

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

**UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA**

ANDREA BARTZ, CHARLES
GRAEBER, and KIRK WALLACE
JOHNSON, individually and on behalf of
others similarly situated,

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

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

**EXPERT DECLARATION OF BEN Y. ZHAO, PHD IN SUPPORT OF PLAINTIFFS
MOTION FOR CLASS CERTIFICATION**

Dated: March 27, 2025

A handwritten signature in black ink, appearing to be 'B. Zhao', with a long horizontal stroke extending to the right.

Ben Y. Zhao

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY’S EYES ONLY**

I. TABLE OF CONTENTS

II. QUALIFICATIONS, EXPERIENCE, AND COMPENSATION	3
III. SCOPE OF ENGAGEMENT	4
IV. SUMMARY OF OPINIONS.....	7
V. COMMON SOURCES AND ANALYSIS SHOW HOW, WHEN, AND WHERE ANTHROPIC OBTAINED COPYRIGHTED BOOKS	10

[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]
[REDACTED]	15

VI. COMMON SOURCES AND ANALYSIS SHOWS WHICH BOOKS ARE PRESENT IN ANTHROPIC’S DATASETS	17
VII. COMMON SOURCES AND ANALYSIS SHOWS HOW ANTHROPIC USED BOOKS DURING ITS TRAINING PROCESS.....	26
VIII. COMMON SOURCES AND ANALYSIS SHOWS HOW BOOKS CONTRIBUTE TO THE OPERATION AND PERFORMANCE OF ANTHROPIC’S LLMS	29

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

1. I, Ben Yanbin Zhao, Neubauer Professor of Computer Science at University of Chicago, submit this declaration in support of Plaintiffs' Motion for Class Certification.

II. QUALIFICATIONS, EXPERIENCE, AND COMPENSATION

2. I am the Neubauer Professor of Computer Science at the University of Chicago, a Fellow of the Association for Computing Machinery, and co-director of the UChicago Security, Algorithms, Networks, and Data (SAND) Lab. I teach and mentor undergraduate and PhD students in numerous courses related to artificial intelligence. Over the years, I have led numerous projects related to artificial intelligence, including projects funded by the National Science Foundation, Department of Defense, and industry research groups. Prior to starting my faculty career 21 years ago, I received my PhD from UC Berkeley, and my Bachelor of Science from Yale University.

3. My research lies at the intersection of machine learning and security, with additional topics covering data-driven systems and human computer interaction. My work is published regularly at the most selective publication venues in computer security and machine learning and has received numerous best paper awards, distinguished paper awards, most-influential paper awards and the USENIX Internet Defense Prize. My early work created the research area now known as distributed hash tables, and those systems are now used in large cloud computing systems today. More recently, I have spent extensive time researching machine learning and deep learning. My more recent projects have involved research related to the use of content by generative AI models. For my body of work, I have been honored as a Fellow of the Association for Computing Machinery (ACM), a TED/AI speaker, MIT Technology Review Young Innovator, and chosen by TIME Magazine as one of the AI/100, the 100 most influential people in AI. My 190 refereed publications have garnered roughly 38,000 citations. A detailed curriculum vitae showing my qualifications, experience, publications, awards and patents, as well as a list of cases

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

in which I have testified at trial, hearing, or by deposition within the last five years is attached to this report as Exhibit A.

4. I am being compensated for my time at the hourly rate of \$800 per hour. My compensation does not depend on my conclusions or the outcome of this case. I have no other interest in this litigation or the parties thereto.

III.SCOPE OF ENGAGEMENT

5. Counsel for Plaintiffs have engaged me to serve in this proposed class action litigation as an expert and to analyze, evaluate, and render opinions on (i) whether named plaintiffs Andrea Bartz, Kirk Wallace Johnson, Charles Graeber, MJ + KJ, Inc., and Andrea Bartz, Inc.'s copyrighted books are present in Anthropic's data sets; and whether there is class-wide evidence and/or a method of proof to determine: (ii) which copyrighted books are present in Anthropic's data sets; (iii) how Anthropic acquired the books in its data sets; (iv) how Anthropic used books during its LLM training process; and (v) how books contribute to the operation and performance of Anthropic's LLMs.

6. I understand that Anthropic, PBC is the Defendant in civil action No. 24-cv-5417. I understand that Andrea Bartz, Kirk Wallace Johnson, Charles Graeber, MJ + KJ, Inc., and Andrea Bartz, Inc. are the Plaintiffs.

