EXHIBIT 18

(FILED UNDER SEAL)

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN JOSE DIVISION

CONCORD MUSIC GROUP, INC., ET AL., Plaintiffs, v. ANTHROPIC PBC, Defendant.	Case Number: 5:24-cv-03811-EKL-SVK Judge Eumi K. Lee Magistrate Judge Susan van Keulen

REBUTTAL EXPERT REPORT OF DR. BEN ZHAO

January 23, 2026

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

TABLE OF CONTENTS

I. OVERVIEW AND SCOPE OF ENGAGEMENT 1

II. REBUTTAL TO DR. BILMES'S OPINIONS 2

A. Dr. Bilmes Does Not Adequately Address LLM Propensity to Memorize 2

1. The Ability to Generalize Does Not Decrease a Model's Tendency to Memorize 5

2. LLMs Memorize Because They Do Not "Learn" In the Same Way Humans Do 7

B. Dr. Bilmes's Claim That Claude Was Not Designed to Copy or Regurgitate Publishers' Lyrics is Flawed 12

C. Dr. Bilmes Relies on Flawed Benchmarks for Evaluating LLM Performance, Reliability, and Alignment 14

D. Dr. Bilmes's Assessment of Claude Is Erroneous 19

E. Dr. Bilmes Lacks Foundation to Conclude that Anthropic's Uses of Publishers' Lyrics is "Transformative" 21

F. Dr. Bilmes Mischaracterizes the Import of Robots.txt in Granting Permissions to Crawl Websites with Publishers' Lyrics 26

G. Dr. Bilmes Underestimates the Extent and Impact of Anthropic's Copying of Publishers' Lyrics Multiple Times Despite its Deduplication and Filtering Efforts 30

1. Many Copies of Publishers' Works in Suit Reside in Claude's Training Data 30

H. Dr. Bilmes Incorrectly Concludes Publishers' Lyrics Are Immaterial to LLM Performance 35

I. Dr. Bilmes Ignores Fundamental Flaws in Claude's Guardrails 41

J. Dr. Bilmes' Analysis of Anthropic's Guardrails is Flawed and Unreliable 42

III. REBUTTAL TO DR. ALTMAN'S OPINIONS 44

A. Dr. Altman Lacks Expertise Concerning Generative AI 44

B. Dr. Altman Ignores Limitations and Dangers of LLMs 45

1. Dr. Altman Does Not Address LLM Propensity to Hallucinate and Commit Errors 45

2. Dr. Altman Fails to Address the Many Harms the Use of LLMs Inflict in Educational Settings 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 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]

2. On the same day that I served my initial report, Defendant Anthropic PBC ("Anthropic") served a series of expert reports, including reports from Dr. Jeffrey Bilmes[2] and Dr. Russ Altman.[3]

3. As part of my retention by Publishers, I have been asked to review and comment on sections of the expert reports of Dr. Bilmes and Dr. Altman that are relevant to my work and expertise in this matter.

4. In this rebuttal expert report, I provide my assessment of and opinions of certain aspects of the work performed by Anthropic's experts and their opinions.

5. 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.

6. 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,

[1] Expert Report of Dr. Ben Zhao, PhD (Dec. 19, 2025) (hereinafter "Zhao Report").
[2] Expert Report of Dr. Jeffrey Bilmes, PhD (Dec. 19, 2025) (hereinafter "Bilmes Report").
[3] Expert Report of Dr. Russ B. Altman, MD (Dec. 19, 2025) (hereinafter "Altman Report").

drawings, tables, and/or animations that may be appropriate to supplement and demonstrate my opinions at trial, deposition, or any hearing at which I may be called to testify.

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

8. 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. Rebuttal to Dr. Bilmes's Opinions

A. Dr. Bilmes Does Not Adequately Address LLM Propensity to Memorize

9. Dr. Bilmes's general conclusion that large language model ("LLM") memorization and regurgitation is aberrational behavior[4] is incorrect. On the contrary, LLM memorization and regurgitation are foundational and fundamental to the dominant LLM architecture (autoregressive transformers).

10. Dr. Bilmes states that "a pre-trained model no longer relies on the original texts but instead only relies on the input tokens and the weights resulting from the training process, i.e., the statistical relationships between likely word sequences."[5] Dr. Bilmes is correct in that a pre-trained model does not specifically recall and then display texts from a database of training corpora in response to user prompts.[6] But, contrary to Dr. Bilmes's assertion, this does not mean a pre-trained model does not rely on the original texts or that it is incapable of memorizing or regurgitating texts from its training corpus, or that it was not intended to have such regurgitation or memorization capabilities.

11. As discussed in my initial report, conditional probability training results in LLMs memorizing unique phrases or sequences of phrases that do not commonly occur across multiple

[4] *See* Bilmes Report, ¶¶21, 79, 287, 289, 296.
[5] *Id.* ¶179.
[6] *See id.* at ¶¶283-84.

training samples, such as specific turns of phrases found in song lyrics.[7] This results in regurgitation, wherein a model reproduces a specific text sequence in the same order that matches text in its training corpus. The model has not memorized the specific text in the sense of literally storing an exact copy of that text, and it does not strictly refer back to and fetch that sequence from its training corpus when it regurgitates. A pre-trained model does not scan for text from its training corpus and reproduce what it finds when asked. However, LLMs *do* memorize highly-likely conditional *probabilities* based on their training corpora (*e.g.*, the sequence of tokens "Drove my Chevy to the levee, but the levee was dry" *very- and highly-likely* follows the sequence "bye-bye, Miss American Pie"), which results in the same outcome: the reproduction of specific phrases or entire texts contained in its training corpus.[8] This memorization of highly-likely conditional probabilities is fundamental to the transformer architecture of LLMs.[9]

12. Various research studies in the LLM and machine learning fields have confirmed this phenomenon. For example, a 2025 study from Stanford and Cornell Universities found that certain LLMs can memorize entire books, and one was able to regurgitate verbatim the entire text of *Harry Potter and the Sorcerer's Stone*, using a prompt containing just a few tokens from the first chapter.[10] A more recent study (2026) found that a number of LLMs (including Claude 3.7 Sonnet) could regurgitate almost entire books verbatim, and in fact that out of the four major LLMs tested (Claude 3.7 Sonnet, Gemini 2.5 Pro, Grok 3, and GPT-4.1), ***Claude 3.7 Sonnet regurgitated the highest percentage*** of *Harry Potter and the Sorcerer's Stone*.[11]

---

[7] *See* Zhao Report ¶¶26–28.

[8] *Id.* ¶¶27–28.

[9] *Id*.

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

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

13. Dr. Bilmes agrees that when data is repeated multiple times in training data it is more likely to be memorized.[12] The likelihood of memorization also increases with texts and phrases that are unique in structure and word choice.[13]

14. This tendency to memorize is true even after a model has been fine-tuned. Dr. Bilmes states that a fine-tuned model "can support downstream applications . . . [and] these adaptations do not reflect memorization or regurgitation of the training examples."[14] This is misleading and incorrect, for three reasons.

15. First, there are multiple methods for adapting LLMs to downstream tasks, including the widely-used approach of supervised fine-tuning (SFT), which trains the LLM to learn task-specific knowledge from new annotated data. Multiple studies have shown that fine-tuning LLMs for downstream tasks leads to substantial memorization of the training data.[15]

16. Second, Dr. Bilmes only discussed two other fine-tuning approaches: reinforcement learning (RL) and reinforcement learning with human feedback (RLHF). RL fine-tuning starts from SFT, then trains a reward model to score model output that mimics human judgement and uses it to further optimize the SFT model. RLHF is a special version of RL where it trains the reward model using human-annotated feedback. But even where RL or RLHF fine-tuning are used in LLM training, the preceding SFT step of the fine-tuning process still leads to memorization of the training data, and data already memorized during the SFT stage tend to remain memorized after the reward-based optimization.[16] This is consistent with the common

[12] Bilmes Report ¶173.

[13] *See* Zhao Report ¶26; *Copyright and Artificial Intelligence, supra* n. 41 at 19–21.

[14] Bilmes Report ¶79.

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

[16] *See* Aneesh Pappu et al., *Measuring Memorization in RLHF for Code Completion*, PROC. OF THE THIRTEENTH INTERNATIONAL CONF. ON LEARNING REPRESENTATIONS (ICLR) (April 2025),

understanding that fine-tuning is a compute-efficient way to modify small portions of the model, but the large majority of the model behavior remains intact through fine-tuning.

17. Finally, as I have mentioned earlier, numerous studies have demonstrated the fundamental behavior of memorization and regurgitation in LLMs.[17] Fine-tuning with new data does not remove such fundamental properties from LLMs.

1. The Ability to Generalize Does Not Decrease a Model's Tendency to Memorize

18. Although Dr. Bilmes claims that "[g]eneralization is fundamentally distinct from, and opposed to, memorization,"[18] the reality is that even a model intended to generalize and not memorize from its training data is nonetheless capable and even likely to regurgitate training data.[19] Generalization and memorization are dictated by the presence and density of training data, where training data that is unique or repeated leads to stronger memorization.[20] The fundamental nature of autoregressive transformers produces the memorization properties given specific patterns in training data, regardless of training or alignment instructions.

19. Generalization requires a comprehensive set of training samples that vary in meaningful dimensions so that clear rules can be extracted, and generalization is not possible when training data is incomplete or biased towards specific conditions. The now-famous wolf vs. husky classification example illustrates this well. A model learning to generalize how to recognize the difference between wolves and husky dogs consistently made mistakes on specific images of

---

https://proceedings.iclr.cc/paper_files/paper/2025/file/22811d2089d4aa0ba66e52a5e47efb65-Paper-Conference.pdf.

[17] *See* Zhao Report ¶¶63, 122.

[18] Bilmes Report ¶87.

[19] *See* Zhao Report ¶26.

[20] *See* Vitaly Feldman, *Does Learning Require Memorization? A Short Tale About a Long Tail,* PROC. OF THE 52ND ANN. ACM SIGACT SYMP. ON THEORY OF COMP. 954 (Jun. 22, 2020), https://doi.org/10.1145/3357713.3384290; *see also* Vitaly Feldman & Chiyuan Zhang, *What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation*, NIPS'20: Proc. of the 34th Int. Conf. on Neural Inf. Proc. Sys. 2881, https://proceedings.neurips.cc/paper/2020/file/1e14bfe2714193e7af5abc64ecbd6b46-Paper.pdf.

wolves.[21] Researchers discovered that because nearly all training images of wolves were taken with snowy backdrops, the model incorrectly generalized that the critical identifying feature of a wolf was the white snow backdrop.[22]

20. This is the same principle that makes generative models memorize. When training data is focused around one unusual element, the model has no alternative baselines to generalize on. For example, given a relatively unique phrase from the lyrics of "Fresh Prince of Bel-Air" by Will Smith such as "Chillin' out, maxin', relaxin' all cool," the model will learn the probability is near 100% that the next phrase is "And all shootin' some b-ball outside of the school," since those two phrases appear next to each other in most contexts. It cannot generalize when there are no other training samples to provide an alternative to memorization. Thus it is the architecture and the training data that dictate memorization, not training methodology or instructions given to the model.

21. Dr. Bilmes claims that "LLMs do not store or retrieve passages of text verbatim."[23] As I have previously explained, while LLMs do not "store" passages of texts in a central database, they nonetheless are prone to regurgitating verbatim text, particularly where unique phrases or sequences do not commonly appear across training samples. The effect is the same as if the LLM had stored the text in a database.

22. Indeed, Dr. Bilmes acknowledges that "[w]hen a model outputs verbatim or near verbatim copies of information contained in the training data, researchers say that the training data may have been memorized," and that "[r]esearchers have found that data repeated many times in

---

[21] *See* Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, "*Why Should I Trust You?": Explaining the Predictions of Any Classifier*, ARXIV:1602.04938v3 (Aug. 9, 2016), https://arxiv.org/pdf/1602.04938.

[22] *Id.*

[23] Bilmes Report ¶98.

the training data is more likely to be memorized."[24] This is consistent with contemporary research,[25] Anthropic's own findings,[26] and my initial report.[27]

23. And as Dr. Bilmes points out, recent research suggests there is an optimal data size for a model, and that merely increasing training datasets does not results in better performance.[28] This undermines Dr. Bilmes's suggestion that a large volume of text "is necessary for the model to develop broad knowledge and generalize across many general-purpose tasks" and that "large text collections . . . and other large-scale web crawling efforts" are necessary to "obtain sufficient data for the types of advanced deep LLM training that has made general purpose LLMs like Claude possible."[29] To the contrary, numerous research efforts such as Microsoft's Phi project have shown that better training can extract more value from existing datasets and improve model quality much more than simply adding more scraped data.[30]

2. LLMs Memorize Because They Do Not "Learn" In the Same Way Humans Do

---

[24] *Id.* ¶173.

[25] See *supra,* n. 20 and accompanying text; Antonia Karamolegkou et al., *Copyright Violations and Large Language Models*, Proc. of the 2023 Conf. on Empirical Methods in Nat. Language Processing 7403 (2023), https://aclanthology.org/2023.emnlp-main.458/; Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models*, 2022 Int. Conf. on Learning Representations (Feb. 15, 2022), https://arxiv.org/pdf/2202.07646.

[26] *See, e.g.,* D. Drain Dep. Tr. 90:2–24.

[27] *See* Zhao Report Section IV.A.

[28] *See* Bilmes Report ¶165 & n.311.

[29] *Id.* ¶100.

[30] Marah Abdin et al., *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*, ARXIV:2404.14219 (Aug. 30, 2024), https://arxiv.org/pdf/2404.14219; Sally Beatty, *Tiny But Mighty: The Phi-3 Small Language Models with Big Potential*, MICROSOFT (Apr. 23, 2024), https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/; *Introducing Phi-4: Microsoft's Newest Small Language Model Specializing in Complex Reasoning*, MICROSOFT FOUNDRY BLOG (Dec. 12, 2024), https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/introducing-phi-4-microsoft%E2%80%99s-newest-small-language-model-specializing-in-comple/4357090.

24. As I explained in my initial report, a large part of why LLMs are susceptible to memorization is due to their neural network architectures.[31]

25. Despite the connotations of phrases like "neural networks" or "deep neural networks," LLMs and other machine learning models trained using such methods do not "learn" in the same way that human beings do.[32] Human beings are capable of both bottom-up processing (reactive responses to external stimuli), as well as top-down regulation. Top-down thinking is a result of human evolution that has integrated "top-down networks" into the prefrontal cortex, allowing decision-making based on knowledge, concepts, expectations and other cognitive mechanisms.[33]

26. Top-down thinking is evident in the ways we learn as children, where the emphasis lies on learning generalizable rules, and how to apply them to inputs. For example, mathematics is taught on principles of functional properties that can be applied to a wide range of inputs. We are taught that operations like multiplication and addition have commutative and associative properties, which can then be applied to the wide range of possible numeric inputs. Given these rules, we are then shown examples to instantiate those rules and demonstrate their application in practice. Ask most middle-school or high-school students how to simplify fractions, and they will tell you the general rules that can be applied regardless of the actual numbers.

27. Most importantly, we are explicitly taught these rules. Unlike AI models, we do not infer them from data points we observe. For example, we teach children that fire or anything that is on fire will burn your hands if you touch it. The rule is applied (and hopefully, internalized) without the child actually seeing direct evidence of the consequences of the action (a burn injury). In contrast, AI models associate pairs of inputs and outputs, and try to generalize rules based on

---

[31] *See* Zhao Report, Section IV.A.2.

