

**Lieff
Cabrer
Heimann
& Bernstein**
Attorneys at Law

SUSMAN GODFREY

May 14, 2025

VIA ECF

Judge William H. Alsup
Phillip Burton Federal Building & U.S. Courthouse
450 Golden Gate Avenue, 16th Floor Clerk's Office
San Francisco, CA 94102

RE: *Bartz, et. al. v. Anthropic PBC*, Case Number 3:24-CV-05417 (N.D.
Cal.): Letter Motion to Compel Production of Documents

Dear Judge Alsup:

Plaintiffs respectfully request an order directing Anthropic to produce: (1) its spreadsheet listing which datasets were used to train which LLM models, and (2) [REDACTED]. Anthropic is prejudicing Plaintiffs by refusing to produce these two key pieces of evidence because they are purportedly “too confidential” to be produced in civil discovery—full stop. Specifically, Anthropic claims Plaintiffs can only view this evidence in a highly-monitored and restrictive inspection environment. For the reasons shown below, this position is unnecessary, untenable, and hampers Plaintiffs’ ability to prepare their case.

I. Dataset Spreadsheet

Background: On March 18, 2025, Anthropic produced a spreadsheet showing the training data mix for each of its models.¹ For each model, the spreadsheet listed [REDACTED]

The spreadsheet was not only an efficient way for the parties’ discovery to proceed, but of crucial importance. From that spreadsheet, Plaintiffs learned about [REDACTED]

On April 1, Anthropic notified Plaintiffs that it was clawing back the spreadsheet and that it should have been designated as HIGHLY CONFIDENTIAL – SOURCE CODE “and/or” HIGHLY CONFIDENTIAL – TRAINING DATA. *See* Ex. A. Anthropic informed Plaintiffs that, going forward, Plaintiffs would have no access to the spreadsheet except through a “secure

¹ Anthropic produced three copies of the spreadsheet: two on March 18, 2025, and another on March 21, 2025. *See* Fredricks Decl. ¶ 2.

Judge William H. Alsup
 May 14, 2025
 Page 2

inspection environment” in San Francisco.² The spreadsheet, however, contains neither source code nor training data.

Over the next month, Plaintiffs met-and-conferred with Anthropic four times. *See* Fredricks Decl. ¶ 4. While Anthropic agreed to provide [REDACTED] it refused to provide Plaintiffs an unredacted version of the spreadsheet in any form—even if designated Attorney’s Eyes Only (“AEO”) under the parties’ protective order—outside the inspection environment. Plaintiffs made every effort to accommodate Anthropic’s confidentiality concerns, even suggesting that Anthropic could provide a *physical* print-out of the spreadsheet (which is beyond what is required for Plaintiffs to access materials from the secure environment³). The parties are at impasse.

Argument: There can be no dispute that the Dataset Spreadsheet is a responsive and relevant document.⁴ Documents identifying which training datasets were used to train which models—[REDACTED]—are obviously relevant to Plaintiffs’ claims that Anthropic committed copyright infringement by using copyrighted books to train its LLMs. The Dataset Spreadsheet confirms that Anthropic assembled the books into giant datasets and trained on them in uniform fashion. [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

While Anthropic has free and full access to this document it relies on for its bid to dismiss Plaintiffs’ claims, Plaintiffs should not be limited to accessing this key document in a secure, vault-like environment. That is especially true because the document in question is a spreadsheet. It is not source code and there is no dispute it is not itself training data.⁶ Under the heavily-negotiated protective order, Anthropic has no basis to impose such onerous restrictions on a document that is not, in fact, source code:

² To access the inspection environment, Plaintiffs must give Anthropic five business days’ notice. And because of the way the files are stored and made accessible, it is difficult (if not impossible) for attorneys without technical expertise to navigate the environment and inspect files. So far, all inspections have been conducted by Plaintiffs’ expert, who flies into San Francisco from Chicago. *See* Fredricks Decl. ¶ 3.

³ The Inspection Protocol provides that “[t]he Inspecting Party may obtain printouts of reasonable portions of the Inspection Data” and “‘printouts’ are understood to refer to a Bates-stamped **electronic** production.” Dkt. 83, Inspection Protocol at 6 (emphasis added).

⁴ The spreadsheet is responsive to at least RFP No. 15 (“Documents reflecting or discussing the use of copyrighted material to train any LLM, including but not limited to material sourced from Books3, LibGen, The Pile, E-Libra, Common Crawl, and YouTube.”)

⁵ For this reason, Anthropic’s suggestion that it can provided a redacted version that strips away all information except for books is not sufficient. At most, Anthropic could perhaps give non-books datasets unique identifiers, but not strip away the related information; the entire document is relevant.

⁶ Indeed, Anthropic itself has denied Plaintiffs’ print requests on this basis. *See* Ex. D (Anthropic refusing print request because “a full list . . . that reveals the names of every dataset used for pretraining . . . is not Anthropic’s training data, nor is it Anthropic’s source code.”).

Judge William H. Alsup
May 14, 2025
Page 3

A document that merely quotes or cites filenames . . . without reproducing any lines of source code is not a Source Code Document and may not be designated as “HIGHLY CONFIDENTIAL – SOURCE CODE” but should be designated “HIGHLY CONFIDENTIAL – ATTORNEYS’ EYES ONLY.”

Dkt. 62, ¶ 12 (emphasis added). The Court should compel Anthropic to produce the spreadsheet.

II. [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED] In any event, Plaintiffs are open to Anthropic’s suggestions of protection, but Anthropic cannot have it both ways: asserting difficulties of proof while limiting Plaintiffs’ access to the evidence.

Respectfully Submitted,

LIEFF CABRASER HEIMANN
& BERNSTEIN, LLP

SUSMAN GODFREY LLP

/s/ Rachel Geman
Rachel Geman

/s/ Justin A. Nelson
Justin A. Nelson

⁷ Ex. C, at 70:21-71:24.