



Joseph Farris
+1 415.471.3454 Direct
Joseph.Farris@arnoldporter.com

June 6, 2025

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

**Re: Anthropic's Response to Plaintiffs' Letter Motion to Compel Production
in *Bartz, et al. v. Anthropic PBC*, No. 3:24-cv-05417-WHA (Dkt. 190)**

Dear Judge Alsup:

Plaintiffs seek an order compelling Anthropic to “produce” (i) an extremely sensitive document that distills and consolidates the extensive research and development efforts for nearly all of Anthropic’s commercial models and several of its non-commercial models (the “Tokens Formula Log”); and (ii) the entirety of its Scanned Books Dataset. Yet, both of these items *were “produced” by Anthropic months ago in the secure inspection environment* created specifically for the Plaintiffs via the carefully negotiated and mutually agreed-upon “Training Data and Source Code Inspection Protocol” (the “Inspection Protocol”) (Dkt. 83) that was entered by this Court (Dkt. 85). To be clear: Plaintiffs have *full access* to them there. Similar protocols are standard in all of the AI copyright cases pending around the country precisely because of the extremely valuable and commercially sensitive nature of the core technical information at issue in these cases. Moreover, Anthropic has produced tens of thousands of other highly sensitive documents outside the inspection environment in reliance on standard designations under the Protective Order (Dkt. 62) and has only sparingly applied the relevant “Highly Confidential - Source Code” and “Highly Confidential - Training Data” designations to extremely sensitive information. Therefore, at bottom, Plaintiffs’ motion is not about “access” to data at all, but rather, Plaintiffs’ unwarranted insistence on putting Anthropic’s most sensitive data at risk of disclosure for their own convenience.

Tokens Formula Log

The Tokens Formula Log is a detailed ledger of many formulations of the training data mixes for Anthropic’s models at specific snapshots in time, and it reflects one of Anthropic’s most competitively sensitive areas of research. *See* Declaration of Mycal Tucker submitted herewith (“Tucker Decl.”) ¶¶ 3-8. It reveals the descriptive names of the datasets used to train most of Anthropic’s commercial LLMs, the type of data and/or provenance of each dataset, data processing measures that Anthropic employs, the number of times the model was exposed to each dataset (i.e., “multi-epoch”), the number of tokens in each dataset, and the percentage of tokens that each dataset contributes to the total number of tokens in the training data mix (i.e., the “weight” of the dataset in the training data mix). It contains, in other words, significant portions of the “recipes” for data used to build many of Anthropic’s LLMs.

Arnold & Porter

Hon. William Alsup
June 6, 2025
Page 2

Information like this plainly qualifies under the Inspection Protocol as “Highly Confidential - Source Code” information, which is defined broadly to include not just “computer code and associated comments and revision histories” but also “formulas, engineering specifications, or schematics that define or otherwise describe in detail the algorithms or structure of software or hardware designs” Dkt. 83, ¶ 1 (citing Dkt. 62, Section II, ¶ 12). This is precisely what the Tokens Formula Log does. In places, it exactly mirrors the information in Anthropic’s source code, including the precise numerical weight afforded to each dataset. Tucker Decl. ¶ 5. Moreover, it reflects years of trial and error and refinement of the formulation and processing of Anthropic’s models over time—and therefore provides a roadmap for a competitor to avoid the investment of time and resources it took to develop Anthropic’s understanding of optimal data procedures and composition. *Id.* ¶¶ 6-7.

The Tokens Formula Document is also highly restricted within Anthropic; only people with a specific need to know this information have access to it. *Id.* ¶ 8. In fact, the versions of this document that were inadvertently produced (and promptly clawed back) in this case were from the custodial files of a single employee on the Pretraining Team. Declaration of Jessica Gillotte (“Gillotte Decl.”) ¶ 3. The risk of an inadvertent disclosure of this information is extreme and cannot be effectively limited by an Attorney’s Eyes Only designation. To illustrate just one potential avenue for disclosure, Plaintiffs have already ***inadvertently filed on the public docket documents designated Attorney’s Eyes Only multiple times***, which resulted in their immediate disclosure to Anthropic’s competitors who are likely monitoring the docket. *Id.* ¶ 11.