7. I have reviewed the First Amended Class Action Complaint ("Complaint") filed in this action. *See* Dkt. 70, *Bartz et al v. Anthropic PBC*, 3:24-cv-5417, (N.D. Cal. December 4, 2024). Plaintiffs Andrea Bartz, Kirk Wallace Johnson, Charles Graeber, MJ + KJ, Inc., and Andrea Bartz, Inc., on behalf of themselves and other similarly situated ("Plaintiffs"), brought this case alleging that Anthropic illegally acquired and illegally trained its LLMs using Plaintiffs' copyrighted books.

8. In their complaint, Plaintiffs allege that Anthropic sought out known repositories of pirated books in order to acquire training data for their LLMs. Those repositories included

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY’S EYES ONLY**

“Books3;” a repository which its creator claimed contained “196,640 books.” *Id.* ¶ 36-37. Those books were sourced from “bibliotik” a notorious “private tracker”—that is a private torrent site for experienced digital pirates. *Id.* ¶ 38. Books3, in turn, was included in a large set of training data released to the public in 2020 known as “The Pile.” *Id.* ¶ 34-42. Soon after its founding, Anthropic used “The Pile” to train a large language model. *Id.* ¶ 43; [REDACTED]

9. Based on documents and discovery responses provided by Anthropic in this case, I understand that Anthropic has obtained books data from other sources in addition to “[REDACTED]” Those sources include the [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

11. I understand Plaintiffs seek certification of the copyright infringement claim on behalf of proposed classes who were allegedly harmed by Anthropic’s conduct and appointment as class representatives.

12. I understand the proposed classes are defined as follows:

Pirated Books Class: All natural persons, estates, literary trusts, and loan-out companies that are legal or beneficial owners of copyrighted works that: (a) were registered with the United States Copyright Office within five years of the work’s publication; (b) were registered with the United States Copyright Office before being downloaded by Anthropic, or within three months of

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

publication; (c) are assigned an International Standard Book Number (ISBN) or Amazon Standard Identification Number (ASIN); and (d) were downloaded by Anthropic as part of [REDACTED]

[REDACTED] **Books Class:** All natural persons, estates, literary trusts, and loan-out companies that are legal or beneficial owners of copyrighted works that: (a) were registered with the United States Copyright Office within five years of the work's publication; (b) were registered with the United States Copyright Office within three months of publication or before being [REDACTED]; (c) are assigned an International Standard Book Number (ISBN) or Amazon Standard Identification Number (ASIN); (d) were [REDACTED]; and (e) were used by Anthropic in LLM training.

13. To perform my analysis—and in addition to drawing from my experience as discussed above and in Exhibit A—I have reviewed and analyzed extensive information and documents produced by Anthropic through discovery in this action. the parties' various pleadings, disclosures, and discovery responses; deposition transcripts; and various documents produced in this case. A list of the materials that I have considered in reaching my opinions is attached as Exhibit B to this report.

14. This report is based on information currently available to me. I reserve the right to supplement my analysis in this report in response to any reports prepared on behalf of Anthropic. I also reserve the right to amend or supplement my opinions based on further discovery and information provided in this case, including any additional depositions or other disclosures that may occur after I submit my report.

15. All of the opinions stated herein are based on my own personal knowledge, professional judgment, and analysis of the materials and information that I considered in preparing this report. I reserve the right to create any additional summaries, tutorials, demonstrations, charts, drawings, tables, and/or animations that may be appropriate to supplement and demonstrate my opinions at trial, deposition, or any hearing at which I may be called to testify.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY’S EYES ONLY**

IV. SUMMARY OF OPINIONS

16. My full opinions are set forth in the body of this report, as a result of my analysis of the information and documents I have considered I have formed the following opinions relating to the Plaintiffs’ motion for class certification:

a. Common sources and analysis show how Anthropic obtained any given copyrighted book—including those of the named Plaintiffs:

- i. Book piracy: Anthropic has amassed three datasets from pirated sources: [REDACTED] [REDACTED] which itself was sourced from the [REDACTED] [REDACTED]. Common evidence shows how and when Anthropic obtained these datasets. [REDACTED], it is a single “download” which Anthropic acquired from a single source. As for the [REDACTED] datasets, [REDACTED] [REDACTED]
- ii. [REDACTED]: In 2024, Anthropic began [REDACTED], including many books it had earlier obtained from [REDACTED] datasets, and [REDACTED] to use in its LLM research, development, training, and commercialization efforts. Common evidence, [REDACTED], shows how Anthropic obtained [REDACTED]

b. Common sources show which books are present in Anthropic’s datasets.

