The point of caching is to keep frequently-accessed data closer to memory registers for faster access. When scanning across large volumes of data (as Anthropic would be), there is no repeat access, and caching provides zero benefit (other than perhaps prefetching data from disk into memory, which would only further accelerate data reads). Modern operating systems send parallel fetches to disk and caches in parallel, so that a cache miss would not add additional latency or delays to the data transfer operation.

92. Moving data physically closer to each server is doable as a one-time operation in order to eliminate all other performance issues and make parallel scans truly operate at the rate of local disk I/O. Today's data center servers generally come equipped with fast flash-based SSDs. Modern NVMe SSDs can achieve read speeds of up to 3,500 MB/s or higher.[133] Note that this type of scanning requires only linear sequential read operations, literally the fastest type of data access for modern file systems, and no writes are needed. Thus, a properly configured server with locally stored data can in fact approach the stated rates of 3500MB/s, which is equivalent to roughly 3.4 GB/s.

---

[132] Bilmes Rebuttal ¶ 84.

[133] *SSD vs HDD: Understanding SSD Speed and Performance Differences*, ADATA (Jul. 17, 2025), https://www.adata.com/en/quikTips/ssd-vs-hdd-speed-and-performance-difference/.

93. Dr. Bilmes is wrong that the primary bottleneck would be data movement through the memory, disk, and network hierarchy,[134] because NVMe SSDs can deliver data into memory roughly 2X faster than grep can process it, making grep itself the main bottleneck.

94. If per-server local disk storage is infeasible, an even easier data partitioning scheme is to simply partition the data across server racks[135] and store each partition on the local top-of-rack storage server. Typical data center rack configurations utilize per-rack storage servers that exceed 2 TB/s in aggregate throughout across the rack.[136] Even assuming a full rack of 40 servers (typically unrealistic due to cooling issues and needed space for management consoles or the storage server itself), per server throughput would easily exceed 50GB/s, which is easily 10X faster than grep. Any configuration that has fewer servers would allow the storage server to allocate more bandwidth per server, producing even faster performance per server. Again, this would easily remove storage and I/O as the performance bottleneck.

95. My proposal of the grep method, with zero additional configuration, demonstrates the extremely low bar to accessibility of string search tools. This is, in my view, the simplest method by which Anthropic engineers could review and remove data from training corpora. One imagines that Anthropic's engineers are also sophisticated computing professionals and can devise other, more complex mechanisms, such as fuzzy fingerprints, in order to reduce false positives.

96. Contrary to Dr. Bilmes's assertion, the methods I propose are in no way contradictory to Dr. Shan's methodologies in searching Anthropic's training data for Publishers' lyrics.[137] I understand Dr. Shan was given a very limited timeframe to search for lyrics on Anthropic's servers and had no ability to configure or reformat the data he reviewed. Dr. Shan's "triage" method was necessary, largely in part because of the slow hardware that was made available to him for the data

[134] Bilmes Rebuttal ¶ 83.
[135] A server rack is a rack specifically designed to hold and organize multiple servers in a way that that is best for the optimization of the equipment and utilization of floor space.
[136] *See, e.g.,* VDURA, www.vdura.com (last visited Feb. 13, 2026) (advertising 2 TB/s per-rack flash).
[137] Bilmes Rebuttal ¶ 85.

examination process.[138] With such a large volume of data, no flexibility in layout, and slow hardware, it is not surprising to me that Dr. Shan had to take additional steps to accelerate the processing time in order to complete his inspection within his limited time window of access to the data. If Anthropic were to implement my proposed methods, one assumes such restrictive conditions would not be present.

97. Dr. Bilmes also fails to understand the algorithm behind fuzzy fingerprints.[139] By nature, fuzzy fingerprints do not produce and search for hashes on every substring of a particular length. The general process involves computing a large number of hashes, each corresponding to a fixed length substring that is generated from moving a fixed length moving window across the document or file. This scan and hash computation happens once per file and requires only a single linear scan. Once all hashes are computed, a deterministic stochastic downsampling process is applied, such as sorting all hashes by numeric value and then selecting top K as the signature. The numeric downsampling process is incongruent with the one-way, randomized generation of the hashes, and the combination produces a small set of random but deterministically selected substring hashes from the document as its fuzzy signature. The downsampled substrings only represent a small portion of the file (ignoring most small or moderate differences between files), cannot be predicted a priori, and cannot be tampered with because of the irreversible nature of the hashes.[140] In terms of parameters, they are easily configured and most moderate values will be highly robust to most application scenarios. Choosing 10 hashes for a set size and a threshold of 5/10 as a match threshold is straightforward and provides very high detection rate and robustness to small errors for nearly all applications. Of course they can be tuned as needed in practice.

---

[138] Expert Report of Dr. Shawn Shan, PhD (Dec. 19, 2025), ¶¶ 33–34, 38; Expert Reply Report of Dr. Shawn Shan, PhD (Feb. 13, 2026), ¶¶ 34–35.

[139] Bilmes Rebuttal ¶¶ 89–90.

[140] Feng Zhou, et al., *Approximate Object Location and Spam Filtering on Peer-to-Peer Systems*, PROC. OF THE ACM/IFIP/USENIX INT.L CONF. ON DISTRIBUTED SYS. PLATFORMS, June 2003, https://link.springer.com/chapter/10.1007/3-540-44892-6_1.

98. Dr. Bilmes's use of the example of a single phrase in "Englishman in New York" by Sting to argue that a fuzzy matching system would be unable to reliably distinguish such language from non-lyric text, therefore risking the removal of substantial amounts of non-lyric material, misses the point.[141] It is highly unlikely that any single substring will be used in generating the fingerprint. But even if a single substring were used, the nature of fuzzy fingerprints is such that a match is only produced when more than half of the fingerprint set match. With the random downsampling, the likelihood of two nonduplicative documents producing more than 50% fuzzy fingerprint matches is astronomically low. Fuzzy fingerprints have been used for many years for various tasks, including finding similar documents using search engines, detecting plagiarism, finding nearly-identical spam emails, and more.[142] This technique is also effective at finding images with small differences generated by attacks targeting image classification models.[143]

99. In fact, Anthropic already employs [redacted] [4] As Dr. Bilmes points out in his initial report, researchers at Anthropic have experimented with removal of complex categories of content from Claude's training data, such as information about **c**hemical, **b**iological, **r**adiological and **n**uclear ("CBRN") weapons.[145] They found removing such material before training led to a meaningful reduction in capability with respect to the targeted category while preserving other capabilities. Dr. Bilmes further claims Anthropic uses [redacted]

[141] Bilmes Rebuttal ¶ 90.