[32] *See* Liz Mineo, *Is AI Dulling Our Minds?*, THE HARVARD GAZETTE (Nov. 13, 2025), https://news.harvard.edu/gazette/story/2025/11/is-ai-dulling-our-minds/.

[33] *See* David Krauss, *Bottom-Up Versus Top-Down Processing: How Interventions Work*. PSYCHOLOGY TODAY, (Nov. 1, 2021), https://www.psychologytoday.com/us/blog/atypical-children-extraordinary-parenting/202111/bottom-up-versus-top-down-processing-how.

the samples of data it has in the training set. If children learned like AI, they would have to put their hands into fire in order to learn the outcome, and they would have to do so in multiple different contexts, before learning that all fire can burn their hands. This is also referred to as "rule-based" and "exemplar-based" generalization in cognitive psychology.[34]

28. By contrast, an AI model "learns" by inferring rules from its available training data, which makes it dependent on the completeness and accuracy of its training data. The wolf vs. husky classifier research is again instructive.[35] As discussed, the classifier trained extensively on images of both wolves and huskies sometimes failed badly by identifying pictures of other animals as wolves where those pictures included snow, because nearly all of its training data on wolves likewise included snow.[36] The model inferred, based on its training data, that the key identifying feature of a wolf is whether there is snow in the background.

29. While LLMs are architecturally more complex than typical deep neural networks used in classification, they operate on similar principles when it comes to generalization. LLMs generate outputs based on features and correlations extracted from training data. If a song is labeled in texts as a hip-hop song, and its lyrics include a particular unusual word which does not appear elsewhere, then it will infer that word is specific to hip-hop music.

30. Another example that demonstrates this is the well-known "reversal curse," where LLMs trained on "A is B" fail to learn the reverse.[37] Across numerous models, asking the model about the name of a family member of a star produces the correct name, *e.g.*, the prompt "Who is Tom Cruise's mother?" results in the output "Mary Lee Pfeiffer." But asking the same model about that family member fails to produce the relationship to the star; *i.e.*, a prompt asking "Who is Mary

---

[34] Ishita Dasgupta et al., *Distinguishing Rule and Exemplar-Based Generalization in Learning Systems*, PROC. OF THE 39TH INT. CONF. ON MACHINE LEARNING (Vol. 162), 4816–4830 (2022), https://proceedings.mlr.press/v162/dasgupta22b.html.

[35] Ribeiro, *supra* n. 21.

[36] *Id.* at 8.

[37] Lukas Berglund, et al., *The Reversal Curse: LLMs Trained on "A Is B" Fail to Learn "B Is A,"* ARXIV:2309.12288v3 (Apr. 4, 2024), https://arxiv.org/pdf/2309.12288v3.

Lee Pfeiffer's son?" produces an error, not "Tom Cruise." This asymmetry is clearly explained by the fact that LLMs are extracting probabilities from sequences of tokens, and see many more examples of training data referring to Tom Cruise's mother, than they do "Mary Lee Pfeiffer's son." Where a human is taught "if A is a man, and A's mother is B, then A is B's son," an LLM has no way to learn these rules outside of its training data, and incomplete data will lead to failures such as the reversal curse.

31. LLMs are dependent on training data to build conditional probability distributions for unforeseen scenarios and questions, and LLMs do not generalize well to topics and questions that are not included within their training data.[38] In order to improve the performance of AI models when more complex reasoning is called for, researchers have developed Large Reasoning Models (LRMs). LRMs are largely combinations of two different techniques applied on top of LLMs: chain-of-thought[39] and self-verification.[40] Unfortunately, thus far LRMs do not fix fundamental issues within LLMs.

32. Chain-of-thought ("CoT") is the core idea that LLMs may make errors in answering many complex queries, and those errors might be reduced if complex queries were reduced down into simpler, smaller components that combined to form the more complex query. Since the original CoT paper was published, numerous efforts have expanded CoT in LLMs to build the basis for today's large reasoning models, particularly for complex domain problems such as mathematics or logical problems. Beyond adapting to specific domains, literature has focused on how to generate partial prompts and how to select the best combination of partial prompts.

---

[38] *See* Steve Yadlowsky, Lyric Doshi & Nilesh Tripuraneni, *Pretraining Data Mixtures Enable Narrow Model Selection Capabilities in Transformer Models*, ARXIV:2311.00871v1 (Nov. 3, 2023), https://arxiv.org/pdf/2311.00871.

[39] Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, 36TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS) (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[40] Yixuan Weng et al., *Large Language Models Are Better Reasoners with Self-Verification*, 2023 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP), 2550–2575 (2023), https://aclanthology.org/2023.findings-emnlp.167.pdf.

33. However, work on CoT has not fundamentally changed the fundamental properties of the underlying LLM. At its core, CoT reduces the complexity of larger problems (sometimes called "problem decomposition"), but does not imbue the LLM with the ability to generalize beyond its training data, nor its fundamental reliance on probability distributions that produce errors and hallucinations. When LLMs produce an error in a smaller question, the errors will propagate and lead to errors in the original larger complex query that generated the chain of smaller queries.

34. Similarly, the introduction of self-verification does not qualitatively change the LLM's inability to learn factual information. Self-verification is the process whereby a single query can be given to the LLM written in many different ways, prompting it to produce a large number of potentially different results, and then aggregating those results and presenting them back to the LLM and asking it to verify them or choose the correct answer from amongst them. Self-verification effectively asks a LLM to perform the same query multiple times thereby reducing (but not removing) the likelihood of mistakes or hallucinations. But these techniques, including CoT, do not represent learning like humans do. They simply train the model to respond to a stack of inferences, one at a time. This is another example of why LLMs do not learn like humans. A child can learn how to improve their math skills once, and utilize it for the rest of their lives. We would not expect a child or any human to "improve their problem solving" by asking them to solve 100 slightly different versions of a problem, then to examine their own answers and pick the one they consider most likely to be correct.

35. This difference between human learning and AI "learning" is intuitive to non-experts as well. For example, the U.S. Copyright Office agrees that AI training is much unlike human learning because of the ease with which it perfectly copies and analyzes content at speeds much faster than any human capability:

> "Nor do we agree that AI training is inherently transformative because it is like human learning. . . . Copyright law should not afford greater latitude for copying simply because it is done by a computer. Moreover, AI learning is different from human learning in ways that are material to the copyright analysis. Humans retain only imperfect impressions of

> the works they have experienced, filtered through their own unique personalities, histories, memories, and worldviews. Generative AI training involves the creation of perfect copies with the ability to analyze works nearly instantaneously. The result is a model that can create at superhuman speed and scale. In the words of Professor Robert Brauneis, 'Generative model training transcends the human limitations that underlie the structure of the exclusive rights.'"[41]

B. Dr. Bilmes's Claim That Claude Was Not Designed to Copy or Regurgitate Publishers' Lyrics is Flawed

36. In Section V and throughout his Report, Dr. Bilmes opines that Anthropic designed Claude to be a "general-purpose LLM" capable of a "wide range of tasks such as summarization, question-answering, and text generation,"[42] or for "text summarization, data processing, and coding."[43] He contends that "Claude was not designed specifically to make or produce music for the entertainment of a listener."[44] Dr. Bilmes's statements are misleading and contradictory.

37. First, Dr. Bilmes opines that Claude is not meant to reproduce or regurgitate its training data because he believes Claude was designed and built as a general-purpose model. But the mere fact that Claude was intended to be a general-purpose model does not mean it was not also intended to reproduce portions of its training data. Notably, the basis for Dr. Bilmes' assertion that Claude is not designed to make or produce music is a single citation to a page on Anthropic's website stating the "main focus" of Claude 4 is "coding, advanced reasoning, and AI agents," and which "[d]oes not mention music generation as a use case."[45]

38. This reasoning is circular and flawed. Just because music generation or music creation is not cited as a specific use case does not mean Claude 4 (or another version of Claude) is incapable of being used in the music creation process or generating lyrics. It also does not mean Anthropic designed Claude 4 and earlier models *not* to generate and reproduce lyrics. The omission

---

[41] *Copyright and Artificial Intelligence, Part 3: Generative AI Training (pre-publication version)* 48, U.S. COPYRIGHT OFFICE (2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf.

[42] Bilmes Report ¶96.

[43] *Id.* ¶290.

[44] *Id.* ¶96.

[45] *Id.* ¶290 n.538.

of a single use case on a web page owned, published, and operated by Anthropic carries little weight, especially when Anthropic has elsewhere promoted Claude's general creative writing capabilities as a specific use case.[46] Indeed, Anthropic employees have added lyrics to part of Anthropic's training data, and included queries for lyrics in their test cases, indicating that indeed, returning lyrics is part of the intended use case for Anthropic's models.[47]

39. Further, Anthropic designed and built Claude using a specific type of model architecture that, as discussed above, is well understood to inherently lead to memorization and regurgitation, especially when the model's training data includes many copies of a unique text. Anthropic included in Claude's training data hundreds, and in some cases thousands, of copies of Publishers' works.[48] The obvious result of these choices is that Claude memorized and regurgitates Publishers' works.

40. As Dr. Bilmes notes, a general purpose model is capable of performing "many general-purpose tasks."[49] But as Dr. Bilmes also notes, LLM developers do not always know what its models are even capable of from the outset, with large-scale LLMs often developing "unanticipated abilities" (emergent properties and abilities) after training.[50] To conclude that a general purpose model is not designed to perform specific tasks but only "general purpose" tasks is nonsensical. The specific tasks that a general purpose model is capable of are all part of its

---

[46] *See, e.g.,* D. Ganguli 30(b)(1) Dep. Ex. 25 (Anthropic_0000013680); Anthropic_0000000588; B. Mann 30(b)(6) Dep. Ex. 187 (Anthropic_0000349548).

[47] *See, e.g.*, Daniela Amodei Dep. Ex. 179 (Anthropic_0000458491); Anthropic_0000226136; Anthropic_0000240654; Anthropic_0000012260; Anthropic_0000525947; Anthropic_0000459162; Anthropic_0000554366; Anthropic_0000554319; Anthropic_0000554980; Anthropic_0000527903; Anthropic_0000554769; Anthropic_0000016458; Anthropic_0000016025; J. Kaplan Tr. 317:21–324:14; D. Ganguli 30(b)(6) Dep. Tr. 94:9–19; Dataset Card for HH-RLHF, HUGGING FACE, https://huggingface.co/datasets/Anthropic/hh-rlhf (last visited Jan. 22, 2026).

[48] Expert Report of Dr. Shawn Shan, PhD (Dec. 19, 2025) (hereinafter "Shan Report") ¶¶47–50.

[49] Bilmes Report ¶100.

[50] *Id.* ¶168.

"general purpose." Such tasks would include more specific use cases, like writing song lyrics or poems.

41. Second, Dr. Bilmes confuses Publishers' allegations and the conduct at issue in this litigation. Dr. Bilmes seems to suggest that because Claude does not generate and play full compositions—notes, rhythms, harmonies, *and* lyrics—it is a model "not designed specifically to make or produce music for the entertainment of a listener." I understand that Publishers are not alleging that Anthropic has infringed its copyrights because Claude "make[s] or produce music for the entertainment of a listener," but rather that Claude reproduces Publisher-owned lyrics as text and creates lyrics after having been trained on those same Publisher-owned lyrics.[51] Thus, the question is not whether Claude is a model "designed specifically to make or produce music for the entertainment of a listener," but whether Claude is a model designed to reproduce or write music *lyrics*. As I have opined in my initial report, Anthropic designed Claude to do exactly these things.

C. Dr. Bilmes Relies on Flawed Benchmarks for Evaluating LLM Performance, Reliability, and Alignment

42. Dr. Bilmes also opined that model builders use standardized benchmarks to evaluate LLMs, and that "[t]hese benchmarks test broad knowledge, reasoning skills, and mathematical and coding abilities, and allow developers to compare models and identify improvements."[52] However, Dr. Bilmes ignores the fact that many studies and reports have repeatedly shown that evaluation of LLMs on these standardized benchmarks is flawed and unreliable, and that high scores achieved by LLMs on these benchmarks are often deceptive, failing to demonstrate actual improvements.[53] The lack of reliability in benchmark results is due to three

---

[51] *See, e.g.*, First Am. Compl. ¶¶8–10, 63, 77–90, 95, ECF No. 336-2 (hereinafter "FAC").

[52] Bilmes Report ¶15.

[53] *See, e.g.,* Nurit Cohen-Inger et al., *Forget What You Know About LLMs Evaluations—LLMs Are Like a Chameleon*, PROCEEDINGS OF THE 2025 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP), 21664–21677 (2025), https://arxiv.org/abs/2502.07445; Shanchao Liang et al., *The SWE-Bench Illusion: When State-of-the-Art LLMs Remember Instead of Reason*, ARXIV:2506.12286v4 (Dec. 1, 2025), https://arxiv.org/pdf/2506.12286.

reasons: LLM overfitting benchmarks, evaluation malpractice, and the failure of benchmarks to capture real-world scenarios.

43. First, LLM scores on standardized benchmarks can be misleading because of the tendency for LLMs to overfit standardized benchmarks. Overfitting is where the subject machine learning model will provide accurate predictions for patterns that closely match training data, but not for new data.[54]

44. This overfitting occurs because LLMs are trained on data that already includes these benchmarks. Since these benchmarks are publicly available on the internet and repeatedly used to compare models, their questions and often their answers inevitably leak into training data.

45. Each benchmark typically includes test questions and answers. For example, the Massive Multitask Language Understanding ("MMLU") benchmark consists of multiple-choice questions from various disciplines:

> How many attempts should you make to cannulate a patient before passing the job on to a senior colleague?
>
> (A) 4 (B) 3 (C) 2 (D) 1
>
> The correct answer is C.
>
> A model airplane flies slower when flying into the wind and faster with wind at its back. When launched at right angles to the wind, a cross wind, its groundspeed compared with flying in still air is
>
> (A) the same (B) greater (C) less (D) either greater or less depending on wind speed
>
> The Answer is B.[55]

In the above example, the questions posed would be copied into the training data along with the answers, and the LLM would train on both sets of information.

---

[54] *What Is Overfitting?*, AMAZON WEB SERIVCES, https://aws.amazon.com/what-is/overfitting/ (last visited Jan. 22, 2026).

[55] Dan Hendrycks et al., *Measuring Massive Multitask Language Understanding*, ARXIV:2009.03300v3 (Jan. 12, 2021), https://arxiv.org/pdf/2009.03300.

46. This leak can occur directly, through inclusion of benchmark data itself, or indirectly, through user interaction data in which users prompt models with benchmark questions for evaluation purposes. Multiple studies (cited *infra*) have shown that LLMs exploit and memorize surface patterns displayed by these benchmark data to obtain high performance scores, rather than build genuine language understanding and reasoning capabilities.

47. As a result, strong benchmark performance often reflects memorization of specific test questions and patterns, rather than general or transferable capabilities. This explains why LLMs can correctly answer the questions listed in published benchmarks, but fail on slightly rephrased versions, or even minor prompt alterations (for example, changing the order of options in multi-choice questions).[56] Such behavior indicates that the models have been overfitted to the specific linguistic patterns and structures used by the benchmarks, such as the sample questions listed above. This phenomenon of overfitting is consistently observed across widely used benchmarks, including those mentioned by Dr. Bilmes.

