

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

**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'
OPPOSITION TO ANTHROPIC'S MOTION FOR SUMMARY JUDGMENT**

Dated: April 24, 2025

A handwritten signature in black ink, appearing to be 'B. Zhao' with a stylized flourish at the end.

Ben Y. Zhao

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

I. TABLE OF CONTENTS

II. QUALIFICATIONS, EXPERIENCE, AND COMPENSATION 3

III. SCOPE OF ENGAGEMENT 4

IV. SUMMARY OF OPINIONS..... 6

V. BACKGROUND ON ANTHROPIC’S BOOKS DATASETS..... 9

 A. Books3 9

 B. LibGen..... 9

 C. Pirate Library Mirror (“PiLiMi”)..... 11

 D. Book Scanning 12

 E. Plaintiffs’ Books Are Present in Each of Anthropic’s Datasets 13

VI. ANTHROPIC FAILS TO OFFER A COMPREHENSIVE, ACCURATE, OR TESTABLE ACCOUNT OF THE LLM TRAINING PROCESS, AND THUS OBSCURES THE RELEVANT FACTS ABOUT THIS PROCESS 14

VII. LLM TRAINING CAN BE UNDERSTOOD AS ENCODING TRAINING DATA THROUGH COMPRESSION..... 20

VIII. ANTHROPIC HIGHLY VALUED BOOKS AS TRAINING DATA FOR ITS LLMS 28

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

1. I, Ben Yanbin Zhao, Neubauer Professor of Computer Science at University of Chicago, submit this declaration in support of Plaintiffs' Opposition to Anthropic's Motion for Summary Judgment.

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 UNDER THE PROTECTIVE ORDER

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 \$850 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 (a) to explain the nature of the LLM training process (including the background of the books data sources and the inclusion of Plaintiffs' books within them); (b) to review Anthropic's discussion of training and how LLMs work in its Motion for Summary Judgment and the Declaration of Anthropic Chief Science Officer Jared Kaplan to determine if Anthropic's statements were or were not supported (and, if not supported, to explain why not); and (c) to explain the value of books to the training and operation 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 "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

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

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; *see supra* § ¶ 19.

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 “Books3.” Those sources include the so-called “shadow libraries” “Library Genesis” (“LibGen”) and “Pirate Library Mirror” (“PiLiMi”) which hosted a mirror of the pirated books library Z-library (“Z-lib”). *See supra* §§ V. I understand that LibGen has been repeatedly ordered shut down by the United States District Court for the Southern District of New York due to copyright infringement. *See id.* I understand that Anthropic has used to LibGen to train a large language model. *See id.* I understand that Anthropic used the Pirate Library Mirror data in LLM training. *See id.* Anthropic has also obtained books by organizing and operating a massive book acquisition and scanning operation. *See id.*

10. I understand that Anthropic has moved for summary judgment on the grounds that its use of Plaintiffs works in training its LLMs is “fair use.” *See* Anthropic’s Motion for Summary Judgment, Dkt. 122.

11. 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.

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

12. 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.

13. 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.

IV. SUMMARY OF OPINIONS

14. 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 Anthropic's Motion for Summary Judgment:

a. Anthropic acquired and has used three sets of pirated books—Books3, Libgen, and Pirate Library Mirror—and a large volume of scanned books.

i. Plaintiffs' books are present in these datasets.

b. Anthropic Fails to Provide a Comprehensive and Accurate Account of the Training Process.

i. Anthropic's summary judgment motion and the supporting factual declaration of its Chief Science Officer Dr. Jared Kaplan characterize the LLM training process in broad strokes and using

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

non-specific language that obscures and thus misstates what is actually occurring during LLM training.

- ii. Further undermining Anthropic’s position is that its conclusions are not testable and they do not account for contrary evidence.

c. The LLM Training Process Can be Understood as Lossily Compressing the Training Data.