[142] Udi Manber, *Finding Similar Files in a Large File System*, WTEC'94: PROC. OF THE USENIX WINTER 1994 TECH. CONF. (Jan. 17, 1994), https://www.cs.arizona.edu/sites/default/files/TR93-33.pdf.

[143] Huiying Li et al., *Blacklight: Scalable Defense for Neural Networks against Query-Based Black-Box Attacks*, 31ST USENIX SEC. SYMP. (Aug. 2022), https://www.usenix.org/system/files/sec22-li-huiying.pdf.

[144] Bilmes Report ¶¶ 114-15, 117–21.

[145] *Id.* ¶ 114.

[146] *Id.* ¶ 118–19.

100. Clearly, curation of Anthropic's training data at scale is possible, and can include filtering and removal of copyrighted content. Anthropic's refusal to do so on the basis that it is allegedly "infeasible" appears to reflect a prioritization of commercial expediency rather than actual impossibility. As I noted in my rebuttal report, the identification of harmful content—such as information on CBRN weapons—is more complex than matching copyrighted lyric strings and copyright symbols, but Dr. Bilmes claims the latter is technically infeasible, while the former is not.[147] This is wrong.

101. Dr. Bilmes also examines the concept of machine "unlearning," and argues that it is a similarly infeasible method of removing discrete pieces of content from an LLM's training data.[148] This analysis overstates my opinion.

102. Instead, it is my opinion that Anthropic can *reduce* the presence of Publishers' lyrics in an *already-trained* model through targeted unlearning methods. In fact, Anthropic researchers have already experimented with methods for targeted unlearning that preserve model capabilities apart from those being targeted.[149] Anthropic's researchers believe such an approach "provides a better trade-off between removing dangerous knowledge and preserving general capabilities compared to simply filtering out dangerous data during training, particularly when the labels distinguishing 'dangerous' from 'safe' content are imperfect.[150]

103. Machine unlearning does not offer the optimal solution for removing Publishers' lyrics from Claude models. Indeed, given the general difficulties and limitations of machine unlearning, it is likely more effective for Anthropic to remove Publishers' lyrics from Claude's training corpora *prior* to training. As I explained in my initial report, Anthropic is capable of implementing an assortment of preventative measures, including blacklisting domains during

---

[147] Zhao Rebuttal ¶ 94.

[148] Bilmes Rebuttal ¶¶ 93–97.

[149] *See* Igor Shilov et al., *Beyond Data Filtering: Knowledge Localization for Capability Removal in LLMs*, ANTHROPIC ALIGNMENT BLOG (Dec. 8, 2025), https://alignment.anthropic.com/2025/selective-gradient-masking/.

[150] *Id.*

crawling [151] and filtering out lyrics by searching for lyric strings and near duplicates. [152] Implementing such efforts to reduce the likelihood of copying and ingestion of Publishers' lyrics prior to pre-training is far more effective than addressing issues retroactively through unlearning methods.

VI. Dr. Bilmes's Guardrail Examinations Remain Narrow in Scope and Critically Flawed in Methodology

104. Dr. Bilmes dismisses my analysis of flaws in Anthropic's guardrail systems on the basis that a flaw in one guardrail mechanism is irrelevant or immaterial, since the overall system is "multilayered."[153] This is not an appropriate response for several reasons.

105. First, while Anthropic's guardrails have multiple parts, [redacted]

106. Second, I have not seen any evidence, and Dr. Bilmes does not cite any evidence, that any of the different components of Anthropic's guardrails are designed to address flaws in the other components, and there is no evidence that any one part of the system does specifically address flaws in other parts of the system. Analysis of each part of the system in isolation can still identify weaknesses in the overall system.

107. Dr. Bilmes's tendency to fall back on this "multilayered" point results in his unjustifiable dismissal of flaws in Anthropic's guardrails. For example, in my opening report, I pointed out significant flaws in Anthropic's Prompt Shield system: that Prompt Shield fails to prevent Claude from outputting copyrighted content where the prompts themselves do not ask for copyrighted content, and that an input-side filter like Prompt Shield can be easily circumvented.[154] In his rebuttal, Dr. Bilmes does not dispute that these are serious potential flaws that undermine

[151] Zhao Report ¶ 144.
[152] Zhao Report ¶ 153; Zhao Rebuttal ¶ 93.
[153] Bilmes Rebuttal Section VI.
[154] Zhao Report ¶¶ 121–22.

the effectiveness of Prompt Shield.[155] Instead, he tries to deflect them by claiming that my observations improperly "rests on an assumption that input-side mechanisms must independently and conclusively prevent copyrighted outputs."[156] I have not made that assumption, nor does my analysis of Prompt Shield's flaws require such an assumption. Rather, my analysis simply indicates that input-side mechanisms do not independently and conclusively prevent copyright infringement, as they are frequently failure-prone and require making a trade-off between false positives and false negatives.[157]

108. Dr. Bilmes's assertion that CCS is "designed to detect reproduction patterns across [redacted]" and "normalize[s] text to account for differences such as capitalization and punctuation"[158] does not lessen my concern that CCS is inherently limited and fragile as a hash-based filter system. Dr. Bilmes does not appear to argue that CCS' normalization process accounts for alternative spellings, transcription errors, or abbreviations of words, which are common in musical lyrics; instead, his assertion only addresses CCS' ability to account for differences in capitalization and punctuation.