48. Despite the known problem of overfitting, certain LLM developers have claimed that their models made significant advances in solving coding and mathematical problems, as reflected by their high scores on standardized benchmarks, including competitive programming questions and a qualifier for the USA Math Olympiad.[57] However, multiple studies have consistently indicated that such gains are largely illusory, because they arise from benchmark memorization rather than genuine improvement in coding and mathematical reasoning.

49. For coding tasks, several LLMs (including the Claude models by Anthropic) claim to achieve high accuracy (>70%) on *SWE-bench*, a widely-used benchmark for programming tasks (e.g. finding bugs in codes and generating corrected codes).[58] However, when evaluated on external benchmarks targeting the same task distribution, their performance drops substantially

---

[56] Cohen-Inger, *supra* n. 53.

[57] *Learning to Reasoning with LLMs*, OPENAI (Sept. 12, 2024), https://openai.com/index/learning-to-reason-with-llms/.

[58] Shanchao Liang et al., *The SWE-Bench Illusion: When State-of-the-Art LLMs Remember Instead of Reason*, arXiv:2506.12286v4 (Dec. 1, 2025), https://arxiv.org/pdf/2506.12286.

(from 73% to 36-51% for Claude models), suggesting model reliance on benchmark-specific memorization rather than genuine problem-solving capability.[59]

50. Similarly, LLMs failed to solve math problems beyond those covered by published benchmarks. Recent reasoning-focused LLMs (including Claude models) are reported to excel at solving mathematical problems from standardized benchmarks like GSM8K, MATH, and the American Invitational Mathematics Examination (AIME), achieving 91.5% accuracy on GSM8K and 94.8% on MATH.[60] However, when researchers evaluated the 11 strongest reasoning-focused LLMs using the 2025 USA Mathematical Olympiad (USAMO) problems (which the models had not been trained on), all of them failed.[61] Out of those, seven models (including the Claude models) scored no higher than 5%, and the best-performing models achieved less than 30%.[62] These models frequently exhibited flawed logical reasoning, relied on unjustified assumptions, and repeatedly applied the same strategies without exploring viable alternatives.[63] Furthermore, while human participants had a clear sense of whether they had successfully (and correctly) solved a problem, all 11 LLMs consistently asserted that they had solved the problems correctly, even though their solutions were incorrect.[64]

51. Detailed analysis reveals that a key reason for such misleading and deceptive behavior in math problems is that current math evaluation benchmarks are fundamentally flawed.[65] These benchmarks focus on the accuracy of final answers while largely neglecting the logical rigor crucial for mathematical problem solving. Trained to optimize performance on these narrow benchmarks, LLMs frequently fail to distinguish actual mathematical reasoning from clearly

---

[59] *Id.*

[60] Hamed Mahdavi et al., *Brains vs. Bytes: Evaluating LLM Proficiency in Olympiad Mathematics*, ARXIV:2504.01995v2 (Apr. 10, 2025), https://arxiv.org/pdf/2504.01995.

[61] Ivo Petrov et al., *Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad*, ARXIV:2503.21934v5 (Sept. 4, 2025), https://arxiv.org/pdf/2503.21934.

[62] *Id.*

[63] *Id.*

[64] *Id.*

[65] Liang, *supra* n. 58.

flawed solutions. Their occasional correct final answers reflect that these models recognize learned patterns during training rather than demonstrating genuine mathematical understanding and reasoning.

52. The second reason for LLM evaluation failures is the absence of a rigorous and transparent methodology for evaluating LLMs, which has led to malpractices by both researchers and companies. In a recent Financial Times interview, Yann LeCun, Meta's outgoing chief AI scientist, said that Meta researchers "fudged a little bit" by using different versions of Llama models on different benchmarks to achieve stronger test results.[66] Studies have found that such malpractice concerning LLM evaluations is not uncommon,[67] further undermining the credibility and interpretability of LLM benchmark evaluations generally.

53. The third source of unreliability in LLM evaluation results is that many benchmarks are constructed from noisy, often unverified datasets. Even widely used LLM benchmarks have been shown to contain errors, ambiguities, and internal inconsistencies. A recent detailed study revealed that the widely used MMLU benchmark contains obvious errors that degrade the credibility of its evaluation result.[68] For example, the benchmark had listed D as the answer to the following question, although A is the correct answer:

---

[66] Mark Sullivan, *Yann LeCun: Meta "Fudged a Little Bit" When Benchmark-Testing Llama 4 Model,* YAHOO FINANCE (Jan. 6, 2026), https://finance.yahoo.com/news/yann-lecun-meta-fudged-little-100000402.html.

[67] Simone Balloccu et al., *Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs*, PROC. OF THE 18TH CONF. OF THE EUROPEAN CHAPTER OF THE ASSOC. FOR COMP. LINGUISTICS VOLUME 1: LONG PAPERS, 67–93 (2024), https://aclanthology.org/2024.eacl-long.5.pdf.

[68] Aryo Pradipta Gema et al., *Are We Done with MMLU?*, PROC. OF THE 2025 CONF. OF THE NATIONS OF THE AMERICAS CHAPTER OF THE ASSOC. FOR COMP. LINGUISTICS: HUMAN LANGUAGE TECH., https://aclanthology.org/2025.naacl-long.262/.

Question: What is the current best option for preventing future outbreaks of Ebola?

A. Rebuild scientific, medical and nursing infrastructure and train staff.
B. Early and accurate diagnosis with molecular kits.
C. Develop effective vaccines.
D. Arrange rapid intervention into West Africa with EU and USA army teams.[69]

54. The study further showed that after re-testing the models on the *corrected* benchmark, ranks of the models changed significantly. For example, Claude's rank (against other LLMs in this benchmark evaluation) ***dropped from first to fifth***.[70]

D. Dr. Bilmes's Assessment of Claude Is Erroneous

55. Dr. Bilmes states that Claude is "safe, transparent, and controllable."[71] In all respects, I disagree.

56. First, Claude's training data is not transparent; Anthropic has not meaningfully disclosed its various training data sources, and has only disclosed very general details (*e.g.*, the model card for Claude 2 stated that its training data was "a proprietary mix of publicly available information from the Internet, datasets that we license from third party businesses, and data that our users affirmatively share or that crowd workers provide").[72]

57. Claude is not "safe" because jailbreaking queries (now built in an entirely automated fashion) are shown to cause LLMs (including the Claude models) to produce harmful content. It is unclear whether such vulnerability can ever be fully patched by LLM providers. The

[69] *Id.*
[70] *Id.*
[71] Bilmes Report ¶12.
[72] Anthropic_0000000588.

nature of the LLM models could make such vulnerabilities inevitable.[73] This vulnerability to jailbreaks continues in newer Claude models, demonstrated by multiple red teaming efforts.[74]

58. Dr. Bilmes's assertion that adversarial attacks and jailbreaks are "unrealistic" in a real-world setting ignores the rapid democratization of jailbreaking techniques through natural language. Unlike traditional software exploits that require deep coding expertise, modern LLM jailbreaks are "low-effort, high-impact" maneuvers "crafted by everyday users with nothing more than cleverly-worded prompts."[75] Research has identified over 1,400 "in-the-wild" jailbreak prompts circulating on public platforms like Reddit, Discord, and specialized prompt-sharing sites where optimized exploits are refined and shared for simple copy-and-paste use.[76] Techniques like Speak Easy demonstrate that even the most robust models can be compromised through "simple" "non-technical" methods involving "interactions commonly encountered in everyday user settings" that decompose harmful requests into seemingly innocent subqueries.[77]

59. Furthermore, accessibility is maximized through low-cost cross-lingual vulnerabilities, where a user can bypass safety filters by simply using a public translation API to convert a prompt into a low-resource language.[78] With open-source automated frameworks like DeepTeam providing these vulnerabilities out-of-the-box on GitHub,[79] the barrier to entry for

---

[73] *See* Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models*, ARXIV:2307.15043v2 (Dec. 20, 2023), https://arxiv.org/pdf/2307.15043; *see also* Mantas Sabeckis & Jurgita Lapienyte, *We Tested ChatGPT, Gemini, and Claude with Adversarial Prompts: Here Are Our Findings and Risks*, CYBERNEWS.COM (Nov. 13, 2025), https://cybernews.com/security/we-tested-chatgpt-gemini-and-claude/.

[74] Mateja Vuradin & Emily Hayes, *How Safe Is Anthropic's "Safest" Model? We Red Teamed Claude Sonnet 4.5*, ZSCALER (Oct. 15 2025), https://splx.ai/blog/red-teaming-claude-sonnet-4-5.

[75] Ahmed B. Mustafa et al., *Anyone Can Jailbreak: Prompt-Based Attacks on LLMs and T2Is*, ARXIV:2507.21820v1 (Jul. 29, 2025), https://arxiv.org/pdf/2507.21820.

[76] Xinyue Shen et al., *"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models*, ARXIV:2308.03825v2 (May 15, 2024), https://arxiv.org/pdf/2308.03825.

[77] Yik Siu Chan et al., *Speak Easy: Eliciting Harmful Jailbreaks from LLMs with Simple Interactions*, ARXIV:2502.04322v3 (Aug. 2, 2025), https://arxiv.org/pdf/2502.04322.

[78] Zheng-Xin Yong, Cristina Menghini & Stephen H. Bach, *Low-Resource Languages Jailbreak GPT-4*, ARXIV:2310.02446v2 (Jan. 27, 2024), https://arxiv.org/pdf/2310.02446.

[79] *See generally,* DEEPTEAM, https://www.trydeepteam.com/ (last visited Jan. 23, 2026).

extracting protected content has effectively vanished. Dismissing these persistent, transferable, and widely accessible techniques as unrealistic fails to account for a digital ecosystem where a typical user can bypass multi-layered guardrails in seconds.[80]

60. Finally, Claude is not "controllable" because it consistently hallucinates, *i.e.*, generates content that is factually incorrect, ungrounded and fabricated, while presenting it in a fluent and confident manner likely to mislead most users. For example, LLMs are known for citing non-existent academic papers or inventing historical events or people not present in its training data or resources. As shown by a recent study by OpenAI researchers, hallucination is a fundamental property of LLMs, and there is ample evidence that hallucination persists in even state-of-the-art LLMs (as illustrated and quantified by multiple ongoing efforts to measure the rate of hallucination).[81] For example, the HHEM study evaluates how often an LLM introduces hallucinations when summarizing a document; it currently shows Claude-haiku-4-5-20251001 has having a 9.8% hallucination rate.[82]

E. Dr. Bilmes Lacks Foundation to Conclude that Anthropic's Uses of Publishers' Lyrics is "Transformative"

61. As an initial matter, "transformativeness" is not a term of art in the fields of computer science, computer engineering, or broadly in generative AI research. Dr. Bilmes's characterization of Claude being "transformative with respect to training inputs"[83] or that LLM training has a "transformative effect"[84] carries no meaning of import to experts such as myself in

---

[80] *See* Wan Chong Choi et al., *A Review of "Do Anything Now" Jailbreak Attacks in Large Language Models: Potential Risks, Impacts, and Defense Strategies,* PREPRINTS.ORG (Sept. 1, 2025), https://www.preprints.org/frontend/manuscript/1955ad799746b0c4ad7ced819c095573/download_pub.

[81] Adam Tauman Kalai et al., *Why Language Models Hallucinate*, OPENAI (Sept. 2025), https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf.

[82] Hallucination Leaderboard, https://github.com/vectara/hallucination-leaderboard/blob/main/README.md (last visited Jan. 23, 2026).

[83] Bilmes Report ¶17.

[84] *Id*. at ¶297.

these fields. However, I understand from counsel that "transformativeness" is a legal concept relevant to one of the fair use factors in copyright law. Dr. Bilmes appears to be attempting to apply a legal concept to his understanding of LLM development and deployment, for which analysis he has disclosed no relevant expertise.

62. In any case, I disagree with Dr. Bilmes's definition of "transformative" and his opinion that data being modified into a mathematical representation is "transformative." As an electrical and computer engineering professor, Dr. Bilmes should be fully aware that data can be arbitrarily converted into different representations without changing its content or being "transformed" in any way. For example, all data stored in a computer is stored in binary at some point, because data storage in computers does not support ASCII letters as a basic data representation. By Dr. Bilmes's reasoning, any data that is stored in a computer, read from storage, and then printed or displayed in ASCII form is "transformed." A binary version of Harry Potter would also be completely unrecognizable to any human being, but that does not mean the substantive literary content has been changed in any way. That is clearly not a valid test for "transformative."

63. Similarly, there is an exceedingly wide variety of arbitrary ways that data can be converted that would meet Dr. Bilmes's definition of "transformative," which he defines in part as high dimensionality or complexity of representation. A computer file, either storing text or images or other types of data, can be encrypted for security purposes, or compressed (either perfect lossless compression or lossy compression), or converted into a different file format (ASCII to Word or RTF or other format), and those changes would all satisfy Dr. Bilmes's criteria for "transformative." Yet all of them could and would be converted back into the original format, and would be identical to the original input. In addition, Dr. Bilmes seems to imply that lack of interpretation implies a type of transformativeness. Again, any non-trivial encryption algorithm would easily meet this bar, as the encrypted results would make the data completely opaque to anyone without the encryption key. In fact, a simple conversion into a different file format would

make it uninterpretable to anyone who simply did not know what file format the data was converted to.

64. To be clear, nothing about the transformer architecture, not the high dimensional vector data representation nor the opaque data representation, is inherently "transformative." Consider this: it is a simple task to take an LLM and to train it solely on a single piece of text, and then to train it repeatedly until it overfits on that piece of text and reproduced it perfectly regardless of the user prompt. At that point, the trained model is a duplicate of the training text, and will serve the same purpose to reproduce the training text, just like a copy of the training text stored in plaintext. We know this is possible given the extensive literature on LLM overfitting.[85] This will work whether the text is a single book, or a book full of lyrics, or a book full of the same song lyric repeated over and over. A simple way to verify this is to just fine-tune an existing model on the piece of text, which is so easy that it has been written up in blogs for home enthusiasts to try.[86] Once this occurs, it does not matter what the internal representation is inside the LLM, or how many dimensions those vectors represent, or how hard it is for anyone to interpret those numbers. None of those things matter, because that trained LLM will now have a perfect representation of the input text and can be trained to regurgitate a copy to every user prompt regardless of the prompt. To be clear, this copying, memorization, and regurgitation phenomenon is already happening naturally without intentional overfitting. Multiple research studies, some with coauthors employed by LLM model trainers, have shown that entire books can be reproduced verbatim from widely used LLMs today.[87] All these reasons illustrate why simply training data into the LLM transformer architecture does not make it transformative.

---

[85] *See supra* pp. 14–17.

[86] Meir Michanie, *Overfitting Is All You Need*, MEDIUM.COM (June 30, 2024), https://medium.com/@meirgotroot/overfitting-is-all-you-need-174caa7f1064.

[87] *See* Cooper, *supra* n. 10; Ahmed, *supra* n. 11.

65. I also strongly disagree with Dr. Bilmes's opinion that "Anthropic uses Plaintiffs' copyrighted materials for a different purpose than the originally intended use by the copyright holders."[88]

66. First, Dr. Bilmes does not cite or identify any basis for ascertaining the "originally intended use" of Publishers' lyrics. Dr. Bilmes makes no representation that he has spoken to any of the Publishers or other copyright holders or has any particular expertise in the purpose of music lyrics. Further, like "transformativeness," the "originally intended use" of training materials is not a concept or term of art that carries any significance in my field or among my peers. This again appears to be an attempt by Dr. Bilmes to apply a legal concept to the facts of this case.