Anthropic’s production of this document in the inspection environment also meets all Plaintiffs’ legitimate needs with respect to the Token Formula Log: (i) it is easily available in the secure inspection environment, an 8-minute walk from the offices of Plaintiffs’ counsel (Lief Cabraser San Francisco), (ii) it is in an ordinary format that is easily reviewable by attorneys, (iii) Anthropic has repeatedly waived the five-day notice period for inspections, (iv) Anthropic also answered expedited RFAs about its contents and produced a redacted version of the Token Formula Log, and (v) despite claiming that further access is of critical importance and being informed on April 1 that the document would instead be available in the inspection environment, ***Plaintiffs have never asked to inspect it there***. *Id.* at ¶¶ 3-6. In light of the readily available, secure access Plaintiffs already have, there is no need for any further accommodation.

Scanned Books Dataset

Plaintiffs’ request for “production” of the Scanned Books Dataset outside of the training data inspection environment is equally unfounded. Until recently, when it was disclosed via this litigation, even the existence of the Scanned Books Dataset was a closely guarded trade secret. Even after that disclosure, the Scanned Books Dataset itself remains one of Anthropic’s

Arnold & Porter

Hon. William Alsup
June 6, 2025
Page 3

most valuable datasets because it reflects a proprietary composition of books that Anthropic sourced, scanned, and created itself—it is available to no other AI company in the world. Tucker Decl. ¶ 9. It is an entirely unique dataset built from the ground up by Anthropic at the cost of tens of millions of dollars. *Id.* And its value derives not only from Anthropic’s substantial investment in creating it, but also in that its composition reflects Anthropic’s research into creating a valuation model for different types of books. *Id.* ¶ 10.

Plaintiffs do not dispute that the Scanned Books Dataset is precisely the type of “Training Data” under the Inspection Protocol that they stipulated to inspect in a specialized environment rather than via document production. Instead they raise three arguments for why there should be an exception to the Inspection Protocol made here, none of which have merit. First, Plaintiffs claim that this is necessary to determine what books are in the Scanned Books Datasets more conveniently, but fail to acknowledge that Anthropic has produced metadata information for its Scanned Books Datasets outside the inspection environment already (which Anthropic did **not** argue were unreliable in its class certification briefing). Gillotte Decl. ¶¶ 12-13. Second, Anthropic’s courtesy production of “raw” versions of the “Internet/Pirated Books Datasets” outside the inspection environment based on Plaintiffs’ request was not a waiver of its right to insist on compliance with the Protocol. Those datasets are publicly available on the internet and Anthropic did not create them via an extraordinary investment as it did with Scanned Books Datasets. Third, Plaintiffs’ suggestion that another court ordered the production of an “**entire training corpus**” in *Tremblay v. OpenAI, Inc.*, No. C 23-03223-AMO (N.D. Cal. Jan. 17, 2025) (Dkt. 284) is a misrepresentation; that dispute related to a single “dataset” of public web data which was not a trade secret.¹ Further, in *Tremblay*, the defendant’s inspection environment repeatedly failed to function adequately, which has not occurred here.

For these reasons, Anthropic respectfully requests that the motion be denied.

Sincerely,

/s/ Joseph Farris

Joseph Farris

¹ *Id.* at 10:2-6 (“I am going to take OpenAI’s suggestion that **the data set** will have to be only given over”) (emphasis added); *id.* Dkt. 254 at 1 (joint letter brief referring to the singular “English Colang Dataset”).

