EXHIBIT 11

(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. SHAWN SHAN

January 23, 2026

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

recognized, this vulnerability is compounded in long-form conversations where the effectiveness of safety instructions naturally degrades over time.[34]

40. The third fundamental limitation of Prompt Shield is its narrow focus on blocking literal reproduction while completely ignoring the model's ability to generate "new" derivative content based on its training. Because the model has already ingested the Publishers' works, it has internalized the specific semantic and stylistic patterns of those artists. To test this vulnerability, I conducted an experiment using the Claude 3 Haiku API, asking the model to generate lyrics in the style of "Coldplay"—arists whose works are among the Works in Suit.[35] In 10 separate attempts, the model did not refuse a single time, consistently producing new lyrics that mimicked the band's signature style.[36] This confirms that Prompt Shield is not designed to prevent prompts or outputs involving "new" lyrics that exploit the model's unauthorized training on protected catalogs. While the guardrail might catch an obvious request for a direct copy, it offers no protection against users seeking to create original compositions that serve as stylistic substitutes for the protected works. This structural failure demonstrates that the model continues to derive significant value from the infringing material even when literal reproduction is suppressed.

41. The "Content Compare System" (CCS) is a fundamentally fragile defense because it relies on [redacted]. As a result, CCS will not operate successfully if texts that Anthropic intend to block in outputs are not present in CCS' hash corpus.[37]

42. Additionally, while CCS is designed to block responses that match a database of copyrighted sequences, it is incapable of detecting those same sequences if the model is instructed to alter the formatting. For example, a user can instruct the model to replace every whitespace with a random string like "4mzmTA," turning the phrase "How are you" into

---

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

[35] *See* claude_coldplay_style.py (on file with Publishers' counsel).

[36] *See* Appendix B.

[37] *See* Zhao Report ¶¶ 131.

Dated: January 23, 2026

Shawn Shan