



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

ANDREA BARTZ, ANDREA BARTZ, INC.,
CHARLES GRAEBER, KIRK WALLACE
JOHNSON, and MJ + KJ, INC., individually and on
behalf of others similarly situated,

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

Case No. 3:24-CV-05417-WHA

Action Filed: August 19, 2024

**DECLARATION OF MYCAL TUCKER
IN SUPPORT OF DEFENDANT
ANTHROPIC PBC'S OPPOSITION TO
PLAINTIFFS' MOTION FOR CLASS
CERTIFICATION**

Judge: Honorable William H. Alsup

I, Mycal Tucker, hereby declare as follows:

1. I am a Research Scientist at Anthropic PBC (“Anthropic”) on the pretraining team. My duties in this role include developing new data processing methods and preparing data to train our commercial models.

2. I received a bachelor of science degree in Computer Science and Aeronautical and Astronautical Engineering from Massachusetts Institute of Technology (“MIT”) in 2015, and a master’s degree in Robotics from MIT in 2016. I received a PhD in Artificial Intelligence from MIT in 2024.

3. I have published papers at top conferences and journals in artificial intelligence and computational linguistics.¹ I have also taught undergraduate- and graduate-level courses at MIT, including Introduction to Artificial Intelligence and Introduction to Computational Psycholinguistics.

I. VALUE OF BOOKS

4. The value of a particular kind of book as part of the training corpus of Anthropic’s Large Language Models (“LLM”) depends on a variety of factors, including the type/genre of the book, the language the book is written in, its date of publication, and whether similar books are already in the training corpus, among many other things. As a general matter, we have found that books related to math, computer science, and science (“MCS”), when included in a model’s pretraining data corpus, tend to result in increased capability for the model to perform well on

¹ E.g., Mycal Tucker et al., *Trading off Utility, Informativeness, and Complexity in Emergent Communication*, in 36th Conference on Neural Information Processing Systems (NeurIPS 2022), available at https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb5f66371c7e4cbf6c223162c62c0f4-Paper-Conference.pdf; Mycal Tucker et al., *Emergent Discrete Communication in Semantic Spaces*, in 35th Conference on Neural Information Processing Systems (NeurIPS 2021), available at https://proceedings.neurips.cc/paper_files/paper/2021/file/5812f92450ccaf17275500841c70924a-Paper.pdf; Mycal Tucker et al., *What If This Modified That? Syntactic Interventions with Counterfactual Embedding*, 862–75, Findings of the Association for Computational Linguistics (Ass’n for Comput. Linguistics 2021), available at <https://aclanthology.org/2021.findings-acl.76.pdf>.

technical tasks. Because Anthropic values training models that are highly capable on technical tasks, this correlation means that we generally value MCS books more than some other categories of books.

5. To assess how adding different types of book data to the training corpus for an LLM affects its compute multiplier (“CM”), Anthropic performs what are called “ablation studies.”² See Ex. 35. In my ablation studies, I removed (or “ablated”) parts of Anthropic’s [REDACTED] and its [REDACTED] (defined below) to see how much different parts of the dataset contributed to the overall dataset performance.

6. [REDACTED]

7. [REDACTED]

²A compute multiplier is a technique, innovation, or data source that makes artificial intelligence training more efficient; it results in a model of the same capability being trained with less compute resources. For example, if a dataset has a compute multiplier of 2, that means one can train for half as much compute (*e.g.*, less data or a smaller model) but still achieve the same performance.

³ The “NLL” score in the document refers to Negative Log Likelihood, which is a standard measure for how accurately an LLM can predict a passage of text that was not part of its training corpus.

8. The benefit to model performance of adding any given type of book data to the training corpus will always depend on what kinds of data are already in the training mix. In other words, the incremental benefit of adding any particular type of data to a training corpus is relative. For example, if one model has been predominantly trained on non-technical data, adding high school-level technical books to that model's training data could be extremely valuable to improving its capabilities. At the same time, that same book data would be relatively less beneficial for increasing model capabilities if it were added to a training mix containing technical data (e.g., college-level technical content). In short, training on different types of book data appears to result in different competency improvements in the LLM, which is another reason that Anthropic values book data differently based on context.

9. Based on these factors, Anthropic developed a valuation model, which identifies how much Anthropic should be willing to pay for particular types of books. *See* Declaration of Douglas A. Winthrop in Support of Defendant Anthropic PBC's Administrative Motion to Seal, Unredacted Ex. 45, ECF No. 119-4 ("Valuation Model"). Anthropic's Valuation Model provides an estimate of our willingness to pay for access to certain numbers of tokens for different types of data (including for non-book training data). There is a subjective component to this analysis, and these values are not static over time. As Anthropic's LLM technology and training corpus develops, these valuations are likely to change in ways that are difficult to predict.