109. Even at its best, CCS is a simple hash lookup scheme[159] that will fail badly when there are errors in the text, as there is in training data of this type (*i.e.,* lyrics). Errors can and do arise from data manipulation and processing, including OCR (optical character recognition) on scanned images, text extraction from PDFs, and errors in text from transcriptions of video dialogue, all of which are present in Anthropic's training data. These errors are in fact the key

[155] Bilmes Rebuttal ¶¶ 131–136.
[156] *Id.* ¶ 136.
[157] Zhao Report ¶¶ 122, 124.
[158] Bilmes Rebuttal ¶ 138.
[159] *See Hash Table Data Structure*, GEEKS FOR GEEKS, https://en.wikipedia.org/wiki/Hash_table (last visited Feb. 13, 2026).

reason why fuzzy fingerprints have been proposed, studied and deployed for text deduplication and similar applications since the late 1990s.[160]

110. Dr. Bilmes also inappropriately dismisses my critique of CCS by asserting that my "criticisms regarding the scope and maintenance of the CCS corpus address issues of corpus administration rather than the technical capability of CCS to detect and block reproduced content once reference material is incorporated."[161] This ignores my discussion of CCS' technical ability to detect block reproduced content once reference material is incorporated.[162] But more fundamentally, CCS' effectiveness depends on *both* the scope and maintenance of its reference corpus *and* its technical ability to block material incorporated into that corpus. Dismissing the former as separate from the latter ignores the former's critical importance to the functioning of the system.

111. Dr. Bilmes has no real response to this problem. Indeed, he agrees that CCS is "effectively useless" for any lyrics not included in the reference corpus.[163] His references to "isolated coverage gaps" and "hypothetical omissions"[164] are misleading and severely downplay the scope of this problem: as he notes elsewhere, UMPG and Concord alone have catalogs containing millions of songs,[165] yet CCS contains lyrics for fewer than [redacted]. Dr. Bilmes then points to the testing in his initial report of 50,000 input-output pairs, supposedly to demonstrate that these coverage gaps are "isolated," but this is not responsive to the issue: that testing

[160] Udi Manber, *Finding Similar Files in a Large File System*, WTEC'94: PROC. OF THE USENIX WINTER 1994 TECH. CONF. (Jan. 17, 1994), https://dl.acm.org/doi/10.5555/1267074.1267076.
[161] Bilmes Rebuttal ¶ 141.
[162] Zhao Report ¶ 129.
[163] Bilmes Rebuttal ¶ 142.
[164] *Id.* ¶ 142.
[165] *Id.* ¶ 185.

specifically concerned only the 499 Works in Suit, all of which Anthropic [redacted][166]

112. Contrary to Dr. Bilmes's claim,[167] the record demonstrates that Anthropic's defenses are indeed static and ephemeral. I am not aware of any evidence, and Dr. Bilmes cites none, that Anthropic has [redacted], despite the obvious fact that many extremely popular songs have been released since then. While Dr. Bilmes cites evidence that he claims shows Anthropic's guardrails have evolved since their first introduction in [redacted] most of that "evolution" seems to have occurred [redacted].[168] I am not aware of any evidence, and Dr. Bilmes cites none, that Anthropic's guardrails have significantly evolved since that initial burst of activity. While Dr. Bilmes cites thousands of commits in the code for CCS and Prompt Shield between October and November 2022 and October 2025 as evidence of "continuous refinement and improvement,"[169] that evidence is of minimal value. Dr. Bilmes does not show the temporal distribution of those commits, nor how many relate to the initial development of those systems versus their ongoing maintenance through the years. He also does not show how many of the commits for CCS relate to lyrics, nor does he show how many of the commits for Prompt Shield relate to copyright infringement. Nor does he show whether those commits related to strengthening, rather than loosening, any of Anthropic's guardrails.

113. As I and Dr. Shan have previously explained, Dr. Bilmes's evaluation of copyright guardrails for Claude 4.1 Opus, Claude 4 Opus, Claude 3.5 Haiku, and Claude 3 Haiku is deeply

---

[166] Anthropic_0000347810; V. Rao 30(b)(6) Dep. Tr. 96:25–97:3; Anthropic's Amended and Supplemental Responses and Objections to Plaintiffs' First Set of Interrogatories (Nos. 9, 11, 12), dated Nov. 12, 2025.

[167] Bilmes Rebuttal ¶ 143.

[168] *Id*. *See also* Anthropic_0000347810; V. Rao 30(b)(6) Dep. Tr. 96:25–97:3; Anthropic's Amended and Supplemental Responses and Objections to Plaintiffs' First Set of Interrogatories (Nos. 9, 11, 12), dated Nov. 12, 2025.

[169] Bilmes Rebuttal ¶ 145.

flawed and only minimally probative of the efficacy of Claude's guardrails, generally, for several reasons.

114. First and foremost, Dr. Bilmes's evaluation completely omits any examination or testing of guardrails for Claude 1 and 2 models and focuses solely on Claude models developed after Publishers filed their litigation and after Anthropic developed bespoke guardrails to prevent the output of Publishers' 499 Works in Suit, specifically.[170] This introduces a significant amount of selection bias into Dr. Bilmes's analysis and compromises the integrity and applicability of Dr. Bilmes's findings.

115. Second, Dr. Bilmes did not attempt any multi-shot examinations of Claude's guardrails during his evaluation.[171] This is significant because, as relevant literature has shown, models often forego safety instructions over data that has been accumulated in-context, allowing users to bypass guardrails simply by asking the same prompt to a model over and over again.[172] Indeed, Dr. Shan's own experiment using the Claude 3 Haiku API to prompt lyrics to one of Publishers' 499 Works in Suit through simple multi-shot adversarial extraction successfully extracted lyrics from Claude three out of 10 times.[173]

116. Further, Dr. Bilmes did not conduct any testing of filters or guardrails on third-party applications powered by Claude, utilizing the Claude API, or on Claude as available via platforms such as Amazon Bedrock, Google Vertex AI, or Quora's Poe. Dr. Bilmes acknowledges that Anthropic has no control over such applications even though they make Claude readily accessible to millions of users.[174]

---

[170] Zhao Rebuttal ¶ 122; Shan Rebuttal ¶ 34.

[171] Bilmes Rebuttal ¶¶ 186, 188 (citing Bilmes Report Section IX.C).

[172] Cem Anil et al., *Many-Shot Jailbreaking*, ANTHROPIC (April 2, 2024), https://www.cdn.anthropic.com/af5633c94ed2beb282f6a53c595eb437e8e7b630/Many_Shot_Jailbreaking_2024_04_02_0936.pdf.