67. Second, Dr. Bilmes's representation that Anthropic's use for Publishers' copyrighted works is "for the purpose of establishing statistical relationships that are represented within the model itself" [89] is misleading and incomplete. LLMs do not train on materials solely for the purpose of establishing statistical relationships. Developers of general-purpose LLMs build and train LLMs to serve "more generalized purposes,"[90] which include use cases, which Anthropic has promoted, like creative writing.[91]

68. To suggest that Anthropic's sole use of Publishers' lyrics is to examine statistical relationships would be like saying the purpose of farming is to put seeds in the ground, or the purpose of extracting oil from the ground is to put it into barrels. Such arguments only represent, at best, half the picture. The actual purpose of farming is to grow crops that are useful for other purposes, such as feeding people. The actual purpose of oil extraction is to use it as a fuel or perhaps to create petroleum products. The same is true of Anthropic's use of Publishers' lyrics. The purpose of using lyrics in training is not to establish statistical relationships among those words. Establishing statistical relationships is only an intermediary step in LLM development and

[88] Bilmes Report ¶288.
[89] *Id.* ¶300.
[90] *Id.* ¶289.
[91] *See, e.g.,* J. Leibrock 30(b)(6) Dep. Tr. at 48:7–24; J. Leibrock 30(b)(6) Dep. Ex. 22.

deployment, with the ultimate goal of building a model that can be used for a wide variety of purposes, including responding to prompts about music lyrics, reproducing lyrics, creating new lyrics, or answering questions about those lyrics.

69. For similar reasons, Dr. Bilmes's argument that Anthropic's use of Publishers' lyrics is transformative because "the purpose and use of Claude's training data . . . is not to reproduce that source material, but rather, to allow Claude to generate entirely novel results" is not compelling or even correct. While Claude may be engineered to generalize and create "novel results," that does not mean it is not *also* engineered to reproduce portions of materials it was trained on. Even if Claude was not designed *specifically* with a use case in mind, as a general purpose LLM it is meant to handle many tasks, including reproduction of source material when prompted.

70. Additionally, the common development of unanticipated emergent abilities and properties in LLMs further undermines the view that Anthropic has and can anticipate all specific uses and purposes for Publishers' lyrics. In Anthropic's co-founder and CEO's own words, "[m]odels are unpredictable."[92]

71. Third, Dr. Bilmes ignores that the "use" for Publishers' lyrics is determined and informed, at least in part, by what Claude ***users*** prompt Claude to do. A Claude user that seeks copies of Publishers' lyrics is utilizing those lyrics for the same "originally intended use" as its creators or copyright holders—to enjoy the words to a song. From both a technical and logical viewpoint, Anthropic copying Publishers' lyrics into Claude's training corpus, training the model, and reproducing the lyrics as output are all intermediary steps to fulfill the Claude user's prompt/request seeking those lyrics. The "use" is not solely limited to mere training inputs to create a general purpose LLM, but what users use that general purpose LLM for.

72. Dr. Bilmes alleges, based on his analysis of Claude user data on Clio, that Claude seeks lyrics only on an infrequent basis.[93] Aside from the many flaws and errors in Dr. Bilmes's

[92] Dario Amodei Dep. Ex. 246, at -605740.

[93] Bilmes Report ¶¶299–300.

methodology identified in Joshua Dennis's Rebuttal Expert Report,[94] the mere fact that users may use Claude for other purposes other than seeking lyrics does not detract from the fact that Claude users nonetheless still use Claude to reproduce Publishers' lyrics.

F. Dr. Bilmes Mischaracterizes the Import of Robots.txt in Granting Permissions to Crawl Websites with Publishers' Lyrics

73. Dr. Bilmes concludes that Publishers permitted Anthropic's copying of their lyrics from the internet because their lyrics were "publicly available (many times over)" on webpages that "permitted Anthropic's crawler to access them."[95] Dr. Bilmes infers such permission was given where he identified an alleged lack of robots.txt file on webpages blocking crawlers at the time Anthropic's crawler accessed them.[96] Dr. Bilmes's conclusion is incorrect and starkly misleading for several reasons.

74. A robots.txt file is not a license to copy; it is a crawling directive that may be obeyed on a voluntary basis. A robots.txt file is a plain text file placed in the root directory of a website that tells web crawlers about which URLs the crawler may or may not access. The concept was proposed in 1994 by Dutch developer Martijn Koster, at a time when web servers had limited capacity and web crawling often caused server overload. By providing simple instructions to crawlers, robots.txt was designed to help reduce unnecessary requests, preserve server resources, and protect servers from being overloaded. The format and structure of robots.txt are defined by the Robots Exclusion Protocol (REP) which became an official Internet Standard (RFC 09309) in 2022. The use of robots.txt is not required for a website to function properly.

75. In addition, robots.txt was designed three decades before generative AI, for a non-commercial Internet that was primarily used for academic research at the time. It was not designed as a response to internet-scale content scrapers downloading data in order to train commercial AI models.

---

[94] *See* Rebuttal Expert Report of Joshua Dennis (Jan. 23, 2026) at Section III.
[95] Bilmes Report ¶308.
[96] *Id.* ¶¶301–10.

76. It is important to note that robots.txt operates entirely on an honor basis and relies on *voluntary compliance* by crawlers. As documented by Google in its introduction to robots.txt document,[97] the use of robots.txt does not and cannot command or enforce crawler behavior; compliance is entirely voluntary and it is up to the crawler to decide whether to follow the instructions. Therefore, the use of robots.txt cannot prevent scrapers from accessing the content. Furthermore, as a fully voluntary protocol, robots.txt also does not instruct the scrapers on what to do with the content. It is not a mechanism to declare, grant, or enforce permission to access or use any content on the site. Such a lack of legal binding and enforcement is one of the key reasons why robots.txt is not uniformly deployed or recognized as a mechanism for granting permissions or expressing consents.

77. In other words, the mere fact that several webpages containing Publishers' lyrics did not have robots.txt does not mean the lyrics were "permissibly accessed" by Anthropic or its crawler.

78. Setting aside the fact that robots.txt was never designed to function as an indicator of permission for access, I disagree strongly with Dr. Bilmes's view that the lack of an explicit denial of access implies free use. Accessibility on the Internet does not imply that content is free of copyright and can be used for commercial purposes. The downloading and use of copyrighted material against copyright is generally referred to as piracy. Dr. Bilmes's view that anything that is not explicitly forbidden is free to use is akin to assuming that any house with an unlocked door implies its contents are free to take. In both cases, basic laws of ownership and copyright protect content and possessions even when no explicit denial of entry is present.

79. Today, some publishers use robots.txt to tell crawlers to not access their online content, either partially or entirely. Yet there is already concrete evidence showing that AI-focused agents and crawlers frequently disregard robots.txt to gather data from these sites, undermining the use of robots.txt as a deterrent for web crawlers, particularly AI crawlers.

---

[97] *Introduction to robots.txt*, GOOGLE, https://developers.google.com/search/docs/crawling-indexing/robots/intro (last visited Jan. 23, 2026).

80. For example, content licensing company TollBit tracked and analyzed AI traffic to the subject publishers' websites, and found that numerous AI agents from multiple sources (not just one company) are bypassing robots.txt and retrieving content from sites.[98] Similarly, an investigation conducted by Wired.com concluded that AI startup Perplexity is surreptitiously scraping websites while ignoring robots.txt.[99]

81. A study conducted by Duke University in 2025 analyzed traffic on 36 Duke-controlled websites over a period of 40 days, and revealed that AI-bots often do not comply with robots.txt directives that "disallow" access to certain subpages, demonstrating the practical limitations of robots.txt instructions.[100] Moreover, many bots never even checked the robots.txt file or checked it infrequently.[101] This means that any changes made to robots.txt to prevent unwanted scraping would not be noticed or respected by the scraper for significant timeframes (if they are respected at all). This result aligns with another study that used a single website to study compliance to robots.txt.[102]

82. As I previously noted, Anthropic itself has admitted that its crawlers did not always obey robots.txt directives not to crawl.[103] Evidence of Anthropic's compliance with robots.txt generally is also sparse; the only evidence Dr. Bilmes cites in support of his claim that Anthropic's ClaudeBot respects 'do not crawl' instructions comes from witness testimony of one of

---

[98] Katie Paul, *Multiple AI companies bypassing web standard to scrape publisher sites, licensing firm says*, REUTERS (June 21, 2024), https://www.reuters.com/technology/artificial-intelligence/multiple-ai-companies-bypassing-web-standard-scrape-publisher-sites-licensing-2024-06-21/.

[99] Dhruv Mehrotra & Tim Marchman, *Perplexity Is a Bullshit Machine*, WIRED (Jun. 19, 2024), https://www.wired.com/story/perplexity-is-a-bullshit-machine/.

[100] Taein Kim et al., S*crapers Selectively Respect robots.txt Directives: Evidence From a Large-Scale Empirical Study*, IMC'25: PROC. OF THE 2025 ACM INT. MEASUREMENT CONF. 541 (Oct. 15, 2025), https://doi.org/10.1145/3730567.3764471.

[101] *Id.*

[102] Enze Liu et al., *Somesite I Used to Crawl: Awareness, Agency and Efficacy in Protecting Content Creators from AI Crawlers*, IMC'25: PROC. OF THE 2025 ACM INT. MEASUREMENT CONF. 78 (Oct. 15, 2025), https://doi.org/10.1145/3730567.3732913.

[103] *See* Zhao Report ¶59.

Anthropic's co-founders and its public-facing website that stops short of committing to entirely refrain or crawl a given website.[104]

83. Second, it is important to note that deploying and maintaining a robots.txt file is a nontrivial undertaking, as it requires technical expertise and ongoing effort. Web hosts must have extensive knowledge of the user agents and entities they wish to restrict from scraping (and must continually update this knowledge as agents and entities change and develop), placing a significant burden on them to keep pace with an ever-changing web ecosystem. Moreover, because the robots.txt file resides in the root directory of a website, only the website administrator has the ability to access and manage it. These complexities make it extremely challenging for individual authors, creators or content owners to use robots.txt. These challenges are further confirmed by a recent survey of 203 professional artists—ostensible copyright holders—regarding their use of robots.txt, in which 59% of those artists had never heard of robots.txt.[105] And of those who knew about robots.txt, such knowledge may be useless as many artists may not have control over their works simply by virtue of not hosting and operating their own website. For example, many artists post their work across multiple web hosting platforms such as wix.com or square.com, where they are not the website owners and thus cannot configure robots.txt files.

84. In sum, Dr. Bilmes mischaracterizes robots.txt by suggesting that a failure to implement it grants permission to crawl data and use such data in training AI models, or that it is a failsafe tool for preventing data crawling and scraping. Instead, robots.txt is not an effective way to deter crawl data, nor is it readily accessible to and configured by individual content creators who post their work on multiple online platforms.

[104] Bilmes Report ¶110 n. 200.
[105] Liu, *supra* n. 102.

G. Dr. Bilmes Underestimates the Extent and Impact of Anthropic's Copying of Publishers' Lyrics Multiple Times Despite its Deduplication and Filtering Efforts

85. Dr. Bilmes claims that Publishers' song lyrics only represent an "immeasurably small relative fraction of Claude's total training corpus," estimating the lyrics of all 499 works constitutes 213,891 tokens of the dataset used for Claude 3.[106] He concludes "[a]t this magnitude, the effect of Plaintiffs' lyrics on Claude's learned parameters or outputs is extremely likely to be near zero, if present at all."[107] I strongly disagree.

1. Many Copies of Publishers' Works in Suit Reside in Claude's Training Data

86. Anthropic has copied the lyrics to Publishers' Works in Suit many times over. According to Dr. Bilmes, Anthropic crawled Publishers' Works in Suit between [redacted] times.[108] The result of this repeated crawling is the vast multiplication of Publishers' lyrics in Claude's training data.[109] Dr. Shawn Shan's analysis identified [redacted] matches to lyrics of Publishers' Works in Suit in Claude's training data, with each Work in Suit being copied by Anthropic no less than 10 times each.[110]

87. Dr. Bilmes attempts to excuse the persistence of Publishers' lyrics in Anthropic's Claude training datasets by arguing that the lyrics exist in other training datasets curated by non-Anthropic parties. In "confirming" his conclusions about "the pervasiveness of Plaintiffs' lyrics," Dr. Bilmes analyzes the Common Pile dataset, which he represents was curated by academic researchers who attempted to use only "openly licensed and public domain text.[111] Dr. Bilmes argues that lyrics appear frequently in this dataset, despite the intent of its creators to include only

[106] Bilmes Report ¶328.
[107] *Id.*
[108] Bilmes Report ¶309.
[109] Shan Report ¶49, Ex. D.
[110] *Id.* ¶40.
[111] Bilmes Report ¶315.

licensed text, and therefore it would be similarly difficult for Anthropic to remove lyrics from its dataset.[112] Dr. Bilmes's reference to the Common Pile is unavailing.

88. First, the Common Pile is a research dataset, not a legal benchmark. It makes no legal claims and provides no guarantee that its contents are lawfully usable. The Common Pile creators explicitly acknowledge licensing errors, license laundering, copyrighted excerpts inside public-domain documents, and license changes over time.[113] In other words, they do not appear to be trying to remove all instances of copyrighted material from the dataset, so its usefulness as a benchmark by which commercial LLMs should be judged is negligible.

89. More importantly, Dr. Bilmes's conclusions about Common Pile are not supported by the data. Dr. Bilmes examined the amount of a given work (by percentage of the total lyrics or number of words) that was found in the Common Pile.[114] This only tells us how much of a work appears when it does. That is not a measure of prevalence.

90. By contrast, Dr. Shan examined the true prevalence of Publishers' Works In Suit in the Claude training data and determined that every work appears in full, multiple times. Some works appear hundreds or thousands of times.[115] To the extent Dr. Bilmes's analysis can be used to assess prevalence in the Common Pile, it appears that only a small portion of most of the Works in Suit appeared in the dataset and 55 did not appear at all. If anything, a comparison of the Claude dataset to the Common Pile demonstrates that substantial elimination of copyrighted works can be achieved if desired.

91. Dr. Bilmes's attempts to excuse the presence of lyrics in Claude's training data by arguing they are "ubiquitous across the internet"[116] also falls flat. This argument misunderstands the mechanics of LLMs: ubiquity is the primary driver of memorization. As I discussed above,

---

[112] *Id.* ¶313.

[113] Nikhil Kandpal et al., *The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text*, ARXIV:2506.05209v1 (Jun. 5, 2025), https://arxiv.org/pdf/2506.05209.

[114] Bilmes Report ¶¶316-17.

[115] Shan Report ¶49.

[116] Bilmes Report ¶313.

research shows that data repeated many times in a training set is significantly more likely to be memorized and regurgitated verbatim.[117] By admitting lyrics are everywhere, Dr. Bilmes effectively concedes the data environment was highly prone to reproducing those lyrics. Knowing this, effective model design aimed at reducing memorization would demand *more* aggressive filtering for such content, not less. The fact that lyrics are "everywhere" is precisely why they require specific, targeted exclusion protocols which Anthropic failed to implement.