- i. For any given copyrighted work, including the Named Plaintiffs’ works, a simple process permits accurate identification of a given work in a given training dataset.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

- ii. First, for many works, including any which Anthropic has admitted to [REDACTED] the inquiry simply involves checking the title, author, and/or ISBN of the work against the lists which Anthropic has already admitted are present in its data sets.
 - iii. For the remainder (and as a further robustness check on the former category), I describe a simple algorithm which can take text as an input and perform a match against the text present in Anthropic's datasets.
 - iv. Using both approaches (as well as Anthropic's admissions) I have confirmed that the Named Plaintiffs' books were located in the [REDACTED] and in the set of [REDACTED]
[REDACTED]
- c. **Common sources and analysis show how Anthropic used books during its training process.**
- i. [REDACTED]
[REDACTED]
[REDACTED]. See Anthropic's Response to Interrogatory No. 3. [REDACTED]
[REDACTED]
[REDACTED]
[REDACTED], there is a

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

common set of facts and sources regarding how the class members' works are used in training Anthropic's LLMs.

d. Common sources and analysis show how books contribute to the operation and performance of Anthropic's LLMs.

- i. As common evidence from [REDACTED] and third-party research demonstrates, the performance of Anthropic's LLMs greatly benefit from [REDACTED] books.
- ii. This is for at least three reasons: i) books are an excellent source of [REDACTED] which enables LLMs to produce high-quality text; ii) books are a great source of long-form text that enable LLMs to process coherent long-form text as an input and produce it as an output; and iii) books provide expressively varied text that enable LLMs to produce text in a wide-array of genres.
- iii. While it is possible that some of the positive qualities may not be present for every book in the proposed classes, [REDACTED] and third-party research have repeatedly evaluated books in a holistic manner that focuses on the quantity of books, rather than the specific attributes of a particular book.

17. The above and following represents my current opinions based on the information I possess at this time; I reserve the right to amend and/or supplement this report in accordance with applicable Court rules, orders, and procedures.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

**V. COMMON SOURCES AND ANALYSIS SHOW HOW, WHEN, AND WHERE
ANTHROPIC OBTAINED COPYRIGHTED BOOKS**

18. I am aware that Anthropic acquired books to train its LLMs from at least: [REDACTED]

[REDACTED]

[REDACTED] Given that several of these datasets have only been recently produced, I reserve the right to supplement my opinions regarding the below. However, I have analyzed each of these common sources, and I will discuss each one in turn.

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

A series of horizontal black bars of varying lengths, some solid and some with yellow segments, representing redacted information.

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

Page 10 of 10

[illegible]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

The image displays a series of horizontal black bars of varying lengths, some with yellow highlights, representing a redacted document. The bars are arranged in a vertical sequence, with some bars having small yellow rectangular highlights at their ends or within their segments. The overall pattern suggests a document where sensitive information has been obscured by black redaction marks.

¹ These estimates are based on the number of unique files present in the datasets.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[illegible]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[illegible]

⁴ This estimate is based on the number of unique files present in the dataset.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

A series of horizontal black bars of varying lengths, representing redacted text. The bars are stacked vertically, with the longest bars in the middle and shorter bars at the top and bottom. The total height of the redacted area is approximately 400 pixels.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

A series of horizontal bars, mostly black, with some yellow segments interspersed, representing a data visualization. The bars vary in length and are arranged in a somewhat irregular, staggered fashion. The yellow segments are small and appear as highlights within the black bars.

⁵ See ¶ 45 below for information regarding these Bates numbers.