[173] Shan Rebuttal ¶ 36.

[174] Jeffrey Diver, *New Poe Apps for Web now, ioS & Android Later*, QUORA (Mar. 2025), https://poeai.quora.com/New-Poe-Apps-for-Web-now-ioS-Android-Later ("We now provide millions of users with access to 100 of the best text, image, video, and audio models from leading AI companies, along with millions of bots built on top of them, all in one place.").

117. Finally, Dr. Bilmes concedes leakage may occur in "limited and atypical settings."[175] But real-world prompts can also constitute "limited and atypical" uses. The claim that Claude guardrails work well "under ordinary usage" but fail under common jailbreak conditions only proves Claude's vulnerabilities in circumstances Anthropic's developers did not consider, and highlights why post-training guardrails will always be insufficient.

118. Additionally, Claude's guardrail efficacy should be assessed on capability and vulnerability, not just Claude usage rates within a limited timeframe. The low frequency of certain prompts/uses today does not eliminate the structural risks over Claude's entire product life cycle. That is, relatively infrequent regurgitation can be significant over the course of many months or years. For example, even accepting as true the [redacted] figure that Dr. Bilmes alleges constitutes requests for music lyrics across all Claude user prompts,[176] such a seemingly small number can constitute thousands or tens of thousands of instances of lyric regurgitation across a total pool of millions of Claude prompts. These consequences will only grow in proportion as Anthropic continues to accelerate LLM deployment across applications and user populations. Just in the last two years, Anthropic's business customers deploying their models has grown by a factor of 3000.[177] What might be thousands or tens of thousands of instances today can quickly turn into tens of millions of instances within a year or two.

119. Finally, Dr. Bilmes erroneously characterizes my opinions as requiring Anthropic's guardrails to meet a "foolproof" standard.[178] I do not advocate for such a standard, but rather observe that Anthropic's post-training mitigation efforts are imperfect and will never fully prevent the output of copyrighted content. [179] But even if I did advocate for a "foolproof" standard, Dr.

---

[175] Bilmes Rebuttal ¶ 227.
[176] Bilmes Report ¶ 299.
[177] Louis Columbus, *How Anthropic's Safety Obsession Became Enterprise AI's Killer Feature*, VENTUREBEAT (Dec. 16, 2025), https://venturebeat.com/security/how-anthropics-safety-obsession-became-enterprise-ais-killer-feature.
[178] Bilmes Rebuttal ¶¶ 121, 158.
[179] Zhao Report ¶¶ 133, 138.

Bilmes does not offer a compelling counter-standard; he does not articulate any clear metric against which the guardrails' efficacy should be measured, nor explain at what point the guardrails could be considered sufficiently effective.

C. Dr. Bilmes's Contention that Publishers' Lyrics Are Immaterial for Regurgitation or Performance Misrepresents My Opinions.

120. Dr. Bilmes argues that I "overstate the value and importance of the Plaintiffs' lyrics in improving Claude's performance given the size of Claude's models and training data."[180] This is a distortion of my opinion regarding the necessity of Publishers' lyrics and conflates the issue of regurgitation of specific lyrics with overall model performance. As I have explained, Publishers' lyrics are necessary for Claude to produce partial or full copies ***of those lyrics***.[181]

121. I opine both that the reproduction of Publishers' lyrics would be extremely unlikely if the model had not been exposed to copies of those lyrics during training,[182] ***and that*** Anthropic underestimates the impact of Publishers' lyrics overall Claude performance and capabilities.[183] Dr. Bilmes collapses the former into the latter and directs critiques of the latter to apply to the former.

122. Dr. Bilmes claims that because I agree with Anthropic's statement that Claude must be shown examples of "texts that approximate [relevant] concepts" before it is able to "respond sensibly to [them]," I also agree that Claude does not need an exact copy of Publishers' lyrics in order to understand song lyrics as concept.[184] This is untrue. My agreement with Anthropic's statement describing LLM training at a high level does not equate to agreement that Claude does not need Publishers' lyrics to understand song lyrics.

123. While Claude may learn about song lyrics ***broadly*** from sources other than Publishers' lyrics, such as "reviews and discussions of music, songs, and poetry,"[185] that does not

[180] Bilmes Rebuttal ¶ 232.
[181] Zhao Report ¶ 70; Zhao Rebuttal ¶¶ 108–111.
[182] Zhao Report ¶ 70.
[183] Zhao Rebuttal ¶ 102.
[184] Bilmes Rebuttal ¶ 232.
[185] *Id.* ¶ 233.

necessarily enable Claude to then reproduce or answer questions about ***Publishers' lyrics specifically***. Nor does Dr. Bilmes offer empirical evidence this is the case.

124. As I have discussed, in order for Claude to ***regurgitate*** Publishers' lyrics, Claude very likely needed to train ***on those same lyrics***.[186] Dr. Bilmes offers no rebuttal to this and only suggests "it is possible for models such as Claude to produce complete text even if its training data only included near matches or subparts of the same text (i.e., not a full n-gram match)."[187] He further suggests "bits and pieces of lyrics can (and do) exist within other non-lyric texts across the Internet."[188] But this does not adequately address the ultimate fact that Claude was indeed trained on copies of Publishers' lyrics. And even if the lyrics were obtained piecemeal, as Dr. Bilmes suggests, Claude still would have trained (and did train) on Publishers' lyrics.

125. Ignoring this point undermines Dr. Bilmes's argument that Anthropic did not need Publishers' lyrics specifically to learn about songs and music lyrics in a broader context. But this is altogether a separate question from whether Claude ***actually trained*** on Publishers' lyrics and whether ***that training has any effect*** on Claude's performance and ability to regurgitate lyrics, create derivative lyrics, or answer general queries about those lyrics.

126. With regard to the latter and more probative question, Dr. Bilmes suggests Claude may have or could have learned about song lyrics from various sources and reiterates that Publishers' lyrics likely make up a small fraction of the total training data for Claude models.[189] However, none of this constitutes reliable or empirical evidence measuring the actual impact of Publishers' lyrics on Claude and its outputs. Ultimately, including for the reasons stated in my rebuttal report, [190] Dr. Bilmes is incorrect in concluding Publishers' lyrics are wholly immaterial to Claude's performance.