DECLARATION OF MYCAL TUCKER

I, Mycal Tucker, hereby declare as follows:

A. Personal Background

1. I am a Research Scientist at Anthropic PBC (“Anthropic”) on the pretraining team. As stated in my Declaration in support of Anthropic’s Opposition to Class Certification (Dkt. 146-5), my duties in this role include developing new data processing methods and preparing data to train our commercial models.
2. I submit this declaration in support of Anthropic’s Response to Plaintiffs’ Discovery Letter Motion to Compel (Dkt. 190). Unless stated otherwise, all facts stated herein are within my personal knowledge. If called upon, I would be willing to testify as set forth in this declaration.

B. The Tokens Formula Log

3. As described in more detail in the declaration of Jared Kaplan, Anthropic’s Chief Science Officer, in support of Anthropic’s Motion for Summary Judgment (Dkt. 119-5) (“Kaplan Decl.”), LLMs are developed from enormous and diverse datasets. The performance of an LLM depends in significant part on the quality and specific formulation of the mix of training data that is used to train it—in addition to other factors, such as the specific design of the model architecture employed, training techniques used, and compute resources devoted to training. The formulation of the training data mix varies greatly between models, and depends on many factors including the types of specific datasets used, overall quantity of data used for training, the ways data is processed or filtered before training, the number of times the tokens in each dataset are passed through the model (where the tokens are passed through the model multiple times, this is called “multi-epoching”), and the proportions of various types of data in the training corpus for any given model. *See also* Kaplan Decl. ¶¶ 35-53.
4. The Tokens Formula Log is a detailed ledger of many formulations of the training data mixes for Anthropic’s models at specific snapshots in time, and it reflects one of Anthropic’s most competitively sensitive areas of research. Specifically, the Tokens Formula Log reveals the descriptive names of the datasets used to train most of Anthropic’s LLMs intended for

1 commercial release, the type of data and/or provenance of each dataset, information that
2 indicates the data processing measures that Anthropic employs, the number of times the model
3 was exposed to each dataset (i.e., “multi-epoch”), the number of tokens in each dataset, and
4 the percentage of tokens that each dataset contributes to the total number of tokens in the
5 training data mix at the specific snapshot in time (i.e., the “weight” of the dataset in the training
6 data mix), among other highly sensitive information. It contains, in other words, significant
7 portions of the “recipes” to build many of Anthropic’s models, including both commercial
8 models and models that were used for research and development.

9 5. In certain places, the structure of the source code for Anthropic’s LLMs mirrors the exact same
10 structure of the information that is reflected in the Tokens Formula Log. For example, the
11 precise numerical weight afforded to each particular dataset in the training mix is reflected in
12 both the source code used to train the model as well as in the Tokens Formula Log.

13 6. The Tokens Formula Log contains over 40 distinct formulations reflecting such information,
14 with each tab generally corresponding to a training data mix for a commercial model
15 (depending on when the snapshot was taken during the pretraining process), training data mix
16 for engineering and research and developmental models, and/or training data mix proposals
17 that were considered for certain models. In other words, over time, the Tokens Formula Log
18 has been frequently updated to incorporate new research findings, refine the formulation of
19 training data mixes of models being prepared for commercial release, plan models for future
20 release, and to reflect proposed different variants of training data mixes for models intended
21 for commercial release. In this respect, it acts like a technical roadmap for how these critical
22 and highly confidential aspects of Anthropic’s models have evolved over time.

23 7. Anthropic has spent years of work and hundreds of millions of dollars developing the
24 information described in distilled format in the Tokens Formula Log, and, if disclosed, this
25 document could provide competitors with significant aid in bypassing Anthropic’s necessary
26 investment into effective data processing techniques and dataset mixes to potentially develop
27 similarly performant LLMs.

1 8. The Tokens Formula Log is also maintained strictly confidentially internally at Anthropic. It
2 is made available to fewer than two dozen employees outside the Legal team and the
3 Pretraining team (which is directly responsible for it). Every individual with access to the
4 Tokens Formula Log at Anthropic was granted access on a strict need-to-know basis and
5 individuals with access to it outside those groups include high-level decision makers such as
6 Jared Kaplan (our Chief Science Officer) and Dario Amodei (our Chief Executive Officer).