10. [REDACTED]

II. THE INTERNET BOOKS DATASETS

11. Early in its history, Anthropic acquired books from three sources on the Internet:

12. [REDACTED] Based on my review of source code records and the dataset itself produced in this case,⁴ Anthropic acquired a version of a dataset known as [REDACTED] (Anthropic's [REDACTED])

⁴ The source code record for the acquisition of [REDACTED] was produced as ANT_BARTZ_000436667.

[REDACTED]

[REDACTED]”) on or around February 25, 2021.⁵ While I cannot determine the exact date that Anthropic’s [REDACTED] was acquired, the source code implementing the download was written at this time, which means it is reasonably likely that it was acquired around that time.

13. There was no separate metadata file acquired in connection with Anthropic’s [REDACTED]. To my knowledge, Anthropic’s [REDACTED] does not contain any metadata other than a filename. And while its [REDACTED] does contain filenames, those file names were generated by unknown people outside of Anthropic. Anthropic has never relied on these file names to identify specific books in its [REDACTED].

14. Anthropic has used its [REDACTED] in the training corpus of certain commercial models.

15. [REDACTED] Based on my review of source code records and the dataset itself produced in this case,⁶ Anthropic acquired its [REDACTED] around June 8, 2021 to June 19, 2021.⁷ While I cannot determine the exact date that it was acquired, the source code implementing the BitTorrent protocol used to acquire it was written around this time, which means it is reasonably likely that it was acquired around that time.

16. Around the same time that Anthropic acquired its [REDACTED] Anthropic also acquired a separate metadata file related to [REDACTED]. Anthropic did not create this metadata. Anthropic has never relied on this metadata to identify specific books in its [REDACTED].

17. Anthropic has used its [REDACTED] in the training corpus of at least one research and development LLM that was never released as a commercial model. Anthropic also included the [REDACTED] in the training corpus of at least one LLM for which the training run was aborted

⁵ Anthropic’s name for this dataset is: [REDACTED]. Anthropic produced this dataset as ANT_BARTZ_000000005.

⁶ The source code record for the acquisition of Anthropic’s [REDACTED] was produced as ANT_BARTZ_000436670.

⁷ There are technically two [REDACTED]. Anthropic produced these datasets as ANT_BARTZ_000436661 and ANT_BARTZ_000436662. For convenience, I refer to them together as [REDACTED].

before the model completed pretraining (*i.e.*, there was never a complete LLM created). It has never been used in the training corpus of a commercial LLM.

18. [REDACTED] Based on my review of source code records and the dataset itself produced in this case,⁸ Anthropic acquired its [REDACTED] between July 8, 2022 and August 10, 2022.⁹ While I cannot determine the exact date that it was acquired, the source code implementing the BitTorrent protocol used to acquire it was written around this time, which means it is reasonably likely that it was acquired around that time.¹⁰

19. Around the same time that it acquired its [REDACTED], Anthropic also acquired a separate metadata file related to [REDACTED]. Anthropic did not create this metadata. Anthropic has never relied on this metadata to identify specific books in its [REDACTED].

20. To the best of my knowledge, Anthropic never used its [REDACTED] in the training corpus for any LLM that completed training. Anthropic did include the [REDACTED] in the training corpus of at least one LLM, but that training run was aborted before the model completed pretraining (*i.e.*, there was never a complete LLM created). It has never been used in the training corpus of a commercial LLM.

\\

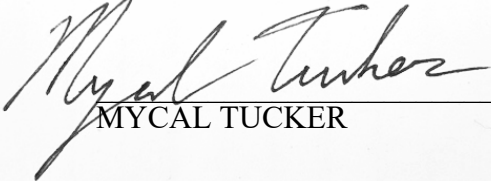
\\

⁸ The source code record for the acquisition of Anthropic's [REDACTED] was produced as ANT_BARTZ_000436671.

⁹ Technically, there are four [REDACTED] Anthropic produced these datasets as ANT_BARTZ_000436663, ANT_BARTZ_000436664, ANT_BARTZ_000436665, and ANT_BARTZ_000436666. For convenience, I refer to them together as [REDACTED].

¹⁰ There appear to have been four separate attempts to download the [REDACTED] including failed attempts.

I declare under penalty of perjury under the laws of the United States of America that the foregoing is true and correct. Executed on April 17, 2025, in San Francisco California.



MYCAL TUCKER



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

CERTIFICATE OF SERVICE

I, Douglas A. Winthrop, am the ECF user whose identification and password are being used to file the foregoing **DECLARATION OF MYCAL TUCKER IN SUPPORT OF DEFENDANT ANTHROPIC PBC’S OPPOSITION TO PLAINTIFFS’ MOTION FOR CLASS CERTIFICATION.**

Dated: April 17, 2025

/s/ Douglas A. Winthrop