EXHIBIT 8

HIGHLY CONFIDENTIAL – ATTORNEYS' EYES ONLY

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

EXPERT REPORT OF DR. SHAWN SHAN

December 19, 2025

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

22. However, to acquire these sufficiently large training corpora, developers often turn to a massive source of easily accessible text: the Internet. Developers typically source large portions of their training corpora by scraping text from webpages,[6] a process by which bots "crawl" through the internet and copy the contents of webpages it encounters.[7] Developers may also acquire datasets constructed by third parties, such as "Common Crawl,"[8] "The Pile,"[9] and "OpenWebText2,"[10] which are themselves largely comprised of petabytes of content scraped from the Internet. Developers may also enter into licenses with rights holders or data aggregators for their library of content, including creative content such as books, magazines, music, and code.

23. Web crawling is an automated process by which bots copy the contents of webpages they come across on the Internet. However, crawling is not always an indiscriminate process. Crawlers may be given instructions to prioritize scraping from certain websites over others; developers may choose to scrape specific domains or webpages depending on their data needs and goals.[11] Crawlers may also be programmed to review a webpage's robots.txt file before it begins scraping. If a webpage's owner does not wish for crawlers to scrape from their website, they may include "do not crawl" instructions (*i.e.,* Robots Exclusion Protocol) in the robots.txt file.[12] Crawlers may or may not obey those instructions, depending on how they are programmed.[13]

---

[6] *See* Kevin Roose, *The Data That Powers A.I. is Disappearing Fast*, N.Y. TIMES (Jul. 19, 2024), https://www.nytimes.com/2024/07/19/technology/ai-data-restrictions.html.
[7] *See generally* UNITED STATES COPYRIGHT OFFICE, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (pre-publication version), 13–14 (2025), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf.
[8] *Overview*, COMMON CRAWL, https://commoncrawl.org/overview (last accessed Dec. 15, 2025).
[9] *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, THE PILE, https://pile.eleuther.ai/ (last accessed Dec. 15, 2025).
[10] *OpenWebText2*, https://openwebtext2.readthedocs.io/en/latest/ (last accessed Dec. 15, 2025).
[11] *See, e.g.*, *What Is a Web Crawler? | How Web Spiders Work*, CLOUDFLARE, https://www.cloudflare.com/learning/bots/what-is-a-web-crawler/ (last accessed Dec. 15, 2025).
[12] *RFC 9309 Robots Exclusion Protocol*, RFC EDITOR, https://www.rfc-editor.org/rfc/rfc9309.html (last accessed Dec. 15, 2025).
[13] *See, e.g.*, Kali Hays, *OpenAI and Anthropic Are Ignoring an Established Rule That Prevents Bots Scraping Online Content*, BUSINESS INSIDER (Jun. 21, 2024), https://www.businessinsider.com/openai-anthropic-ai-ignore-rule-scraping-web-contect-robotstxt.

Dated: December 19, 2025

Shawn Shan