7 **C. The Scanned Books Dataset**

8 9. The Scanned Books Dataset is one of Anthropic's most valuable datasets because it reflects a
9 proprietary composition of books that Anthropic sourced, scanned, and created itself—which
10 is available to no other AI company in the world. The history of the creation of the Scanned
11 Books Dataset has been set forth in detail in two declarations by Anthropic's Vice President
12 of Product Partnerships Tom Turvey (Dkt. 123-2 and Dkt. 205-4). In short, it is a unique
13 dataset unavailable to any other AI company that was created by Anthropic via an
14 extraordinary effort and at the cost of tens of millions of dollars. Specific dollar amounts
15 reflecting the extent of this investment are provided in Mr. Turvey's Declaration of 3/27/2025,
16 which is subject to a pending motion to seal. Dkt. 123-2 at 22. If the Scanned Books Dataset
17 became available to Anthropic's competitors, they would be able to use it to train competing
18 LLMs without investing the tens of millions of dollars that Anthropic invested to create it.

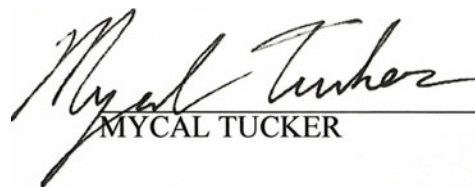
19 10. Further, the Scanned Books Dataset is also worth more than just the sum of its parts, as it
20 reflects the specific mix of books that Anthropic has curated and chosen to invest into. This is
21 because, as described in more detail in my prior Declaration (Dkt. 146-5) (*see* ¶¶ 4-10), the
22 value of a particular kind of book as part of the training corpus of Anthropic's LLMs depends
23 on a variety of factors, including the type/genre of the book, the language the book is written
24 in, its date of publication, and whether similar books are already in the training corpus, among
25 many other things. Based on these factors, Anthropic developed a valuation model, which
26 identifies how much Anthropic should be willing to pay for particular types of books. *See*
27 Declaration of Douglas A. Winthrop in Support of Defendant Anthropic PBC's Administrative
28

1 Motion to Seal, Unredacted Ex. 45, Dkt. No. 205-6 (“Valuation Model”). The composition of
2 the Scanned Books Dataset therefore also reflects Anthropic’s investment and research in
3 finding the right book data mix.

4 11. As with all of Anthropic’s training data, the Scanned Books Training Datasets are maintained
5 in a highly secure environment in restricted cloud-based S3 storage containers (S3 “buckets”).
6 If someone tried to export the Scanned Books Training Datasets from the S3 buckets (e.g., by
7 copying them to a local hard drive or thumb drive), even within Anthropic’s network,
8 Anthropic’s security infrastructure would prohibit that action. In practice, this means
9 Anthropic employees are effectively prohibited from having more than a small sample of the
10 Scanned Books Training Datasets on the hard drives of their local Anthropic laptops because
11 technical measures throttle any attempts to download more than a small portion of that data. It
12 is worth noting that the model weights for Anthropic’s LLMs, which are also widely regarded
13 as some of the most competitively sensitive information in the LLM industry, are protected in
14 the same way.

15 I declare under penalty of perjury under the laws of the United States of America that the
16 foregoing is true and correct. Executed on June 6, 2025, in San Francisco, California.

17
18
19
20
21
22
23
24
25
26
27
28



MYCAL TUCKER

DECLARATION OF JESSICA L. GILLOTTE

I, Jessica L. Gillotte declare:

1. I am an associate at Arnold & Porter Kaye Scholer LLP and am one of the attorneys representing Defendant Anthropic PBC (“Anthropic”) in this matter. I am an attorney duly licensed to practice law in the State of California. I make this declaration in support of Anthropic’s Response to Plaintiffs’ Letter Motion to Compel Production dated May 14, 2025.