---

[186] Zhao Report ¶ 70; Zhao Rebuttal ¶¶ 110–112.
[187] Bilmes Rebuttal ¶ 235.
[188] *Id.* ¶ 235.
[195] Bilmes Rebuttal ¶¶ 232–38.
[190] *See* Zhao Rebuttal ¶¶ 102–15.

VII. Dr. Altman Understates the Harms Posed by LLMs and the Challenges in Addressing Them

127. In his rebuttal report, Dr. Altman agrees that "LLMs introduce meaningful challenges and their risks warrant serious attention"[191] but suggests these challenges may be managed through "technical innovation, market adaptation, and reasonable governance."[192] He further claims to find no basis for my characterization of LLMs like Claude carrying intrinsic risks and threatening inherent harms.[193] He concludes that "the transformative applications for and benefits for this technology will continue to expand, particularly in science, medicine, and other public-serving domains."

128. In the first two respects, I disagree for the reasons described below. On the third point, the relevance of Dr. Altman's opinions continue to escape me. Dr. Altman seems to again avoid any engagement with the issues in this litigation or the uses challenged (or defended) by the Parties. For example, Dr. Altman offers no explanation for why any of his cited applications, benefits, or capabilities of Claude specifically relate to Anthropic's use of Publishers' lyrics or how Claude's functionality is enhanced or made possible through the inclusion of Publishers' lyrics in the training corpora or in Claude outputs.

129. Nevertheless, I provide the following responses in reply to Dr. Altman's rebuttal opinions below.

A. AI Safety Problems Are Not Adequately Addressed by Simple Changes in Policy or Governance

130. Dr. Altman frames the potential misuses, harms, and failure modes as governance and policy issues while ignoring various model-level issues and harmful behaviors that are not curable by mere governance fixes. In short, Dr. Altman incorrectly asserts the concerns I raise in my report "reflect governance questions rather than intrinsic technological defects."[194]

---

[191] Altman Rebuttal ¶ 21.
[192] *Id.* ¶ 13.
[193] *Id.* ¶ 14.
[194] Altman Rebuttal ¶ 30.

131. I do not raise mere "technological defects" but rather intrinsic risks that stem from AI model behaviors and capabilities. As I explained in my initial report, some of these harms stem from emergent model behaviors, such as the ability and tendency of Anthropic's own Claude models to resort to "malicious insider behaviors . . . including blackmailing officials and leaking sensitive information to competitors."[195] This reveals a model reliability issue, not a failure of policy. In fact, such dangerous behaviors seem to emerge in Claude, despite the implementation of various guardrails, mitigation, and governance efforts—including the use of Anthropic's Constitutional AI framework—by Anthropic.[196]

132. Dr. Altman argues that the presence of "safety guardrails in LLMs, including Anthropic's Claude models, and the fact that they are continuously refined and improved, undercuts [my] characterization of LLM systems as 'uncontrolled.'"[197] While the implementation of such guardrails is generally a good thing, voluntary controls do not constitute actual controls that Anthropic is obligated to maintain, particularly if Anthropic believes they are no longer necessary. For example, in one Anthropic document from [redacted], Anthropic co-founder Tom Brown questioned whether [redacted]"[198] The same co-founder also appears to have advocated for [redacted] [199] In short, Anthropic's efforts to mitigate harms, "including model evaluation, red-teaming, usage restrictions, and ongoing safety research"[200] do not constitute strict

[195] Zhao Report ¶ 165; *Agentic Misalignment: How LLMs could be insider threats,* ANTHROPIC (June 20, 2025),https://www.anthropic.com/research/agentic-misalignment.
[196] Bilmes Rebuttal ¶ 54.
[197] Altman Rebuttal ¶ 29.
[198] T. Brown Dep. Ex. 279 (Anthropic_0000217026) at -217028.
[199] T. Brown Dep. Ex. 280 (Anthropic_0000599647).
[200] Altman Rebuttal ¶ 52.

governance safeguards since they are non-binding and subject to change at any time, depending on Anthropic's whims.

133. Dr. Altman also applauds Anthropic for its "explicit acknowledgement of [] risks, and its extensive, proactive efforts to anticipate and prevent them." But the mere fact that Anthropic identified and disclosed many significant harms that Claude can inflict on society does not mean these harms will not come to bear or that Anthropic should not be required to do more to prevent them.

B. Historical Analogs Are Not Probative of the Unprecedented Risks of LLMs

134. Dr. Altman's suggestion that governance remedies can serve as a kind of panacea to all the harms I identified is entirely unsupported by relevant empirical evidence.

135. Dr. Altman analogizes the adoption of AI tools to other historical technological shifts, including the rollout of electrical power in the 19th century, the discovery of x-rays and radioactive materials, development of GPS technologies, and the adoption of automobiles.[201] Dr. Altman claims "initial disruptions associated with new technologies are addressed over time with technical innovation, market adaptation, good governance, and in some cases, regulatory intervention," suggesting the adoption of AI tools will go a similar way.[202]

136. Aside from the obvious logical fallacy that past performance is not indicative of future results, the introduction of LLMs and similar AI tools is much unlike any other technological shift humanity has previously encountered. As I have previously discussed,[203] LLMs are disruptive to an extent never seen before, affecting all aspects of our lives, and as one Anthropic co-founder notes, capable of job displacement and potentially resulting in the extinction of mankind.[204] These

---

[201] *Id.* ¶¶ 25–28.

[202] *Id.* ¶ 24.

[203] Zhao Report Section VII.

