EXHIBIT 6

(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

REPLY EXPERT REPORT OF DR. BEN ZHAO

February 13, 2026

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

B. Memorization Is Inherent, Not Aberrational

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

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

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

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

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

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

[141] Bilmes Rebuttal ¶ 90.

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

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

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

[145] *Id.* ¶ 114.

[146] *Id.* ¶ 118–19.

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

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

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

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

---

[147] Zhao Rebuttal ¶ 94.

[148] Bilmes Rebuttal ¶¶ 93–97.

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

[150] *Id.*

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



Ben Zhao