92. Dr. Bilmes states that because filtering is "time- and resource-intensive," Anthropic "only filters on samples of the data" for certain categories.[118] By choosing to filter only a sample rather than the full corpus, Anthropic statistically guarantees that the vast majority of "ubiquitous" content (like the Publishers' lyrics) would remain in the training set, and that the model trained on that data would memorize that content and regurgitate it when prompted. A safety system designed to check only a fraction of the input is inherently designed to fail at scale. It is perhaps because of these inadequate practices that Anthropic concedes that its deduplication efforts are imperfect and that "the training corpus still contains . . . low quality data" even after filtering.[119]

93. Dr. Bilmes also claims it is "technologically infeasible at scale" to remove every near-duplicate instance of data.[120] But this argument conflates commercial expense with technical feasibility. Tech giants routinely perform massive-scale fuzzy deduplication effectively[121] (*e.g.*, Google Search indexing). If Anthropic has the computational resources to *train* a model on trillions of tokens—a process requiring massive capital and hardware—it has the resources to perform robust fuzzy deduplication (such as MinHash or SimHash) on that same data. Dr. Bilmes treats "scale" as an immutable constant that excuses errors, rather than a parameter Anthropic voluntarily

---

[117] See *supra* n.20 and accompanying text.

[118] Bilmes Report ¶311.

[119] Anthropic's Am. Resp. & Obj. to Interrogatory 9 (Nov. 12, 2025).

[120] Bilmes Report ¶312.

[121] Saniya Sudhakaran & Meera Treesa Mathews, *A Survey on Data Deduplication in Large Scale Data*, INT'L. JOURNAL OF COMPUTER APPLICATIONS Vol. 165, No. 1, 1–4 (May 2017), https://www.ijcaonline.org/archives/volume165/number1/27534-2017913696.

chose. If the scale of the data were too large to safely curate for copyright infringement, the decision to proceed at that scale was a calculated risk to prioritize product capability over legal safety, not a result of technological inability.

94. This assertion is also belied by Dr. Bilmes' own description of Anthropic's successful efforts to filter "harmful" content and deduplicate its dataset.[122] Indeed, identifying harmful content—which often requires parsing nuance and intent—is significantly more complex than matching specific copyrighted strings and symbols, yet Anthropic has prioritized solving the former while claiming impossibility for the latter. The fact that Anthropic sanitizes its dataset for safety but not for intellectual property rights demonstrates that the failure to exclude copyrighted material is a choice of prioritization, not a limitation of technology.

95. I also question whether Anthropic's alleged "standard practices" would have removed copies of Publishers' lyrics in the first place. Dr. Bilmes argues that Anthropic used "standard techniques" to limit the impact of Publishers' lyrics on its training datasets, first by filtering out "low-quality" or "junk" data and then deduplicating the data so no single piece of data exerts an "outsized influence" on the model.[123] But Dr. Bilmes does not explain whether Anthropic *actually* filtered Publishers' lyrics out as "low-quality" or "junk" data, or whether Anthropic considered the lyrics to the Works in Suit to actually be "low-quality" or "junk." I have not seen any evidence demonstrating that Anthropic's efforts to filter Publishers' lyrics as "junk" data would have been successful or meaningful.

96. Dr. Bilmes states that Anthropic uses classifiers to filter out "low-quality" or "junk" data to improve model performance, listing boilerplate HTML and spam as examples.[124] This reveals a critical contradiction: if lyrics are indeed "low-quality" data that degrade model performance, Anthropic's filtering systems—designed specifically to remove such data—failed to catch them. The fact that millions of lyric tokens remain in the dataset suggests that Anthropic

[122] Bilmes Report ¶¶114–15.
[123] *Id.* ¶¶311–12.
[124] *Id.* ¶311.

either does not actually view lyrics as "junk" (and thus intentionally retained them) or that their filtering systems are fundamentally ineffective at their stated purpose. Anthropic cannot simultaneously claim their filters are robust enough to clean the data while excusing the presence of "low-quality" lyrics as an inevitable accident.

97. Dr. Bilmes' dismissal of lyrics as low quality because they lack factual density is also curious in this context. Dr. Bilmes emphasizes in his Report that the purpose of training data is not for the LLM to memorize that content but simply to learn semantic and syntactic structures and relationships.[125] If that is so, the factual density of lyrics should not affect their utility as a source of training data. Indeed, their unique structure should be valuable and that value is reflected in the weighting of lyrics in the model, as discussed below.

98. Dr. Bilmes' discussion of the value of lyrics also fails to align with his overall conclusion that the value of training data is relative, not absolute.[126] In other words, whether a dataset is valuable for training depends not just on intrinsic qualifies of that dataset but also on what other data the model's training datamix includes. Even high-quality data is not valuable if that data already exists in the dataset. Conversely, even lower-quality data can be valuable if it is unique.

99. Dr. Bilmes concludes that "it would be technologically difficult to remove all instances of the lyrics from the training data."[127] But as I explained in my initial report, Anthropic is more than capable of identifying copies of music lyrics in its training corpora—if it chose to—using relatively simple methods.[128] I have not seen any evidence that Anthropic has ever employed any of these methods or intended to do so.

100. Further, Dr. Bilmes arrives at this conclusion based on perceived difficulties associated with Anthropic's *deduplication* of data, not removal of Publishers' lyrics *specifically*.

---

[125] *Id.* ¶21.
[126] *Id.* ¶323.
[127] *Id.* ¶313.
[128] Zhao Report ¶153.

But Dr. Bilmes does not present any evidence showing that Anthropic intended to exclude Publishers' lyrics from its Claude training corpora, rather than merely deduplicating copies of their lyrics to improve LLM performance. In other words, Anthropic seemingly has never tried to exclude Publishers' lyrics in the first place.

101. In sum, it is highly likely that Publishers' lyrics make up a larger portion of the overall dataset than Dr. Bilmes estimates. And, as I have explained previously and below, a relatively small dataset of unique text and phrases, such as Publishers' lyrics, can have a significantly high impact on LLM performance and memorization.

H. Dr. Bilmes Incorrectly Concludes Publishers' Lyrics Are Immaterial to LLM Performance

102. Dr. Bilmes argues that Publishers' lyrics have comparatively "less value for LLM training datasets" because they are "small, narrow in scope, heavily repetitive, redundant, or low in factual content."[129] While he concedes that "small and narrow-in-scope data can be valuable if it provides unique information," he opines that the "value of a chunk of data to LLM training is always relative to the much larger dataset the chunk is contained within."[130] This is incorrect because a relatively small dataset can have a significant impact on overall LLM performance, particularly if it contains highly unique texts and phrases. Such datasets are also likely to induce more memorization and regurgitation from models trained on them. As I have explained above, it is not corpus percentage that drives memorization (and regurgitation), but rather repetition in the training corpora. Even if Publishers' lyrics only constitute a tiny fraction of Claude's training corpora, the risk of memorization is nonetheless high because of their uniqueness and their repeated presence.

103. Dr. Bilmes' opinion rests on the quantitative fallacy that because the lyrics are a small fraction of the total tokens, they are "immaterial."[131] This ignores the fundamental reality

[129] Bilmes Report ¶323.
[130] *Id.*
[131] *Id.* ¶¶20, 372.

that model learning is driven by **data uniqueness and weighting**, not just raw volume. Unlike generic web text, which is often fungible, song lyrics constitute a unique, non-substitutable data class possessing distinct linguistic properties—such as rhyme, meter, and metaphor—that cannot be learned from other sources.[132] Dr. Shan's inspection of Claude's training data revealed that Anthropic explicitly values this distinctiveness by assigning English text (the category containing these lyrics) a training weight of [redacted].[133] Research consistently demonstrates that models are highly sensitive to such unique, high-quality data signals, which can disproportionately steer model capabilities even in small quantities.[134] By combining this [redacted] with the lyrics' inherent uniqueness and massive repetition (over [redacted] copies), Anthropic ensures these works exert an outsized influence on the model that far exceeds their statistical footprint.

104. Dr. Bilmes also claims Publishers' lyrics are not a high-quality training dataset because "Plaintiffs' lyrics (and lyrics in general) while stylistically rich, are short, repetitive, idiosyncratic, relatively homogenous, and highly constrained in topical breadth."[135] But these are words taken from co-founder Jared Kaplan's deposition testimony—Anthropic's own words.[136] Revealingly, Dr. Bilmes cites no other basis for this blanket assertion. Further, as an expert in the field, I am not aware of any research supporting this conclusion. Regardless, it is precisely *because* Publishers' lyrics are "stylistically rich" and "idiosyncratic" that they are valuable as training data. Also, unlike poems, lyrics are constructed by taking account of rich musical information: melody, rhythm, instrumentation, the quality of the singer's voice, other qualities of the recording, [137] as

---

[132] *See* Kaplan Decl. ¶35 (Aug. 22, 2024), ECF No. 209.

[133] Shan Report ¶40.

[134] *See* Aladin Djuhera et al., *Fixing It in Post: A Comparative Study of LLM Post-Training Data Quality and Model Performance*, 39TH ANN. CONF. ON NEURAL INF. PROC. SYS. (NEURIPS) (Dec. 15, 2025), https://arxiv.org/pdf/2506.06522.

[135] Bilmes Report ¶324.

[136] *Id.* ¶324 n.599.

[137] Matthew Zapruder, *The Difference Between Poetry and Song Lyrics*, BOSTON REV. (Dec. 6, 2012), https://www.bostonreview.net/forum/poetry-brink/difference-between-poetry-and-song-lyrics/.

explained below. Thus they belong to a diverse, heterogeneous domain that is different from poems and thus cannot be covered by poetry data.

105. Dr. Bilmes also misinterprets the findings from the Longpre et al. paper (2023),[138] which he represents found that "removing heterogenous datasets from the pre-training corpus had a greater impact on model performance than removing more narrow specialized domains."[139] Dr. Bilmes confuses the results of two QA benchmarks (MRQA and UnifiedQA) as the full measure of LLM performance, broadly. The Longpre researchers did not examine the full extent of LLM performance using a comprehensive set of benchmarks in their evaluations; they only used MRQA and UnifiedQA.[140] Additionally, the Longpre authors actually concluded that, "[t]he best performing models use *all* the pretraining data sources" and that the removal of targeted source domains, such as Pile's Code or Academic domains, may not have as significant an impact on performance because they were "not well matched with the QA evaluation sets."[141]

106. The fact that Publishers' lyrics might make up a relatively small portion of the total training corpus does not mean those lyrics have a proportionally small impact on Claude's training processes, outputs, and capabilities. As I've previously explained, memorization is driven by repetition in the training data and the uniqueness of the text. Given the uniqueness of Publishers' lyrics (*e.g.*, "Drove my Chevy to the levee, but the levee was dry"), it therefore follows that Claude is more likely to memorize Publishers' lyrics than other types of texts.

107. Anthropic's own employees admit that Claude is prone to memorization,[142] that memorization occurs particularly when a model is trained on the same piece of text multiple

---

[138] Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*, arXiv:2305.13169 (Nov. 13, 2023), https://arxiv.org/pdf/2305.13169.

[139] Bilmes Report ¶321.

[140] Longpre, *supra* n. 138.

[141] *Id*.

[142] *See, e.g.,* D. Drain Dep. Ex. 140 (Anthropic_0000507796) ("Large LMs memorize a LOT"); Anthropic_0000458427 (July 2021 memo discussing whether "we can reduce data recitation / solve the copyright problem internal to the model via RL").

times,[143] and that it is desirable to memorize culturally important texts.[144] These admissions contradict Dr. Bilmes' view that Publishers' lyrics were not memorized by Claude merely because they only represent a small fraction of training data. Rather, given Claude's LLM architecture and the existence of multiple copies of Publishers' Works in Suit in Claude's training data, the opposite is more likely to be true: Claude is more likely to memorize unique lyrics that may represent a smaller fraction of Claude's training data.

108. Second, Dr. Bilmes' conclusion is incorrect because, while Publishers' lyrics might not be material to certain performance markers, they are material to certain use cases, such as regurgitating lyrics, creating new lyrics or answering questions about *specific* lyrics.

109. Dr. Bilmes claims Claude could generate poetry not by training on poetry or lyrics, but by training on data that provide high-level "insights" on poetic verse structure. However, Dr. Bilmes only cites testimony from one of Anthropic's witnesses in support of this assertion.[145] I strongly disagree. It would be near-impossible to generate lyric-like outputs if Claude did not train on lyric-like inputs. High level "insights" would not be sufficient because lyrics are different from poems in terms of how words are constructed. As American poet Matthew Zapruder writes,

> "Words in a poem take place against the context of silence . . . whereas . . . lyrics take place in the context of a lot of deliberate musical information: melody, rhythm, instrumentation, the quality of the singer's voice, other qualities of the recording, etc. Without all that musical information, lyrics usually do not function as well, precisely because they were intentionally designed that way. The ways the conditions of that environment affect the construction of the words (refrain, repetition, the ways information that can be communicated musically must be communicated in other ways in a poem, etc.) is where we can begin to locate the main differences between poetry and lyrics."[146]

---

[143] D. Drain Dep. Tr. 90:2–24.

[144] *Id.* 75:3–76:15.

[145] Bilmes Report ¶326.

[146] Zapruder, *supra* n. 137. Matthew Zapruder is an American poet, editor, translator, and professor. He has won many awards, including the 2007 William Carlos Williams Award from the Poetry Society of America, and his poetry collection *The Pajamais* was chosen by Library Journal as one of the top ten poetry volumes of 2006. Matthew Zapruder, MATTHEW ZAPRUDER, https://matthewzapruder.com/about/ (last visited Jan. 23, 2026).

110. Furthermore, the fact that Claude specifically reproduces Publishers' exact lyrics in the same exact sequence would only be possible because the lyrics in the training data formed the foundation for which Claude was able to learn from and output those same lyrics. This is because Claude is not merely meant to "understand . . . concepts" about song lyrics generally.[147] It was also engineered to regurgitate and replicate those same lyrics it trained on.

111. Dr. Bilmes concludes that "Claude's ability to generate new lyric-like text likely has little (if anything at all) to do with whether Plaintiffs' lyrics (or any lyrics) were included within its training data and instead exemplifies its capacity for abstract pattern synthesis."[148] However, this is highly speculative because of limited LLM interpretability and Anthropic's inability to confirm how its models actually function or what they are capable of.[149] Neither Dr. Bilmes nor Anthropic offers any evidence that confirms whether Claude is generating lyric-like text because of training data *other* than Publishers' lyrics. The only surefire way to test this hypothesis would be to train Claude without Publishers' lyrics, prompt Claude for those same lyrics, and compare the outputs. But Dr. Bilmes makes no claim that he or Anthropic have done this.

112. More importantly, the nature of autoregressive transformers is that they pattern-match against training data by counting the frequency (and therefore probability) that a particular sequence of tokens follows other tokens. However, the number of possible sequences of words in any application of the English language is still very, very large. Features common to lyrics, including structure, phrasing, tempo, and rhyme are all constraints that an LLM would not recognize without training on song lyrics. The potential space of all combinations of words is unimaginably large, and the likelihood of an LLM naturally producing lyric-like text without seeing song lyrics is infinitesimally small. It is also impossible for LLMs to produce lyrics with

[147] Bilmes Report ¶330.
[148] *Id.* ¶338.
[149] *See, e.g.,* Dario Amodei Dep. Ex. 246, at Anthropic_0000605740 [redacted]

specific words or phrasing that is only present in specific song lyrics. For example, a user may ask for lyric-like text about a boy moving from Philadelphia to Southern California. But without training on lyrics from Fresh Prince of Bel-Air, an LLM would not generate the words "maxin'" or the phrases "Nah, forget it, yo Holmes, to Bel-Air!" LLMs do not create patterns or tokens they have never seen. A word that does not exist in training data would not match any existing tokens in a trained LLM, and cannot be naturally generated.