[204] T. Brown Dep. Tr. 41:3–42:6 (Q: "And my question focused on the risks, so what are those catastrophic risks to humanity?" A: "I think that I have publicly posted it is—again, in addition to the tremendous upside potential from AI, it seems like AI has the potential to cause issues as bad as the extinction of humanity." Q: "So one of the risks of AI, including Claude, is something as

impacts are particularly felt in the market for creative works. For example, as I previously explained, the capacity for AI to generate "slop" in amounts never seen before has profound effects on artists and creators, flooding markets for books and fictional writing.[205] Indeed, Anthropic co-founder Tom Brown readily admits that "Claude can read much more quickly than a human, has been—has seen much more overall data than a human, and can write more quickly than a human" and that in "many places [] Claude and other language models are superhuman."[206] Dr. Altman declines to engage or address these already-present harms to labor markets, instead vaguely concluding that "growing pains [] can and should be addressed through prudent governance."[207]

137. Dr. Altman's views seem to be at odds with Anthropic's own founders' views.[208] For example, Anthropic's CEO, Dario Amodei, who has repeatedly asserted that AI will eliminate 50% of entry-level white collar jobs within five years,[209] argues that AI is unlike other advances in

---

bad as the extinction of humanity, is that correct?" A: "Again, in—in addition to the—almost the unprecedented potential for upside, I agree that there's potential downsides as bad as extinction." Q: "What are the other catastrophic risks of AI?" A: "Well, yeah, I guess—I think again, as this is a technology that is extremely omni-purpose. It has the potential for immense upsides; many different possible downsides, which I think we have been trying to talk about publicly for a long time and try to mitigate, including job displacement, potential misuse, and potential misalignment from the AI."); *see also id.* at 347:24–348:11 (Q: "What are the other risks of the AI technology that your company is developing?" A: "Yeah, I think that I expect AI to have, like, tremendous potential for benefits similar to something like the Industrial Revolution. I think also it has the potential for great harms with not only our ones that—like the—kind of the job replacement that we saw in the Industrial Revolution, but also kind of, I think, potential issues for misuse and for, like, misaligned AI causing trouble itself."), *id.* at 348:12–350:3.

[205] Zhao Report ¶¶ 159–60.

[206] T. Brown Dep. Tr. 235:12–15; 241:21–23.

[207] Altman Rebuttal ¶ 13.

[208] *See, e.g.,* T. Brown Dep. Tr. 55:3–17 (Q: "Do you agree, Mr. Brown, that the risks of AI technology are significantly higher than any risks that took place during the Industrial Revolution?" A: "I think—I think they're—they're very different risks." Q: "Was there a risk of human extinction during the Industrial Revolution?" A: "Yeah, I think—I think no. So I think that's an example of a place right now that has increased risk. And that's something that we care a lot about, though. That's part of the—the founding mission of Anthropic, to reduce that risk.").

[209] *See, e.g.,* Dario Amodei, *The Adolescence of Technology* (Jan. 2026), https://www.darioamodei.com/essay/the-adolescence-of-technology; Sawdah Bhaimiya, *Anthropic CEO Dario Amodei Wans AI May Cause 'Unusually Painful' Disruption to Jobs*, CNBC (Jan 17, 2026), https://www.cnbc.com/2026/01/27/dario-amodei-warns-ai-cause-

technology—contrary to Dr. Altman's views—including because the pace of progress of AI is "much faster than for previous technological revolutions" and "AI appears to be advancing from the bottom of the ability ladder to the top . . . . affecting people with certain intrinsic cognitive properties, namely lower intellectual ability."[210]

138. In addition to highlighting the "sheer number of risks" and harms arising out of AI, Dario Amodei further argues that common-sense governance proposals involving "chip export controls and judicious regulation of AI" have largely been rejected by policymakers in the US because of the "literally trillions of dollars per year" to be made with AI.[211] In Dario Amodei's words, "[t]his is the trap: AI is so powerful, such a glittering prize, that it is very difficult for human civilization to impose any restraints on it at all."

139. Dr. Altman also suggests the disruptions to education I previously identified are actually the fault of "outdated assessment methods rather than a loss of learning capacity" and that the responsibility actually lies with institutions that must "adapt[] their instructional and evaluation frameworks.[212] Dr. Altman argues many of the issues I previously identified are "already being actively addressed," and lists cherry-picked examples of how some educators have integrated AI into their classrooms or implemented new methods of assessing their students.[213]

140. In my view, Dr. Altman continues to ignore that many of these institutional adaptations may not be readily available to under-funded institutions or that educators in K-12 schools may lack the resources to detect, monitor, or control the use of AI in classrooms. Nor does Dr. Altman propose any viable solutions that may be implemented at scale, choosing instead to highlight just some of the ways some well-funded institutions and educators, like Stanford

---

unusually-painful-disruption-jobs.html; Ana Altchek & Sarah Perkel, *Anthropic CEO Says AI Could Wipe Out Half of All Entry-Level White-Collar Jobs*, BUSINESS INSIDER (May 28, 2025), https://www.businessinsider.com/anthropic-ceo-warning-ai-could-eliminate-jobs-2025-5;

[210] Dario Amodei, *The Adolescence of Technology* (Jan. 2026), https://www.darioamodei.com/essay/the-adolescence-of-technology.

[211] *Id.*

[212] Altman Rebuttal ¶ 35.

[213] *See* Altman Rebuttal ¶¶ 41–51.

University, have chosen to mitigate some of the harms of AI in educational settings.[214] It is the stakeholders with the least resources and most vulnerable who will face the most severe of the challenges of LLMs.

141. I also do not find Dr. Altman's endorsement of Claude's benefits in the fields of science and medicine to be reliable or compelling. For example, the majority of Claude applications in science and medicine that Dr. Altman cites in Section II.B. are vendor case studies and testimonials ***from Anthropic and its partners***.[215] Dr. Altman does not offer any empirical evidence or cite to any independent evaluations of Claude's capabilities (or drawbacks) in these settings.

[214] *See id.* at ¶ 42, 49.

[215] *See id.* at nn.114–15, nn.119–21.

Executed in Chicago, IL on this 13th day of February, 2026.