[illegible]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

**VI. COMMON SOURCES AND ANALYSIS SHOWS WHICH BOOKS ARE
PRESENT IN ANTHROPIC'S DATASETS**

42. As explained above, Anthropic possesses several datasets [REDACTED] which it has used in its research and development efforts, including in training its LLMs. *See supra* ¶¶ 18-41. [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

⁸ The estimate is made using the number of metadata lines in the file, each of which appears to correspond to a book.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]		[REDACTED]	
[REDACTED]		[REDACTED]	
[REDACTED]		[REDACTED]	
[REDACTED]		[REDACTED]	
[REDACTED]		[REDACTED]	
[REDACTED]		[REDACTED]	

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]

[REDACTED]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

		
		

















52. If this information is not present for a particular book, the “Book Hash” search algorithm described below provides a comprehensive and accurate alternative approach to identifying the presence of any given text in Anthropic’s datasets.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

53. Approach 2: The Book Hash Approach: The book search algorithm described in the following paragraphs is capable of conclusively determining whether any given book is present in any given subset of data.¹⁰

54. This approach relies on so-called “one-way hashing algorithms,” or “hash codes.” One way hashing algorithms, also called “collision-resistant hashes,” are a decades-old tool used in computer security, including to secure critical internet infrastructure, such as verifying the sender of a signed secure email message and verifying that a file downloaded from the Internet was received without tampering or corruption. *See, e.g.*, “Hashing in Cybersecurity,” CrowdStrike, Jan. 16, 2024 (available at: <https://www.crowdstrike.com/en-us/cybersecurity-101/data-protection/data-hashing/>). These hash codes are series of binary numbers that are compact, yet are capable of uniquely representing a particular piece of data, i.e. a sentence or text excerpt from a book.

55. Secure one-way hash code generators, e.g. the standard “Secure Hash Algorithm-256” (“SHA-256”) algorithm, accept any input (including text). *See* “FIPS PUB 180-4: Secure Hash Standard. Federal Information Processing Standards Publication 180-4,” National Institute of Standards and Technology: U.S. Department of Commerce (March 2012), <http://csrc.nist.gov/publications/fips/fips180-4/fips-180-4.pdf>. The hash-code generators then run that input through a series of mathematical functions resulting in a fixed length—in the case of SHA-256, a 256 bit—binary code that, with exceedingly high probability, can only be generated

[REDACTED]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY’S EYES ONLY**

by that particular input. *Id.* at 21. This process of running an input text through the hash-code generator is called “hashing” the input.

56. In the case of the SHA-256 algorithm, the probability of a particular hash code for any text input also matching the hash code for some non-identical text input is roughly $1/(2^{128})$, or roughly 2.93 times 10 to the -39th power. *See* “Recommendation for Applications Using Approved Hash Algorithms,” National Institute of Standards and Technology, p. 8 (Aug. 2012), <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-107r1.pdf>. Statistically speaking, a person is more likely to be hit by 7 meteorites, *see* “How Likely are You to Get Hit by a Meteor,” *Christian Science Monitor*, (Feb. 09, 2016), <https://www.csmonitor.com/Science/Spacebound/2016/0209/How-likely-are-you-to-get-hit-by-a-meteor>, than to randomly find two different text inputs that produce the same hash code.¹¹

57. The uniqueness of hash-codes is a useful feature. For the book search algorithm, it means that any given piece of text can be hashed to a unique value and compared against the hashes of Anthropic’s datasets. Where the two hash-codes match, that means the underlying text is identical and thus present in the dataset under investigation.

58. As an example of this approach which I believe to be robust and accurate, I have provided the following outline of how such an algorithm would function:

59. First, the text of the book under investigation is separated into distinct sentences. Next, special characters (such as unicode characters for, e.g., umlauts), and empty lines are removed.

60. Of the remaining set of sentences, and in order to guard against the prevalence of common, short length, sentences, e.g. “Good Morning!”, any sentences of fewer than ten words

¹¹ Comparison calculated as $\ln(1/2^{128})/\ln(1/250000)$.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

are removed. At the end of this process, the book under investigation will be typically reduced to a few thousand longer sentences.

61. Next, a unique hash code is computed for each of these longer sentences. The result is a set of hash codes that uniquely identifies these long sentences from the book in question.

62. Once that set of unique, long-sentence hash codes for the test book is created, the same process is performed with the data in question. The data is cleaned in the same way as the test book data is cleaned and unique hash codes are computed for the longer sentences.

63. Finally, a simple comparison checks the hash codes of the test book against the hash codes of the data to confirm the presence or absence of the book.

64. To guard against the possibility of data files that summarize or even quote portions of a book, but which are not full copies, the algorithm takes the test book hash codes, divides them into equal subsets, and checks the matching rate for each subset of hash codes in the data. This is equivalent to performing a check that different portions of the test book are equally present in the data. If each subset of test book hash codes are found within the data at approximately the same overall rate as the book's hash codes as a whole, then the book—rather than a summary or lengthy quotation—is present within the data.