The Tokens Formula Log

2. On March 18, Anthropic inadvertently produced the Tokens Formula Log with an incorrect confidentiality designation outside of the Source Code / Training Data inspection environment. The document was designated as Highly Confidential - Attorneys’ Eyes Only, but should have instead been designated as Highly Confidential - Source Code and/or Highly Confidential - Training Data, and should have instead been produced only in the training data and source code inspection environment (the “inspection environment”).

3. When Anthropic discovered the incorrect disclosure, it immediately informed Plaintiffs on April 1 that the Tokens Formula Log was “extremely sensitive” and that it was clawing back the three copies of the document produced with incorrect confidentiality designations pursuant to Section V.3 (“Inadvertent Failures to Designate”) of the Protective Order (Dkt. 62), as adopted by this Court (Dkt. 63). All three copies of the document were from the custodial files of one employee on the Pretraining Team. Anthropic further requested that Plaintiffs destroy its copies of the document, but clearly stated that the Tokens Formula Log would “be made available in the training data/source code review environment,” which is an 8-minute walk from the offices of Plaintiffs’ counsel according to Google Maps (Lieff Cabraser San Francisco Office).

4. On April 3, Plaintiffs emailed Anthropic and objected to its clawback request. However, Plaintiffs did not ask to set up a time to view the Tokens Formula Log in the training environment.

5. On April 4, Anthropic again asked Plaintiffs to “[l]et Anthropic know when [they] would like to go to the inspection environment to view [the Tokens Formula Log].” In the same

1 email, Anthropic also stated that it was “willing to discuss other reasonable accommodations that
2 will allow [Plaintiffs] to access a version of this document, for example, if [Anthropic] bring[s] a
3 secure copy of this document to certain technical depositions so [Plaintiffs] can question witnesses
4 about it.”

5 6. The parties met and conferred on numerous topics on April 8, April 15, April 28,
6 and May 12. During these meetings, Plaintiffs demanded Anthropic produce the Tokens Formula
7 Log outside the inspection environment. However, at no point did Plaintiffs ever ask to inspect the
8 Tokens Formula Log in the inspection environment. While the Inspection Protocol requires
9 Plaintiffs to give five days’ notice before an inspection, Anthropic has repeatedly waived that
10 requirement throughout the course of this litigation and is willing to continue to grant reasonable
11 waivers of that time limitation for purposes of inspection of the Tokens Formula Log.

12 7. On May 5, Anthropic further offered that it was willing to admit to the contents of
13 Request for Admissions (“RFA”) on shortened time in advance of the MSJ Hearing to confirm
14 that the training data mix for Anthropic’s non-commercial and aborted “Finch” model included
15 datasets associated with Books3, PiLiMi, and LibGen—consistent with the claimed need for this
16 information in the Tokens Formula Log, raised in Paragraph 4 of the Declaration of Collin
17 Fredericks in Support of Plaintiffs’ Opposition to Anthropic’s Motion for Summary Judgment
18 (Dkt. 153).

19 8. On May 9, Plaintiffs served three RFAs on precisely those topics. During a meet and
20 confer on May 12, Plaintiffs again demanded that Anthropic agree to produce the Tokens Formula
21 Log outside of the training data and source code inspection environment.

22 9. On May 13, Anthropic agreed to respond to Plaintiffs’ RFAs on an expedited basis
23 to confirm that data sourced from Books3/LibGen/PiLiMi were included in the training data mix
24 for Anthropic’s non-commercial and aborted model Finch, and to produce a redacted version of
25 the “Finch” portion of the Tokens Formula Log showing solely the names of primarily books
26 datasets, the number of tokens associated with those book datasets, the number of epochs for each
27
28

1 of those datasets, and the percentage of the total tokens in the total training mix that each of those
2 datasets make up, as listed in the portion of the Tokens Formula Log for the Finch model.