Ben Zhao

REPLY EXPERT REPORT OF BEN ZHAO

APPENDIX 1: DOCUMENTS CONSIDERED

DISCOVERY, PLEADINGS, AND PRODUCED DOCUMENTS

Action, Discovery, or Pleading
Anthropic's Amended and Supplemental Responses and Objections to Plaintiffs' First Set of Interrogatories (Nos. 9, 11, 12) (Nov. 12, 2025)
Expert Report of Dr. Ben Zhao (Dec. 19, 2025)
Expert Report of Dr. Shawn Shan (Dec. 19, 2025)
Expert Report of Jeffrey Bilmes, Ph.D (Dec. 19, 2025)
Expert Report of Joshua W. Dennis (Dec. 19, 2025)
Rebuttal Expert Report of Dr. Ben Zhao (Jan. 23, 2026)
Expert Rebuttal Report of Dr. Jeffrey Bilmes, PhD (Jan. 23, 2026)
Expert Rebuttal Report of Dr. Russ Altman (Jan. 23, 2026)
First Amended Complaint, ECF No. 337

Deposition Transcripts and Exhibits
Benjamin Mann 30(b)(6) Transcript and Exhibits (November 20, 2025)
Dawn Drain 30(b)(l) Transcript and Exhibits (November 12, 2025)
Pranay Sangani Deposition Transcript and Exhibits (November 10, 2025)
Thomas Brown 30(b)(1) Transcript and Exhibits (January 22 & January 26, 2026)
Vinay Rao Deposition Transcript and Exhibits (November 7, 2025)

BATES-NUMBERED DOCUMENTS
Anthropic_0000217026
Anthropic_0000343228
Anthropic_0000346083
Anthropic_0000347810
Plaintiffs_000000000–2677
Plaintiffs_0000004584–5062
Plaintiffs_0000005198–6967
Plaintiffs_0000009389–9525

INTERNET AND ACADEMIC REFERENCES

1. *7 Rings – Ariana Grande*, ST LYRICS, https://stlyrics.org/songs/a/arianagrande71622/7rings3172807.html (last visited Feb. 12, 2026).

2. Ana Altchek & Sarah Perkel, *Anthropic CEO Says AI Could Wipe Out Half of All Entry-Level White-Collar Jobs*, BUSINESS INSIDER (May 28, 2025), https://www.businessinsider.com/anthropic-ceo-warning-ai-could-eliminate-jobs-2025-5.

3. A. Feder Cooper, et al., *Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models,* ARXIV:2505.12546v4 (Nov. 30, 2025), https://arxiv.org/pdf/2505.12546

4. Adam Tauman Kalai et al., *Why Language Models Hallucinate,* ARXIV:2509.04664 (Sep. 4, 2025), https://arxiv.org/pdf/2509.04664.

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

6. Ahmed Ahmed et al., *Extracting Books from Production Language Models,* ARXIV:2601.02671V1 (Jan. 6, 2026) https://arxiv.org/pdf/2601.02671.

7. Ahmed B. Mustafa et al., *Anyone Can Jailbreak: Prompt-Based Attacks on LLMs and T2Is*, ARXIV:2507.21820 (July 29, 2025), https://doi.org/10.48550/arXiv.2507.21820.

8. Andrew Ng, *Weekly Issues: Issue* 332, THE BATCH (Dec. 17, 2025), https://www.deeplearning.ai/the-batch/issue-332/.

9. Antoine Bellemare-Pepin et al., *Divergent Creativity in Humans and Large Language Models*, SCIENTIFIC REPORT, 16, 1279 (Jan. 21, 2026), https://doi.org/10.1038/s41598-025-25157-3.

10. Catherine Thorbecke, *AI Tools Make Things Up a Lot, and That's a Huge Problem*, CNN (Aug. 29, 2023), https://www.cnn.com/2023/08/29/tech/ai-chatbot-hallucinations.

11. Cem Anil et al., *Many-Shot Jailbreaking*, ANTHROPIC (Apr. 2, 2024), https://www-cdn.anthropic.com/af5633c94ed2beb282f6a53c595eb437e8e7b630/Many_Shot_Jailbreaking__2024_04_02_0936.pdf.

12. Cem Dilmegani & Aleyna Daldal, *AI Hallucination: Compare Top LLMs Like GPT-5.2*, AIMULTIPLE (Jan. 23, 2026), https://research.aimultiple.com/ai-hallucination.

13. Chandler Smith et al., *Evaluating Generalization Capabilities of LLM-Based Agents in Mixed-Motive Scenarios Using Concordia*, ARXIV:2512.03318V1 (Dec. 3, 2025), https://arxiv.org/pdf/2512.03318v1.

14. Chen Wang et al., *Do LLMs Have the Generalization Ability in Conducting Causal Inference?*, ARXIV:2410.11385V1 (Oct. 15, 2024), https://arxiv.org/pdf/2410.11385v1.

15. Damien Charlotin, *AI Hallucination Cases Database*, https://www.damiencharlotin.com/hallucinations/ (last visited Feb. 12, 2026).

16. Emilia Napolano, *Sora Inappropriate and Harmful Content Creation Easily Bypassed Through Simple Prompt Engineering*, ZENODO (Apr. 28, 2025), https://zenodo.org/records/15295087.

17. Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (Mar. 2021), 610–623, https://dl.acm.org/doi/10.1145/3442188.3445922.

18. Feng Zhou, et al., *Approximate Object Location and Spam Filtering on Peer-to-Peer Systems*, PROC. OF THE ACM/IFIP/USENIX INT.L CONF. ON DISTRIBUTED SYS. PLATFORMS, June 2003, https://link.springer.com/chapter/10.1007/3-540-44892-6_1.

19. Grégoire Deletang et al., *Language Modeling is Compression*, 12TH INT'L. CONF. ON LEARNING REP. (2024), https://openreview.net/pdf?id=jznbgiynus.

20. Giorgio Franceschelli et al., Training Foundation Models as Data Compression: On Information, Model Weights, and Copyright Law, ARXIV:2407.13493V3 (Oct. 7, 2024), https://arxiv.org/abs/2407.13493v3.

21. 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-aremathematically-inevitable-not-just-engineering-flaws.html.

22. *Hallucination Leaderboard*, https://github.com/vectara/hallucination-leaderboard (last visited Feb. 12, 2026).

23. Huying Li et al., *Blacklight: Scalable Defense for Neural Networks against Query-Based Black-Box Attacks,* Proc. of USENIX Sec. Symp. (Aug. 2022), https://www.usenix.org/conference/usenixsecurity22/presentation/li-huiying.

24. Ian Webster, *Jailbreaking LLMs: A Comprehensive Guide (With Examples)*, PROMPTFOO (Jan. 7, 2025), https://www.promptfoo.dev/blog/how-to-jailbreak-llms.