65. To the extent there is any remaining doubt, a final robustness check may be performed which simply involves a viewing the portion of Anthropic's data identified alongside the book in question to verify the presence of the book in Anthropic's data. Example excerpts of the results of this matching visual inspection are included as Exhibit D to my report and the full comparisons are being produced alongside this report. Both the code to the Book Hash search algorithm and the code to create the comparisons are being produced alongside this report.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

66. The following table summarizes the results of the Book Hash search for Plaintiff's works on Anthropic's [REDACTED]. The full tables are included as Exhibit C to this report:

	2019	2020	2021
1. Overall	100	100	100
2. Gender			
3. Age			
4. Ethnicity			
5. Income			
6. Education			
7. Marital Status			
8. Religion			
9. Occupation			
10. Health Status			
11. Living Arrangements			
12. Travel History			
13. Employment Status			
14. Home Ownership			
15. Vehicle Ownership			
16. Insurance Coverage			
17. Charitable Contributions			
18. Political Affiliation			
19. Volunteer Work			
20. Substance Use			
21. Mental Health			
22. Physical Health			
23. Stress Levels			
24. Life Satisfaction			
25. Overall Well-being			

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[REDACTED]	[REDACTED]	[REDACTED]
------------	------------	------------

67. In summary, this book matching algorithm takes unique sentences of substantial length from a test book, computes hash codes that uniquely represent those sentences, then matches those hash codes against the hash codes generated from the data file. Where there is a substantial overlap, the book is present in the data.

B. Named Plaintiffs Works are Present in Anthropic's Datasets

68. Using the above approaches I have confirmed that the works by Named Plaintiffs listed in the table below are present in Anthropic's datasets, additional detail regarding the presence of the works in the pirated datasets is presented in Exhibit C to my report:

[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]
[REDACTED]	[REDACTED]	[REDACTED]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

—		
—		
—		
—		
—		

VII. COMMON SOURCES AND ANALYSIS SHOWS HOW ANTHROPIC USED BOOKS DURING ITS TRAINING PROCESS

69. Class-wide evidence and analysis are capable of showing how Anthropic used books during its training process. At this time I have not had opportunity to inspect Anthropic's pretraining source code or benefit from additional discovery as to Anthropic's training process. I thus reserve the right to supplement the following as discovery proceeds. However, Anthropic itself has stated [REDACTED]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[illegible]

A series of 15 horizontal bars of varying lengths, mostly black, with one bar in the middle containing a yellow and black checkered pattern. The bars are arranged in a vertical stack, with the checkered bar being the only one of its kind. The other bars are solid black and vary in length, with some being nearly full-width and others being significantly shorter. The checkered bar is positioned in the middle of the stack, between the 7th and 8th bars from the top. The checkered pattern consists of alternating yellow and black squares, with the yellow squares being slightly larger than the black squares. The bars are set against a white background.

[REDACTED]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

VIII. COMMON SOURCES AND ANALYSIS SHOWS HOW BOOKS CONTRIBUTE TO THE OPERATION AND PERFORMANCE OF ANTHROPIC’S LLMs

79. As explained above, Anthropic used [REDACTED]. *see supra* ¶¶ 68-78. This unsurprisingly resulted in significant benefits to Anthropic’s LLMs that is supported by common evidence from Anthropic and third-party research.

80. To begin with, in numerous internal documents and transcripts of internal meetings, Anthropic has repeatedly acknowledged that [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

81. Indeed, Anthropic has further acknowledged [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

82. As described below, the performance of Anthropic's LLMs greatly benefit from [REDACTED] for at least three reasons.

83. First, [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

84. [REDACTED] other leading AI labs, researchers, and academics. For example, researchers at OpenAI have stated that while webscrapes are a "promising source of diverse and nearly unlimited text . . . they have significant data quality issues"