113. Dr. Bilmes also suggests "[c]ompared to other sources of text within Claude's training corpora, lyrics, in general and including Plaintiffs' lyrics, do not constitute a 'high-quality' dataset specifically for general-purpose LLM training. Song lyrics are closer to poetic verse than the information-dense material that drives performance improvements in large-scale language models."[150] First, Dr. Bilmes does not offer firm bases for these representations; he merely cites to testimony provided by Anthropic's witnesses—which concede Anthropic never tested the hypothesis that song lyrics are less useful for training purposes than other long-form texts[151]— and does not identify any other evidence in support of this theory.

114. Second, Dr. Bilmes's (and Anthropic's) suggestion that lyrics are not "high-quality" datasets because they are closer to poetic verse than other so-called "information-dense material" is highly questionable. Even from a linguistic and literary perspective, song lyrics and poems perform different functions and possess different and unique characteristics. In other words, lyrics constitute heterogeneous and diverse texts distinct from poetry and do not constitute "low-quality" data on that basis alone.

115. Ultimately, the impact of Publishers' lyrics on Claude's performance cannot be reliably assessed unless Anthropic were to remove Publishers' lyrics from Claude's training data and measure Claude's performance before and after the removal. But Anthropic has conducted no

[150] *Id.* ¶324.

[151] *Id.* ¶324 nn. 598-99, ¶326 n. 602.

such tests to determine whether Publishers' lyrics actually constitute "low-quality" data on the basis it believes song lyrics are "low-quality" data.[152] This is circular and flawed reasoning.

I. Dr. Bilmes Ignores Fundamental Flaws in Claude's Guardrails

116. Guardrails in Anthropic's LLM system are separate software components that work independently from the Claude LLM model itself. They include user-facing input filters that analyze and filter user prompts for specific topics such as toxic content or copyrighted materials, and post-inference filters that analyze Claude's output for any undesirable content, including output that violates copyright. As add-on components to the LLM, their creation and use by Anthropic is evidence that Anthropic is aware that the core LLM is fundamentally prone to producing content that can be toxic, unsafe, or violate copyright. This behavior is precisely what Anthropic's "constitutional AI" was designed to prevent. If constitutional AI and similar "alignment training" were successful, these input and output filters would be unnecessary and redundant. Anthropic's reliance on these tools is evidence that the LLM itself is inherently and fundamentally flawed, and filters are necessary as band aids to patch its imperfections. Moreover, if either the input filter or the output filter were fully effective, it would render the other unnecessary and redundant. The fact that both filters are active is further evidence that neither filter is correctly removing toxic or copyrighted content from the query or output.

117. Dr. Bilmes seems to suggest that it was not reasonable for Anthropic to be expected to include guardrails before 2023 by opining "AI guardrails are a very new and active area of AI research."[153] However, memorization and regurgitation by LLMs was a well-known issue before 2023.[154] Further, the earliest version of Anthropic's Content Compare System (CCS), a guardrail

---

[152] J. Kaplan Dep. Tr. 153:22–154:9.

[153] Bilmes Report ¶210.

[154] Brian Tarran, *Why Large Language Models Should Come with a Content Warning*, REAL WORLD DATA SCIENCE (Nov. 23, 2022), https://realworlddatascience.net/the-pulse/editors-blog/posts/2022/11/23/LLM-content-warning.html; Ziwei Ji et al., *Survey of Hallucination in Natural Language Generation,* 55 ACM COMPUTING SURVEYS (Mar. 3, 2023), https://doi.org/10.1145/3571730; Lance Eliot, *AI Ethics Lucidly Questioning This Whole*

allegedly designed to prevent the output of copyrighted works, was created and launched in [redacted].[155] I am not aware of any evidence that Publishers' lyrics were included in the hash corpus of this early version of CCS.

118. Tellingly, Dr. Bilmes does not offer any opinion as to why or how Claude's model regurgitated verbatim or near-verbatim copies of Publishers' works in its outputs, particularly before the Publishers filed suit but after guardrails meant to prevent copyrighted output were active.

119. These inherent vulnerabilities are also separate and in addition to the glaring issues that "third-party clients that Anthropic does not control may also configure their products to ignore or impair the effectiveness of Anthropic's guardrails."[156] In other words, Anthropic's guardrails may be removed by Anthropic's clients, through which users may access Claude without Anthropic's guardrails. This is yet another vulnerability that throws into doubt the true efficacy of Anthropic's guardrails in preventing the output of copyrighted content.

J. Dr. Bilmes' Analysis of Anthropic's Guardrails is Flawed and Unreliable

120. Dr. Bilmes alleges to have evaluated Claude's copyright guardrails by conducting tests on four Claude models made available through Claude's API: Claude 4.1 Opus, Claude 4 Opus, Claude 3.5 Haiku, and Claude 3 Haiku.[157] Based on the results of those tests, Dr. Bilmes claims his analysis "firmly establishes that Anthropic's multilayered defenses against copyright reproduction are effective"[158] and opines "Claude is not designed to and does not reliably reproduce Plaintiffs' lyrics, even when directly prompted."[159]

Hallucinating AI Popularized Trend That Has Got to Stop, FORBES (Aug. 24, 2022), https://www.forbes.com/sites/lanceeliot/2022/08/24/ai-ethics-lucidly-questioning-this-whole-hallucinating-ai-popularized-trend-that-has-got-to-stop/.

[155] Anthropic's Am. and Supp. Resps. and Objs. to Pls.' First Set of Interrogatories (Nos. 9, 11, 12), dated Nov. 12, 2025, ("Nov. 12, 2025 Interrogatory Responses") at 13.

[156] Bilmes Report ¶279.

[157] *Id.* ¶342.

[158] *Id.* ¶350.

[159] *Id.* ¶341.

121. However, Dr. Bilmes's tests of Claude's copyright guardrails are deeply flawed, unreliable, and at best, reveal a very limited picture of Claude's ability to prevent the output of copyrighted content, including Publishers' lyrics.

122. First and most notably, Dr. Bilmes did not conduct any tests of models prior to Claude 3. Despite identifying Claude 1 and Claude 2 models as models all "at issue," Dr. Bilmes did not test Claude 2 for his guardrails analysis.[160] In fact, it appears all of the models Dr. Bilmes *did* test for his guardrails analysis were Claude models developed or released *after* Publishers filed suit in October 2023, and after Anthropic had developed and deployed bespoke filters to prevent the output of Publishers' Works in Suit *specifically*.[161]

123. Second, Dr. Bilmes limits his analysis to just the 499 Works in Suit, again for which Anthropic made a bespoke solution to prevent their output in Claude responses after Publishers filed suit. Dr. Bilmes's test therefore does not account for Claude's ability to prevent the output of other copyrighted content, such as the rest of the catalog owned by Publishers or other music publishers.

124. Based on the above and Dr. Shan's tests and analysis of Claude's guardrails in his rebuttal report,[162] I conclude that to the extent Dr. Bilmes's guardrails testing shows anything, it shows: (1) guardrails are necessary to prevent infringement because without them the model alone will memorize/regurgitate; (2) when Anthropic wants the model to stop producing infringing outputs, it knows how to do so; (3) the guardrails are an after-the-fact ad hoc "solution" adopted specifically in response to this litigation; (4) Anthropic's failure to include song lyrics in the initial CCS corpus, the limited number of song lyrics included in the revised CCS corpus, plus Anthropic's failure [redacted] (even though many popular songs have been released since then) shows that Anthropic still refuses to take this problem seriously.

---

[160] *Id.* ¶390.

[161] Nov. 12, 2025 Interrogatory Responses at 13–14; Bilmes Report ¶391.

[162] Shan Rebuttal Report ¶¶34–43.

III. Rebuttal to Dr. Altman's Opinions

A. Dr. Altman Lacks Expertise Concerning Generative AI

125. In his initial report, Dr. Altman offers opinions in his initial report about the "public benefits and beneficial uses associated with Large Language Models (LLMs)" and "[Claude] is being used across a variety of domains."[163] Specifically, he makes numerous statements about the benefits of LLMs when applied to different disciplines and areas of study, as well as the general field of education. In each, he brings forth individual examples of AI applications and uses them as demonstrative proof that LLMs are significantly beneficial to society at large. This is, however, at odds with Anthropic's own statements and documents which warn against the potential and real dangers and risks of AI and LLMs.[164]

126. Dr. Altman is a trained medical professional, and has published extensively in medical and biology related journals. But he does not appear to be trained in computer science and AI or have any appointments in Computer Science except by courtesy. He has published infrequently in recent years on AI related topics, and nearly all of those projects are in collaborative efforts, and on deep neural networks, not on the generative models such as LLMs about which he opines in his report.

127. This is significant because deep neural networks (DNNs) are a fundamentally different architecture of AI models. DNNs became popular after they were shown to be almost as good as humans in the ImageNet challenge (September 2014), a benchmark that involved identifying the subjects in a large dataset of images. DNNs are widely regarded as the previous generation of AI models that predated generative AI models. Instead of generating content as LLMs do, DNNs take content as input, and output labels that classify the content. As such, they are referred to as discriminative (as opposed to generative) models.

---

[163] Altman Report ¶2.

[164] *See* Dario Amodei Dep. Exs. 250, 251.

128. The relevance of unrelated and disparate benefits of LLMs, broadly, or how Claude is being used outside of the context of this litigation, is also unclear to me.

129. Nonetheless, I provide my assessment of and opinions of certain aspects of Dr. Altman's opinions below.

B. Dr. Altman Ignores Limitations and Dangers of LLMs

1. Dr. Altman Does Not Address LLM Propensity to Hallucinate and Commit Errors

130. Dr. Altman is correct in his statement that LLMs are AI systems that "learn to predict text by analyzing language patterns across large sets of documents." His example demonstrates that indeed, in a medical context, an LLM is effectively predicting potential diagnoses and medical statements using the probability distribution learned from training data. LLMs are autoregressive transformers that extract probabilities between words and contextual cues, in order to predict words, phrases and larger pieces of text. However, this is fundamentally at odds with most scientific pursuits, which rely on application of the basic scientific method, validating theories with rigorous experimentation and data analysis. While some scientific applications do involve a factor of randomness, *e.g.*, searching for new biological compounds, most scientific applications require deterministic and reliable computation to formulate and validate scientific theories. Predictive models that make mathematical errors and hallucinate facts are not conducive to factual scientific discovery.

131. Dr. Altman cites his work on FDA bioequivalence, where his work applies LLMs to the task of bioequivalence evaluation, using LLMs to "read and summarize these reports and to check for completeness and consistency." He later again cites LLMs use to summarize data in medical histories as a way to save time.[165] While LLMs analyze large volumes of text faster than humans can, they are also known to make biased mistakes, as well as hallucinate non-existent statements that do not follow the text input. A recent study peer reviewed and published in the

[165] Altman Report ¶44.

Royal Society of Open Science found that "when summarizing scientific texts, LLMs may omit details that limit the scope of research conclusions, leading to generalizations of results broader than warranted by the original study."[166] The authors directly compared LLMs, including Anthropic's Claude, against human beings when summarizing 4900 scientific texts, and found that LLMs "were nearly five times more likely to contain broad generalizations" with strong statistical confidence ($p < 0.001$).[167] This and similar studies published in well regarded journals (including Proc. of National Academy of Sciences) have led to open questions of whether or how LLMs should be used in pursuit of science.[168]

132. Numerous studies in medical journals have captured and analyzed the prevalence of textual misinformation and hallucinations in applying LLMs to medical text, including

[166] Uwe Peters & Benjamin Chin-Yee, *Generalization Bias in Large Language Model Summarization of Scientific Research*, 12 ROYAL SOCIETY OF OPEN SCIENCE (Apr. 30, 2025), https://royalsocietypublishing.org/rsos/article/12/4/241776/235656/Generalization-bias-in-large-language-model.

[167] *Id.*

[168] *See, e.g.,* Marcel Binz et al., *How Should the Advancement of Large Language Models Affect the Practice of Science?*, 122(5) PROC. NAT'L ACAD. SCI. USA, Jan. 27, 2025, https://www.pnas.org/doi/10.1073/pnas.2401227121; Bradley D. Menz et al., *Current Safeguards, Risk Mitigation, and Transparency Measures of Large Language Models Against the Generation of Health Disinformation: Repeated Cross Sectional Analysis*, 384 BMJ (Mar. 20, 2024), https://www.bmj.com/content/384/bmj-2023-078538; Catherine A. Gao et al., *Comparing Scientific Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers*, 6(1):75 NPJ DIGIT. MED. (Apr. 26, 2023), https://pubmed.ncbi.nlm.nih.gov/37100871/; Jan Homolak, *Opportunities and Risks of ChatGPT in Medicine, Science, and Academic Publishing: A Modern Promethean Dilemma*, 64 CROAT. MED. J. no. 1 (Feb. 2023), https://pmc.ncbi.nlm.nih.gov/articles/PMC10028563/.

hallucinations in voice transcription.[169] Others have already accepted LLM hallucinations in medicine as inevitable, and are building and evaluating systems to detect these hallucinations.[170]

2. Dr. Altman Fails to Address the Many Harms the Use of LLMs Inflict in Educational Settings

133. Dr. Altman asserts that LLMs are a positive force transforming education, with examples in his classes where students use LLMs responsibly. He argues that "when used responsibly, LLMs have the potential to close gaps in access to high-quality education and make learning more personalized and equitable."[171]

134. Numerous research studies across a wide range of educational environments dispute Dr. Altman's account, instead showing dramatic disruption across all educational levels, leading to "existential risk" to educational systems. First, numerous studies assert that even when properly used, the speed at which LLMs provide information often comes at a cost of lower understanding. Controlled user studies involving LLMs show that students using LLMs experience lower cognitive load, and students using LLMs showed weaker reasoning and shallower understanding.[172] Similarly, other controlled studies show that the more workers trust generative

---

[169] *See, e.g.,* Ananya, *AI Transcription Tools 'Hallucinate,' Too*, SCIENCE (Apr. 26, 2024), https://www.science.org/content/article/ai-transcription-tools-hallucinate-too; Hussam Alkaissi & Samy I. McFarlane, *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing*, CUREUS (Feb. 19, 2023), https://www.cureus.com/articles/138667-artificial-hallucinations-in-chatgpt-implications-in-scientific-writing/; Yubin Kim et al., *Medical Hallucination in Foundation Models and Their Impact on Healthcare*, MEDR XIV (Nov. 2, 2025), https://doi.org/10.1101/2025.02.28.25323115.

[170] Suhas BN et al., *Fact-Controlled Diagnosis of Hallucinations in Medical Text Summarization*, PROC. OF INTERSPEECH 2025 (Aug. 2025), https://www.isca-archive.org/interspeech_2025/bn25_interspeech.pdf.

[171] Altman Report ¶42.

[172] *See* Matthias Stadler, Maria Bannert & Michael Sailer, *Cognitive Ease at a Cost: LLMs Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry*, COMPUTERS IN HUMAN BEHAVIOR (Nov. 2024), https://www.sciencedirect.com/science/article/pii/S0747563224002541.