25. Igor Shilov et al., *Beyond Data Filtering: Knowledge Localization for Capability Removal in LLMs*, ANTHROPIC ALIGNMENT BLOG (Dec. 8, 2025), https://alignment.anthropic.com/2025/selective-gradient-masking/.

26. Igor Shilov et al., *The Mosaic Memory of Large Language Models*, NATURE COMM. (2026), https://doi.org/10.1038/s41467-026-68603-0.

27. Jeffrey Diver, *New Poe Apps for Web now, ioS & Android Later*, QUORA (Mar. 2025), https://poeai.quora.com/New-Poe-Apps-for-Web-now-ioS-Android-Later.

28. John Kubiatowicz et al., *OceanStore: An Architecture for Global-scale Persistent Storage*, 35 ACM SIGPLAN NOTICES 190, https://dl.acm.org/doi/10.1145/356989.357007.

29. *Imagine Dragons – Radioactive (Remix) Lyrics and Video (Ft. Kendrick Lamar)*, LYRICS-VIDEO-MUSIC, https://lyrics-video-music.blogspot.com/2014/01/imagine-dragons-radioactive-remix.html (last visited Feb. 12, 2026).

30. *Introducing Contextual Retrieval*, ANTHROPIC (Sep. 19, 2024), https://www.anthropic.com/engineering/contextual-retrieval.

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

32. Jing Huang et al., *Demystifying Verbatim Memorization in Large Language Models*. STANFORD AI LAB BLOG (Apr. 28, 2025), https://ai.stanford.edu/blog/verbatim-memorization.

33. John Hughes et al., *Best-of-N Jailbreaking*, ARXIV:2412.03556 (Dec. 4, 2024), https://arxiv.org/abs/2412.03556.

34. Joseph Spracklen et al. *We Have a Package for You! A Comprehensive Analysis of Package Hallucinations by Code Generating LLMs*. 34TH USENIX SEC. SYMP. (Aug. 2025), https://www.usenix.org/conference/usenixsecurity25/presentation/spracklen.

35. Katherine Lee et al., *Deduplicating Training Data Makes Language Models Better*, ARXIV:2107.06499 (Mar. 24, 2022), https://arxiv.org/pdf/2107.06499.

36. Lei Huang et al., *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*, arXiv:2311.05232 (Nov. 19, 2024), https://arxiv.org/pdf/2311.05232.

37. Louis Columbus, *How Anthropic's Safety Obsession Became Enterprise AI's Killer Feature*, VENTUREBEAT (Dec. 16, 2025), https://venturebeat.com/security/how-anthropics-safety-obsession-became-enterprise-ais-killer-feature.

38. Nikhil Kandpal et al., *Deduplicating Training Data Mitigates Privacy Risks in Language Models*, ARXIV:2202.06539 (Feb. 12, 2022), https://doi.org/10.48550/arXiv.2202.06539.

39. 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.

40. Patrick Lewis et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, arXiv:2005.11401v4 (Apr. 12, 2021), https://doi.org/10.48550/arXiv.2005.11401.

41. *Reduce Hallucinations*, CLAUDE API DOCS, https://platform.claude.com/docs/en/test-and-evaluate/strengthen-guardrails/reduce-hallucinations.

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

43. Sara Sams, *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.

44. Sawdah Bhaimiya, *Anthropic CEO Dario Amodei Warns AI May Cause 'Unusually Painful' Disruption to Jobs*, CNBC (Jan. 17, 2026), https://www.cnbc.com/2026/01/27/dario-amodei-warns-ai-cause-unusually-painful-disruption-jobs.html.

45. Sean Rhea et al., *Pond: The OceanStore Prototype*, Proc. of the 2nd USENIX Conf. on File and Storage Technologies (FAST'03) (Mar. 2003), https://www.usenix.org/conference/fast-03/pond-oceanstore-prototype.

46. *SSD vs HDD: Understanding SSD Speed and Performance Differences*, ADATA (Jul. 17, 2025), https://www.adata.com/en/quikTips/ssd-vs-hdd-speed-and-performance-difference.

47. Tom Brown et. al, *Language Models are Few-Shot Learners*, ARXIV:2005.14165V4 (July 22, 2020), https://arxiv.org/abs/2005.14165.

48. *Tracing the Thoughts of a Large Language Model*, ANTHROPIC (Mar. 27, 2025), https://www.anthropic.com/research/tracing-thoughts-language-model.

49. Udi Manber, *Finding Similar Files in a Large File System*, WTEC'94: PROC. OF THE USENIX WINTER 1994 TECH. CONF. (Jan. 17, 1994), https://dl.acm.org/doi/10.5555/1267074.1267076.

50. Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, JOURNAL OF EMPIRICAL LEGAL STUDIES (Mar. 2025), https://dho.stanford.edu/wp-content/uploads/Legal_RAG_Hallucinations.pdf.

51. Wan Zhang & Jing Zhang, *Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review*, 13 MATHEMATICS 2025, 856 (Mar. 4, 2025), https://www.mdpi.com/2227-7390/13/5/856.

52. VDURA, 2 TB/s Per-Rack Flash, www.vdura.com (last visited Feb. 13, 2026).

53. Yakai Li et al., *Prefill-Level Jailbreak: A Black-Box Risk Analysis of Large Language Models*, ARXIV:2504.21038V2 (Aug. 25, 2025), https://arxiv.org/abs/2504.21038v2.

54. Yihao Huang et al., *Perception-Guided Jailbreak Against Text-to-Image Models*, ARXIV:2408.10848 (Aug. 20, 2024), https://doi.org/10.48550/arXiv.2408.10848.

55. Yuntao Bai, et al., *Constitutional AI: Harmlessness From AI Feedback*, ARXIV:2212.08073V1 (Dec. 15, 2022), https://arxiv.org/pdf/2212.08073.

56. Zhongxiang Sun et al., *ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability* (Jan. 21, 2025), ARXIV:2410.11414V2, https://arxiv.org/pdf/2410.11414.

57. Ziwei Xu, *Hallucination is Inevitable: An Innate Limitation of Large Language Models*, ARXIV:2401.11817 (Jan. 22, 2024), https://doi.org/10.48550/arXiv.2401.11817.