3 10. On May 14, Anthropic served its responses to the RFAs (admitting them), and
4 produced the requested redacted version of the Tokens Formula Log.

5 11. Multiple times in this litigation, Plaintiffs have inadvertently filed documents with
6 information designated Attorney's Eyes Only on the public docket. *See* Dkts. 86, 158. When this
7 occurs, documents containing information designated highly confidential are immediately emailed
8 by services like PacerPro and Docket Bird to all members of the public who have signed up for
9 these services for this case. This means that, if the Tokens Formula Log were filed on the public
10 docket, Anthropic's competitors would immediately obtain a copy before the clerk could lock the
11 docket. Because Anthropic's key competitors are involved in similar copyright litigation and
12 representatives from at least three leading industry competitors have been observed in the
13 courtroom gallery during hearings in this matter, it is likely that they are closely following the
14 docket and receiving filings in this case.

15 **Scanned Books Dataset**

16 12. In January 2025, Plaintiffs first demanded production of Anthropic's Scanned
17 Books Dataset outside the inspection environment. While Anthropic objected based on the
18 extremely sensitive and valuable nature of the dataset, it nonetheless agreed, as a compromise, to
19 produce metadata associated with its Scanned Books Dataset for its Claude 3.5 Haiku model (the
20 first model that was trained on the Scanned Books Dataset) outside the inspection environment.
21 Plaintiffs have also had access to the Scanned Books Dataset for Anthropic's Claude 3.5 Haiku
22 model in the inspection environment since March (Plaintiffs last conducted an inspection on April
23 11 after that data was made available).

24 13. In February 2025, Anthropic further agreed to produce metadata associated with all
25 of the books it had purchased and scanned, whether or not the books had been used to train any
26 model, and then did so on March 1. On March 25, Anthropic produced further scanned book
27 metadata when it released another Claude model trained on the Scanned Books Dataset (Claude
28

1 3.7 Sonnet). And Anthropic has made clear at all times that Plaintiffs can inspect its entire
2 Scanned Books Dataset in the inspection environment.

3 14. On May 8, notwithstanding the metadata Anthropic had produced and the
4 availability of the Scanned Books Datasets in the training data environment, Plaintiffs demanded
5 production of the entirety of Anthropic's Scanned Books Dataset outside the training date
6 inspection environment.

7 15. At the parties' May 12 meet and confer, Anthropic reiterated the need for the
8 Scanned Books Dataset to be inspected rather than produced as documents. At that time,
9 Anthropic made a further offer of compromise to produce a smaller exemplar of scanned books in
10 this dataset so that Plaintiffs' expert could use it off-site for the purposes of testing his hash
11 function. Plaintiffs rejected that offer. Anthropic nonetheless affirmed that it was still willing to
12 discuss this and other reasonable compromises short of producing the entire Scanned Books
13 Dataset. However, Plaintiffs have offered no compromise position and have only demanded the
14 wholesale production of the Scanned Books Dataset.

15 I declare under penalty of perjury pursuant to the laws of the State of California and the
16 United States of America that the foregoing is true and correct.

17 Executed this 6th day of June 2025 in San Francisco, CA.

18
19 /s/ Jessica L. Gillotte
JESSICA L. GILLOTTE

CERTIFICATE OF SERVICE

I, Joseph Farris, am the ECF user whose identification and password are being used to file the foregoing **ANTHROPIC’S RESPONSE TO PLAINTIFFS’ MOTION TO COMPEL AND DECLARATIONS OF MYCAL TUCKER & JESSICA GILLOTTE IN SUPPORT OF DEFENDANT ANTHROPIC’S RESPONSE TO PLAINTIFFS’ MOTION TO COMPEL.**

Dated: June 6, 2025

/s/ Joseph Farris

JOSEPH FARRIS