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY'S EYES ONLY**

and describing a filtration process emphasizing “document quality.” *See* “Language Models are Unsupervised Multitask Learners,” Radford et al, OpenAI, April 2019, p. 3 (available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf).¹⁵ In 2021, researchers at Google, working to build a “high-quality 1.6 trillion token dataset,” stated that while “[w]eb pages constitute the vast quantity of data in this unlabelled corpus, [] their quality ranges from professional writing to low-quality comment and forum pages.” *See* “More Efficient In-Context Learning with GLaM,” Andrew Dai, Nan Du, Google Research, December 2021 (available at: <https://research.google/blog/more-efficient-in-context-learning-with-glam/>). To remedy this issue, the researchers “developed a text quality filter that was trained on a collection of text from Wikipedia and books (both of which are generally higher quality sources) to determine the quality of the content for a webpage” and “applied this filter to generate the final subset of webpages and combined this with books and Wikipedia to create the final training dataset.” *Id.* Google’s findings mirrored those of Microsoft researchers who wrote in the 2020 release of their Turing-NLG model “that the bigger the model and the more diverse and comprehensive the pretraining data, the better it performs at generalizing to multiple downstream tasks even with fewer training examples.” *See* “Turing-NLG: A 17-billion-parameter language model by Microsoft,” Rosset, Cory, Microsoft Research, Feb. 13, 2020 (available at: <https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/>). And, in the 2022 paper announcing Microsoft’s follow-up Megatron-Turing NLG model, Microsoft and Nvidia researchers similarly

¹⁵ Another OpenAI researcher has similarly concluded that dataset quality is what matters most in training AI models, writing that “model behavior is not determined by architecture, hyperparameters, or optimizer choices. It’s determined by your dataset, nothing else. Everything else is a means to an end in efficiently deliver[ing] compute to approximating that dataset.” *See* <https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/>.

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY’S EYES ONLY**

filtered their webscraped documents for “high-quality” by training a classifier to label documents most similar to “OpenWebText2, Wikipedia, and Books3,” as well as using Books3 itself as part of the training dataset. *See* “Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model,” Smith et al. (Microsoft & Nvidia), Feb. 4, 2022 p. 9 (available at: <https://arxiv.org/pdf/2201.11990>). The authors noted that “[c]areful processing of high-quality, high-volume and diverse datasets directly contributes to model performance in downstream tasks.” *Id.* at 2.

85. Second, [REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED] As a 2024 research paper from Princeton University explains, “books are good sources of long-context data” and “are more broadly beneficial for in-context learning, summarization and re-ranking.” “How to Train Long-Context Language Models (Effectively),” Gao et al., p. 5, Oct. 3, 2024, <https://arxiv.org/pdf/2410.02660>. Other researchers—including those affiliated with the organization that compiled and released The Pile—similarly find that “increased training dataset diversity improves general cross-domain knowledge and downstream generalization capability for large-scale language models” and that books “are invaluable for long-range context modeling research and coherent storytelling.” “The Pile: An 800GB Dataset of Diverse Text for Language Modeling”, Gao et. al, EleutherAI, pp. 1, 4, Dec. 2020, <https://arxiv.org/pdf/2101.00027>).¹⁶

¹⁶ Indeed, one of the authors of “The Pile” expressly stated that quality concerns drove the exclusion of social media data from Reddit and Twitter in their datasets. *See* Reddit Post by

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY’S EYES ONLY**

87. Intuitively, few online sources provide coherent, curated long form text similar to that of published books. The use of such text during training are critical to performance on LLM tasks that require longer contexts, including large scale text summarization and generation of long text documents (long essays, articles or books). *Id.*

88. Third,

[REDACTED]

StellaAthena (Stella Biderman, Pile Co-Author) “The purpose of the Pile is to combine diverse datasets that are of higher quality than “generic internet text.” We felt that adding Twitter or Reddit posts wouldn’t advance that agenda much.”

<https://www.reddit.com/r/MachineLearning/comments/kokk8z/comment/ghvgoq8/>; Ex. 53
<https://github.com/stellaathena> (Github profile belonging to Stella Biderman titled “StellaAthena”).

**CONTAINS MATERIAL DESIGNATED
HIGHLY CONFIDENTIAL ATTORNEY’S EYES ONLY**

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

[REDACTED]

90. In sum, [REDACTED]

books as a class of training data for LLMs possess significant and common value. These values are crucial to the operation and performance of [REDACTED]. While it is possible that some of the positive qualities described above may not be present for every book in the proposed classes, [REDACTED] repeatedly evaluated books in a holistic manner that focuses on the quantity of books given the data and rules-of-thumb indicating their value to LLM training, rather than the specific attributes of a particular book. These same factors regarding the quality and value of books as a source of training data supported [REDACTED]. Thus, the benefits of books to Anthropic’s LLMS are “common” in the sense of applying generally to the books that comprise the proposed classes.