AI, the less there is critical thinking involved by the workers.[173] Similar arguments have been put forward by faculty on how AI blunts students' ability to think.[174]

135. Perhaps even more impactful than shallow learning, the adoption of LLMs is creating a boom in AI-based plagiarism across all levels of education. Researchers have long known that LLMs will make plagiarism easier than ever for students.[175] But the actual pervasive spread of LLM-based cheating has shocked everyone, leading some to call LLMs a "plague upon education."[176] By 2025, the effects of LLM across schools and universities are no longer in doubt. The widespread expectation is that AI-based cheating is rampant everywhere.[177]

136. There is ample evidence that this impact is not limited to the US. More recent studies in the UK show similar types of cheating is also rampant elsewhere. A study conducted by the Guardian in the UK found roughly 7000 proven cases of cheating using AI, and that is expected to be only the tip of the iceberg. The results showed the rate of cheating using AI *tripled* between

---

[173] *See, e.g.,* Hao-Ping (Hank) Lee et al., *The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects From a Survey of Knowledge Workers*, CHI '25: PROCEEDINGS OF THE 2025 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (Apr. 25, 2025), https://dl.acm.org/doi/10.1145/3706598.3713778.

[174] *See, e.g.,* Anastasia Berg, *Why Even Basic A.I. Use Is So Bad for Students*, N.Y. TIMES (Oct. 29, 2025), https://www.nytimes.com/2025/10/29/opinion/ai-students-thinking-school-reading.html.

[175] Diana Kwon, *AI Is Complicating Plagiarism. How Should Scientists Respond?*, NATURE (Jul. 30, 2024), https://www.nature.com/articles/d41586-024-02371-z; Michael Barbaro et al., *Suspicion, Cheating and Bans: A.I. Hits America's Schools*, N.Y. TIMES (Jun. 28, 2023), https://www.nytimes.com/2023/06/28/podcasts/the-daily/ai-chat-gpt-schools.html.

[176] Jeremy Weissman, *ChatGPT Is a Plague Upon Education*, INSIDE HIGHER ED. (Feb. 8, 2023), https://www.insidehighered.com/opinion/views/2023/02/08/chatgpt-plague-upon-education-opinion.

[177] Matt Barnum & Deepa Seetharaman, *There's a Good Chance Your Kid Uses AI to Cheat*, WALL ST. J. (Mar. 15, 2025), https://www.wsj.com/tech/ai/chatgpt-ai-cheating-students-97075d3c?st=SPUr9W&reflink=desktopwebshare_permalink; James D. Walsh, *Everyone Is Cheating Their Way Through College. ChatGPT Has Unraveled the Entire Academic Project*, N.Y. MAG. (May 7, 2025), https://nymag.com/intelligencer/article/openai-chatgpt-ai-cheating-education-college-students-school.html.

2022-2023 and 2023-2024, and that early results in 2025 projected another 50% increase in 2025.[178]

137. Beyond trying to quantify the level of AI-enabled cheating, the question now shifts to whether LLMs are simply killing education. Journalists talking to faculty and students across multiple institutions now report that it is no longer about students cheating and faculty trying to detect them, but it's more about "ambivalence and disillusionment and despair."[179] Perhaps nothing exemplifies this effect more than when faculty realize that even apologies from students caught cheating using AI, were written with AI.[180]

138. What's more, the difficulty of detecting LLM-generated text is making these negative impacts even more severe. In the early days of LLMs, researchers developed several detectors that used metrics like vocabulary perplexity to distinguish between LLM-generated text and human-generated text. Unfortunately, those efforts all proved fragile, and often biased against specific populations, *e.g.*, non-native speakers were often falsely flagged as using LLMs.[181] As more detectors were proposed, they still showed vulnerability to tools designed explicitly to make LLM-text look more realistic and harder to detect. In fact, an entire ecosystem of "humanizer" tools now exists (QuillBot, Refinely, Originality.ai, UnAIMyText, HumanizeAI, TwainGPT, AI Humanize, Humbot, etc.), and their sole purpose is to assist students in cheating using LLMs while

---

[178] Michael Goodier, *Revealed: Thousands of UK University Students Caught Cheating Using AI*, THE GUARDIAN (Jun. 15, 2025), https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey.

[179] Sean Illing, *Is ChatGPT Killing Higher Education?*, VOX (Jul. 5, 2025), https://www.vox.com/the-gray-area/418793/chatgpt-claude-ai-higher-education-cheating.

[180] Neil Vigdor and Hannah Ziegler, *Their Professors Caught Them Cheating. They Used A.I. to Apologize*, N.Y. TIMES (Oct. 29, 2025), https://www.nytimes.com/2025/10/29/us/university-illinois-students-cheating-ai.html.

[181] *See* Weixin Liang et al., *GPT Detectors Are Biased Against Non-Native English Writers*, PATTERNS (Jul. 14, 2023), https://doi.org/10.1016/j.patter.2023.100779.

avoiding detection by LLM-detectors.[182] Adding to all this is the fact that honest students are now caught in the crossfire, and are seriously concerned about how to prove that they did not use AI in case they are flagged by AI detectors in a false positive.[183] This is hardly the beneficial tool for education that Dr. Altman paints LLMs to be.

139. Yet another highly reputable report on this topic is the very recently released report on Education and AI, from the Center for Universal Education at the Brookings Institution, titled *A New Direction for Students in an AI World: Prosper, Prepare, Protect*.[184] In it, educational experts and researchers consider the potential benefits and harms of AI to education. They identify numerous potential harms, noting that AI can 1) "undermine students' cognitive development," 2) "impede students' social and emotional development," 3) "degrade trust in education," 4) "threaten students' safety," 5) "erode students' autonomy and agency," and 6) "deepen equity divides."[185] In weighing the risks and benefits, the report concludes, "[c]urrently, AI's risks overshadow its benefits."[186]

140. Compared to these research studies and factual reporting by news outlets, journalists, and researchers, Dr. Altman's support for AI's benefit to education generally come in the form of speculative statements on what LLMs "can do."[187]

141. In support of his assertions that LLMs are providing benefits to numerous industries, Dr. Altman cites a number of new projects, companies, products and tools designed to

---

[182] Anangsha Alammyan, *I Tried 30+ AI Humanizers. Here Are the Best 10 Tools to Bypass AI Detectors,* MEDIUM (Apr. 24, 2025), https://medium.com/freelancers-hub/best-ai-humanizer-to-bypass-ai-detectors-628590da5ccf.

[183] Callie Holtermann, *A New Headache for Honest Students: Proving They Didn't Use A.I.*, N.Y. TIMES (May 17, 2025), https://www.nytimes.com/2025/05/17/style/ai-chatgpt-turnitin-students-cheating.html.

[184] Mary Burns et al., *A New Direction for Students in an AI World: Prosper, Prepare, Protect*, CENTER FOR UNIV. ED. AT BROOKINGS (Jan. 2026), https://www.brookings.edu/wp-content/uploads/2026/01/A-New-Direction-for-Students-in-an-AI-World-FULL-REPORT.pdf.

[185] *Id.* at 19.

[186] *Id.* at 119.

[187] *See* Altman Report ¶¶87–88.

use LLMs to accelerate work in a variety of industries. However, nearly all of these are nascent projects that are experimental in nature. Mere efforts to build LLM tools is not proof that they are effective or providing the benefit they are intended to provide. In fact, a number of recent research studies by academics and industry groups show that the overwhelming majority of these projects by industry have failed to achieve their goals. A popular study conducted by MIT showed that "95% of generative AI projects at companies are failing."[188] Another report by a Wall Street analyst from Goldman Sachs suggests that AI projects will only generate half the profit needed to justify their hefty costs.[189] Companies who laid off human workers with hopes that they can be replaced by AI to increase overall productivity have reported actually backtracking after discovering that AI produced low-quality work product.[190]

142. In any event, Dr. Altman does not connect any of the purported benefits of LLMs he describes to the inclusion of song lyrics (either Publishers' lyrics specifically or song lyrics more generally) in LLM training data or their output by Claude. Nor would there be any reason to expect that including lyrics would significantly affect the use cases Dr. Altman describes. The use cases for which song lyrics are most relevant—regurgitating requested lyrics, answering questions about lyrics, writing new song lyrics and poetry, *etc.*—are far afield from the use cases Dr. Altman discusses.

143. Moreover, if Dr. Bilmes's opinion that song lyrics have an immaterial effect on LLM performance is correct, then including or excluding song lyrics from LLM training data would be expected to have no effect on the use cases Dr. Altman discusses. Even if the benefits

---

[188] Sheryl Estrada, *MIT Report: 95% of Generative AI Pilots at Companies Are Failing*, FORTUNE (Aug. 18, 2025), https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/.

[189] Jim Edwards, *AI May Generate Only Half the Profit Needed to Justify the Investment, Goldman Analyst Warns*, FORTUNE (Jan. 7, 2026), https://fortune.com/2026/01/07/ai-companies-profit-capex-investment-goldman-sachs-stocks/.

[190] Joe Wilkins, *Company Regrets Replacing All Those Pesky Human Workers with AI, Just Wants Its Humans Back*, FUTURISM (May 13, 2025), https://futurism.com/klarna-openai-humans-ai-back.

from LLMs Dr. Altman discusses were all real (and, as discussed above, they are not), Anthropic could avoid infringing Publishers' works without sacrificing any of those benefits.

Dated: January 23, 2026

Ben Zhao

REBUTTAL EXPERT REPORT OF BEN ZHAO

APPENDIX 1: DOCUMENTS CONSIDERED

DISCOVERY, PLEADINGS, AND PRODUCED DOCUMENTS

Action, Discovery, or Pleading
First Amended Complaint (April 25, 2025) ECF No. 337
Anthropic's Amended and Supplemental Responses and Objections to Plaintiffs' First Set of Interrogatories (Nov. 12, 2025)
Declaration of Jared Kaplan in Support of Defendant's Opposition to Plaintiff's Renewed Motion for Preliminary Injunction (Aug. 22, 2024), ECF No. 209
Expert Report of Plaintiff Expert Dr. Ben Zhao, PhD (Dec. 19, 2025)
Expert Report of Defense Expert Jeffrey Bilmes, Ph.D (Dec. 19, 2025)
Expert Report of Defense Expert Russ Altman, M.D. (Dec. 19, 2025)
Expert Report of Plaintiff Expert Shawn Shan (Dec. 19, 2025)
Rebuttal Expert Report of Joshua Dennis (Jan. 23, 2026)
Rebuttal Expert Report of Shawn Shan (Jan. 23, 2026)

Deposition Transcripts and Exhibits
Ben Mann 30(b)(6) Deposition Exhibit 187 (Anthropic_0000349548)
Daniela Amodei Deposition Exhibit 179 (Anthropic_0000458491)
Dario Amodei Deposition Exhibit 246 (Anthropic_0000605700)
Dario Amodei Deposition Exhibit 250
Dario Amodei Deposition Exhibit 251 (Anthropic_0000478585)
Dawn Drain 30(b)(1) Deposition Transcript (November 12, 2025)
Dawn Drain 30(b)(1) Deposition Exhibit 140 (Anthropic_0000507796)
Deep Ganguli 30(b)(1) Dep. Exhibit 25 (Anthropic_0000013680)
Jane Leibrock 30(b)(6) Deposition Transcript (October 15, 2025)
Jane Leibrock 30(b)(6) Deposition Exhibit 22
Jared Kaplan 30(b)(1) Deposition Transcript (November 21, 2025)

Bates-Stamped Documents
Anthropic_0000000588
Anthropic_0000012260
Anthropic_0000016025
Anthropic_0000016458
Anthropic_0000226136
Anthropic_0000240654
Anthropic_0000349548
Anthropic_0000459162
Anthropic_0000458427
Anthropic_0000525947

Anthropic_0000527903
Anthropic_0000554319
Anthropic_0000554366
Anthropic_0000554980
Anthropic_0000554769
Anthropic_0000605740

INTERNET AND ACADEMIC SOURCES

1. 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.
2. Ahmed Ahmed et al., *Extracting Books from Production Language Models*, ARXIV:2601.02671v1 (Jan. 6, 2026) https://arxiv.org/pdf/2601.02671.6, 2026), https://arxiv.org/abs/2601.02671.
3. Ahmed B. Mustafa et al., *Anyone Can Jailbreak: Prompt-Based Attacks on LLMs and T2Is*, ARXIV:2507.21820v1 (Jul. 29, 2025), https://arxiv.org/pdf/2507.21820.
4. Aneesh Pappu et al., *Measuring Memorization in RLHF for Code Completion*, PROC. OF THE THIRTEENTH INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR) (April 2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/22811d2089d4aa0ba66e52a5e47efb65-Paper-Conference.pdf.
5. Andy Zou et al., *Universal and Transferable Adversarial Attacks on Aligned Language Models*, ARXIV:2307.15043v2 (Dec. 20, 2023), https://arxiv.org/pdf/2307.15043.
6. Aladin Djuhera et al., *Fixing It in Post: A Comparative Study of LLM Post-Training Data Quality and Model Performance*, 39TH ANN. CONF. ON NEURAL INF. PROC. SYS. (NEURIPS) (Dec. 15, 2025), https://arxiv.org/pdf/2506.06522.
7. Anangsha Alammyan, *I Tried 30+ AI Humanizers. Here Are the Best 10 Tools to Bypass AI Detectors,* MEDIUM (Apr. 24, 2025), https://medium.com/freelancers-hub/best-ai-humanizer-to-bypass-ai-detectors-628590da5ccf.
8. Ananya, *AI Transcription Tools 'Hallucinate,' Too. Study Finds Surprisingly Harmful Fabrications in OpenAI's Speech-To-Text Algorithm*, SCIENCE (Apr. 26, 2024), https://www.science.org/content/article/ai-transcription-tools-hallucinate-too
9. Anastasia Berg, *Why Even Basic A.I. Use Is So Bad for Students*, N.Y. TIMES (Oct. 29, 2025), https://www.nytimes.com/2025/10/29/opinion/ai-students-thinking-school-reading.html.
10. Antonia Karamolegkou et al., *Copyright Violations and Large Language Models*, Proc. of the 2023 Conf. on Empirical Methods in Nat. Language Processing 7403 (2023), https://aclanthology.org/2023.emnlp-main.458/.

11. Aryo Pradipta Gema et al., *Are We Done with MMLU?* PROC. OF THE 2025 CONF. OF THE NATIONS OF THE AMERICAS CHAPTER OF THE ASSOC. FOR COMP. LINGUISTICS: HUMAN LANGUAGE TECH., https://aclanthology.org/2025.naacl-long.262/.
12. AWS, *What is Overfitting?*, https://aws.amazon.com/what-is/overfitting/ (last visited Jan. 22, 2025).
13. Bradley D. Menz et al., *Current Safeguards, Risk Mitigation, and Transparency Measures of Large Language Models Against the Generation of Health Disinformation: Repeated Cross Sectional Analysis*, 384 BMJ, Mar. 20, 2024, https://www.bmj.com/content/384/bmj-2023-078538
14. Brian Tarran, Why Large Language Models Should Come with a Content Warning, REAL WORLD DATA SCIENCE (Nov. 23, 2022), https://realworlddatascience.net/the-pulse/editors-blog/posts/2022/11/23/LLM-content-warning.html
15. Callie Holtermann, *A New Headache for Honest Students: Proving They Didn't Use A.I.*, N.Y. TIMES (May 17, 2025), https://www.nytimes.com/2025/05/17/style/ai-chatgpt-turnitin-students-cheating.html.
16. Catherine A. Gao et al., *Comparing Scientific Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers*, 6 NPJ Digit. Med., Apr. 26, 2023, https://pubmed.ncbi.nlm.nih.gov/37100871/
17. D. Hendrycks, et al., *Measuring Massive Multitask Language Understanding*, ARXIV:2009.03300v3 (Jan. 12, 2021), https://arxiv.org/pdf/2009.03300.
18. Dataset Card for HH-RLHF, HUGGING FACE, https://huggingface.co/datasets/Anthropic/hh-rlhf (last visited Jan. 22, 2026).
19. David Krauss, *Bottom-Up Versus Top-Down Processing: How Interventions Work*. PSYCHOLOGY TODAY, (Nov. 1, 2021), https://www.psychologytoday.com/us/blog/atypical-children-extraordinary-parenting/202111/bottom-up-versus-top-down-processing-how.
20. DEEPTEAM, https://www.trydeepteam.com/ (last visited Jan. 23, 2026).
21. Dhruv Mehrotra and Tim Marchman, *Perplexity Is a Bullshit Machine*, WIRED (Jun. 19, 2024), https://www.wired.com/story/perplexity-is-a-bullshit-machine/.
22. Diana Kwon, *AI Is Complicating Plagiarism. How Should Scientists Respond*? NATURE (Jul. 30, 2024), https://www.nature.com/articles/d41586-024-02371-z
23. Enze Liu et al., *Somesite I Used to Crawl: Awareness, Agency and Efficacy in Protecting Content Creators from AI Crawlers*, IMC'25: PROC. OF THE 2025 ACM INT. MEASUREMENT CONF. 78 (Oct. 15, 2025), https://doi.org/10.1145/3730567.3732913.
24. F. Mireshghallah et al., *An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models*, PROC. OF THE 2022 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (2022).
25. H. Mahdavi, et al., *Brains vs. Bytes: Evaluating LLM Proficiency in Olympiad Mathematics*, ARXIV:2504.01995v2 (April 10, 2025), https://arxiv.org/pdf/2504.01995.

