EXHIBIT 5

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

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.

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.

Dated: January 23, 2026

Ben Zhao