26. Hallucination Leaderboard, https://github.com/vectara/hallucination-leaderboard/blob/main/README.md (last visited Jan. 23, 2026).
27. Hao-Ping (Hank) Lee et al., *The Impact of Generative AI on Critical Thinking: Self-Reported Reductions in Cognitive Effort and Confidence Effects from a Survey of Knowledge Workers*. CHI 2025 (Apr. 2025), https://dl.acm.org/doi/10.1145/3706598.3713778.
28. Hussam Alkassi & Samy I. McFarlene, *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing*, 15 CUREUS (Feb. 19, 2023), https://www.cureus.com/articles/138667-artificial-hallucinations-in-chatgpt-implications-in-scientific-writing/
29. I. Dasgupta et al., *Distinguishing rule and exemplar-based generalization in learning systems*, PROCEEDINGS OF THE 39TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING (Vol. 162), 4816–4830 (2022).
30. I. Petrov, et al., *Proof or Bluff? Evaluating LLMs on 2025 USA Math Olympiad*, ARXIV:2503.21934v5 (Sept. 4, 2025), https://arxiv.org/pdf/2503.21934.
31. *Introducing Phi-4: Microsoft's Newest Small Language Model Specializing in Complex Reasoning*, MICROSOFT FOUNDRY BLOG (Dec. 12, 2024), https://techcommunity.microsoft.com/blog/azure-ai-foundry-blog/introducing-phi-4-microsoft%E2%80%99s-newest-small-language-model-specializing-in-comple/4357090.
32. *Introduction to Robots.txt*, GOOGLE, https://developers.google.com/search/docs/crawling-indexing/robots/intro.
33. J. Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, 36TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS) (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.
34. James D. Walsh, *Everyone Is Cheating Their Way Through College. ChatGPT Has Unraveled the Entire Academic Project*, N.Y. MAG. (May 7, 2025), https://nymag.com/intelligencer/article/openai-chatgpt-ai-cheating-education-college-students-school.html.
35. Jan Homolak, *Opportunities and Risks of ChatGPT in Medicine, Science, and Academic Publishing: A Modern Promethean Dilemma*, Croat. Med. J. no. 1, (Feb. 2023), https://pmc.ncbi.nlm.nih.gov/articles/PMC10028563/.
36. Jeremy Weissman, *ChatGPT Is a Plague Upon Education*, INSIDE HIGHER ED. (Feb. 8, 2023), https://www.insidehighered.com/opinion/views/2023/02/08/chatgpt-plague-upon-education-opinion.
37. Jim Edwards, *AI May Generate Only Half the Profit Needed to Justify the Investment, Goldman Analyst Warns*, FORTUNE (Jan. 7, 2026), https://fortune.com/2026/01/07/ai-companies-profit-capex-investment-goldman-sachs-stocks/.

38. Joe Wilkins, *Company Regrets Replacing All Those Pesky Human Workers with AI, Just Wants Its Humans Back*, FUTURISM (May 13, 2025), https://futurism.com/klarna-openai-humans-ai-back.
39. L. Berglund, et al., *The Reversal Curse: LLMs trained on "A Is B" Fail to Learn "B Is A,"* arXiv:2309.12288v3 (April 4, 2024), https://arxiv.org/pdf/2309.12288v3.
40. Lance Eliot, *AI Ethics Lucidly Questioning This Whole Hallucinating AI Popularized Trend That Has Got to Stop*. FORBES (Aug. 24, 2022), https://www.forbes.com/sites/lanceeliot/2022/08/24/ai-ethics-lucidly-questioning-this-whole-hallucinating-ai-popularized-trend-that-has-got-to-stop/.
41. *Learning to Reasoning with LLMs*, OPENAI (Sept. 12, 2024), https://openai.com/index/learning-to-reason-with-llms/.
42. Liz Mineo, *Is AI Dulling Our Minds?*, THE HARVARD GAZETTE (Nov. 13, 2025), https://news.harvard.edu/gazette/story/2025/11/is-ai-dulling-our-minds/.
43. Mantas Sabeckis & Jurgita Lapienyte, *We Tested ChatGPT, Gemini, and Claude with Adversarial Prompts: Here Are Our Findings and Risks*, CYBERNEWS.COM (Nov. 13, 2025), https://cybernews.com/security/we-tested-chatgpt-gemini-and-claude/.
44. Marah Abdin et al., *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*, ARXIV:2404.14219 (Aug. 30, 2024), https://arxiv.org/pdf/2404.14219.
45. Marcel Binz et al., *How Should the Advancement of Large Language Models Affect the Practice of Science?*, 122 Proc. Natl Acad. Sci. USA, Jan. 27, 2025, https://www.pnas.org/doi/10.1073/pnas.2401227121.
46. Marco Tulio Ribeiro, Sameer Singh & Carlos Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, ARXIV:1602.04938v3 (Aug. 9, 2016), https://arxiv.org/pdf/1602.04938.
47. Mark Sullivan, *Yann LeCun: Meta "Fudged a Little Bit" When Benchmark-Testing Llama 4 Model,* YAHOO FINANCE (Jan. 6, 2026), https://finance.yahoo.com/news/yann-lecun-meta-fudged-little-100000402.html.
48. Mary Burns et al., *A New Direction for Students in an AI World: Prosper, Prepare, Protect*, CENTER FOR UNIV. ED. AT BROOKINGS (Jan. 2026), https://www.brookings.edu/wp-content/uploads/2026/01/A-New-Direction-for-Students-in-an-AI-World-FULL-REPORT.pdf).
49. Mateja Vuradin & Emily Hayes, *How Safe Is Anthropic's "Safest" Model? We Red Teamed Claude Sonnet 4.5*, ZSCALER (Oct. 15 2025), https://splx.ai/blog/red-teaming-claude-sonnet-4-5.
50. Matt Barnum & Deepa Seetharaman, *There's a Good Chance Your Kid Uses AI to Cheat*, WALL ST. JOURNAL (Mar. 15, 2025), https://www.wsj.com/tech/ai/chatgpt-ai-cheating-students-97075d3c?st=SPUr9W&reflink=desktopwebshare_permalink
51. Matthew Zapruder, *The Difference Between Poetry and Song Lyrics*, BOSTON REV. (Dec. 6, 2012), https://www.bostonreview.net/forum/poetry-brink/difference-between-poetry-and-song-lyrics/.

52. Matthias Stadler et al., *Cognitive Ease at a Cost: LLMs Reduce Mental Effort but Compromise Depth in Student Scientific Inquiry*, 60 COMPUTERS IN HUMAN BEHAVIOR (Nov. 2024), https://www.sciencedirect.com/science/article/pii/S0747563224002541.
53. Meir Michanie*, Overfitting Is All You Need*, MEDIUM.COM (Jun. 30, 2024), https://medium.com/@meirgotroot/overfitting-is-all-you-need-174caa7f1064.
54. Michael Barbaro et al., *Suspicion, Cheating and Bans: A.I. Hits America's Schools*, N.Y. TIMES (Jun. 28, 2023), https://www.nytimes.com/2023/06/28/podcasts/the-daily/ai-chat-gpt-schools.html.
55. Michael Goodier, *Revealed: Thousands of UK University Students Caught Cheating Using AI*, THE GUARDIAN (Jun. 15, 2025), https://www.theguardian.com/education/2025/jun/15/thousands-of-uk-university-students-caught-cheating-using-ai-artificial-intelligence-survey.
56. *Multiple AI Companies Bypassing Web Standard to Scrape Publisher Sites, Licensing Firm Says*, Reuters.com (Jun. 2024), https://www.reuters.com/technology/artificial-intelligence/multiple-ai-companies-bypassing-web-standard-scrape-publisher-sites-licensing-2024-06-21/,
57. N. Cohen-Inger, et al., *Forget What You Know about LLMs Evaluations—LLMs are Like a Chameleon*, PROCEEDINGS OF THE 2025 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP), 21664–21677 (2025); S. Liang et al., *The SWE-Bench Illusion: When State-of-the-Art LLMs Remember Instead of Reason*, arXiv:2506.12286v4 (Dec. 1, 2025), https://arxiv.org/pdf/2506.12286.
58. N. Kandpal et al., *The CommonPile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text*, ARXIV:2506.05209v1 (Jun. 5, 2025), https://arxiv.org/pdf/2506.05209.
59. Neil Vigdor & Hannah Ziegler, *Their Professors Caught Them Cheating. They Used A.I. to Apologize*, N.Y. TIMES (Oct. 29, 2025), https://www.nytimes.com/2025/10/29/us/university-illinois-students-cheating-ai.html.
60. Nicholas Carlini et al., *Quantifying Memorization Across Neural Language Models*, 2022 Int. Conf. on Learning Representations (Feb. 15, 2022), https://arxiv.org/pdf/2202.07646.
61. S. Balloccu, et al., *Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs*, PROC. OF THE 18TH CONF. OF THE EUROPEAN CHAPTER OF THE ASSOC. FOR COMP. LINGUISTICS VOLUME 1: LONG PAPERS, 67–93 (2024), https://aclanthology.org/2024.eacl-long.5/.
62. S. Liang et al., *The SWE-Bench Illusion: When State-of-the-Art LLMs Remember Instead of Reason*, arXiv:2506.12286v4 (Dec. 1, 2025), https://arxiv.org/pdf/2506.12286.
63. S. Yadlowsky, *Pretraining Data Mixtures Enable Narrow Model Selection Capabilities in Transformer Models*, ARXIV:2311.00871v1 (Nov. 3, 2023), https://arxiv.org/pdf/2311.00871.
64. S. Zheng et al., *Exploring Memorization in Fine-Tuned Language Models*, PROCEEDINGS OF THE 62ND ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (Volume 1: Long Papers).

65. Sally Beatty, *Tiny But Mighty: The Phi-3 Small Language Models with Big Potential*, MICROSOFT (Apr. 23, 2024), https://news.microsoft.com/source/features/ai/the-phi-3-small-language-models-with-big-potential/.
66. Sean Illing, *Is ChatGPT Killing Higher Education?*, VOX (Jul. 5, 2025), https://www.vox.com/the-gray-area/418793/chatgpt-claude-ai-higher-education-cheating.
67. Shayne Longpre et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*, arXiv:2305.13169 (Nov. 13, 2023), https://doi.org/10.48550/arXiv.2305.13169.
68. Sheryl Estrada, *MIT Report: 95% of Generative AI Pilots at Companies Are Failing*, FORTUNE (Aug. 18, 2025), https://fortune.com/2025/08/18/mit-report-95-percent-generative-ai-pilots-at-companies-failing-cfo/.
69. Suhas BN et al., *Fact-Controlled Diagnosis of Hallucinations in Medical Text Summarization*, PROC. OF INTERSPEECH 2025 (Aug. 2025), https://www.isca-archive.org/interspeech_2025/bn25_interspeech.pdf.
70. Taein Kim et al., S*crapers Selectively Respect robots.txt Directives: Evidence from a Large-Scale Empirical Study*, IMC'25: PROC. OF THE 2025 ACM INT. MEASUREMENT CONF. 541 (Oct. 15, 2025), https://doi.org/10.1145/3730567.3764471.
71. U.S. COPYRIGHT OFFICE, COPYRIGHT AND ARTIFICIAL INTELLIGENCE, PART 3: GENERATIVE AI TRAINING (pre-publication version) 19–20 (2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf.
72. Uwe Peters & Benjamin Chin-Yee, *Generalization Bias in Large Language Model Summarization of Scientific Research*, Royal Society of Open Science (Apr. 30, 2025), https://royalsocietypublishing.org/rsos/article/12/4/241776/235656/Generalization-bias-in-large-language-model.
73. Vitaly Feldman, *Does Learning Require Memorization? A Short Tale About a Long Tail,* PROC. OF THE 52ND ANN. ACM SIGACT SYMP. ON THEORY OF COMP. 954 (Jun. 22, 2020), https://doi.org/10.1145/3357713.3384290.
74. Vitaly Feldman and Chiyuan Zhang, *What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation*, NIPS'20: Proc. of the 34th Int. Conf. on Neural Inf. Proc. Sys. 2881, https://proceedings.neurips.cc/paper/2020/file/1e14bfe2714193e7af5abc64ecbd6b46-Paper.pdf.
75. W. Choi et al., *A Review of "Do Anything Now" Jailbreak Attacks in Large Language Models: Potential Risks, Impacts, and Defense Strategies,* PREPRINTS.ORG (Sept. 1, 2025), https://www.preprints.org/frontend/manuscript/1955ad799746b0c4ad7ced819c095573/download_pub.
76. Weixin Liang et al., *GPT Detectors Are Biased Against Non-Native English Writers*, 4 PATTERNS (Jul. 14, 2023), https://doi.org/10.1016/j.patter.2023.100779.

77. *Why Language Models Hallucinate*, OPENAI (Sept. 2025), https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf.
78. X. Shen et al., *"Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models*, arXiv:2308.03825v2 (May 15, 2024), https://arxiv.org/pdf/2308.03825.
79. Y. Chan et al., *Speak Easy: Eliciting Harmful Jailbreaks from LLMs with Simple Interactions*, ARXIV:2502.04322v3 (Aug. 2, 2025), https://arxiv.org/pdf/2502.04322.
80. Y. Weng et al., *Large Language Models are Better Reasoners with Self-Verification*, 2023 CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP), 2550–2575 (2023).
81. Yubin Kim et al., *Medical Hallucination in Foundation Models and Their Impact on Healthcare*, MEDR XIV (Nov. 2, 2025), https://doi.org/10.1101/2025.02.28.25323115.
82. Z. Yong et al., *Low-Resource Languages Jailbreak GPT-4*, ARXIV:2310.02446v2 (Jan. 27, 2024), https://arxiv.org/pdf/2310.02446.
83. Ziwei Ji et al., *Survey of Hallucination in Natural Language Generation*, 55 ACM COMPUTING SURVEYS (Mar. 3, 2023), https://doi.org/10.1145/3571730.