- i. At this time, and in light of the fact that discovery is ongoing, my declaration focuses on demonstrating one crucial example of how Anthropic overlooks contrary explanations in its description of LLM training and the resulting LLMs.
- ii. In particular, the scale of LLM memorization—which Anthropic’s internal documents admit but which Anthropic’s motion does not explain—is best understood as a result of the training process. Third-party research and Anthropic’s own documents and source code demonstrate that the LLM training process can be characterized as a process of compressing its training data. Specifically, LLM training involves optimizing the model’s “neurons” such that they encode a lossily compressed representation of training data.
- iii. This understanding is demonstrated by multiple sources that in my professional opinion are individually and collectively reliable, and that comport with my own research and background. These sources are: (1) independent research from leading AI labs, including

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

Google, (2) the use of loss functions—i.e. measures of compression efficiency—to steer the model training process (as discussed more below), (3) Anthropic’s documents stating that its models memorize large fractions of their data, and (4) Anthropic’s copyrighted output prevention system, which deploys large quantities of pirated books from Libgen as the “check set” for potentially copyrighted output.

d. Books are Extremely Valuable as Training Data and Contribute to the Functioning of Anthropic’s LLMs.

- i. Evidence from Anthropic and third-party research demonstrates, the performance of Anthropic’s LLMs greatly benefit from training on a large corpus of books and that books are one of (if not the) highest quality source of training data.
- ii. This is for at least three reasons: i) books are an excellent source of “good writing,” 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.

15. Thus, I disagree with the statements of Anthropic and its Chief Science Officer about the value of books and how LLMs work because, in my professional opinion, the characterizations are contradicted by materials in this case, by research, and by principles of science (which require testing and explanation).

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER**V. Background on Anthropic’s Books Datasets**

16. I am aware that Anthropic acquired books to train its LLMs from at least: (1) Books3, (2) LibGen (also known as “Library Genesis”), (3) “Pirate Library Mirror” (also known as “PiLiMi”), and (4) its Book Scanning Project.

A. Books3

17. I understand that Books3 is a dataset of pirated books created by Shawn Presser and made available by EleutherAI as part of a larger dataset called The Pile. *See* Tweet from Shawn Presser, Oct. 25, 2020, <https://x.com/theshawwn/status/1320282149329784833?lang=en>. Mr. Presser claimed that Books3 contained 196,640 books.

18. Anthropic has used at least one books dataset sourced from Books3, called pile.books3_source to train models. *See* Anthropic’s First Supplemental Response to Interrogatory No. 2.

19. Anthropic admits that it used books from Books3 in LLM training, including, in Anthropic’s words, “at least one commercially released and generally accessible Claude model.” *See* Anthropic’s Response to Request for Admission No. 10. At least as of May 2024, Anthropic continued to use “the books3 dataset from the pile” in its LLM training work, listing it as “likely our most favoured general prose loss eval.” ANT_BARTZ_000209219.

B. LibGen

20. I understand that LibGen, also known as “Library Genesis,” is a website providing access to many pirated copyrighted books, which has been repeatedly ordered to shut down for violating copyright law by United States federal courts. *See Elsevier Inc. v. Sci-Hub*, No. 15-CV-4282 (RWS), 2017 WL 3868800 (S.D.N.Y. June 21, 2017); *Cengage Learning, Inc. v. Library Genesis*, No. 23-cv-8136 (CM), Dkt. 36 (S.D.N.Y. Sept. 24, 2024).

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

21. Anthropic has used at least two books datasets sourced from LibGen, called libgen/fiction and libgen/nonfiction to train a model. *See* Anthropic’s Supplemental Response to Interrogatory No. 4. Anthropic has produced these datasets as ANT_BARTZ_000436661 and ANT_BARTZ_000436662, respectively. Based on my analysis of the datasets Anthropic produced both libgen/fiction and libgen/nonfiction appear to contain millions of books.¹

22. Anthropic acquired its LibGen datasets by torrenting them. *See* Anthropic’s Supplemental Response to Interrogatory No. 16. Torrenting is a decentralized method of peer-to-peer file sharing, often associated with the unauthorized downloading and distributing of pirated content over the BitTorrent protocol. *See* Office of the United States Trade Representative, 2022 Review of Notorious Markets for Counterfeiting and Piracy, p. 4, [https://ustr.gov/sites/default/files/2023-01/2022%20Notorious%20Markets%20List%20\(final\).pdf](https://ustr.gov/sites/default/files/2023-01/2022%20Notorious%20Markets%20List%20(final).pdf). Anthropic has produced the source code it used to torrent the LibGen datasets as ANT_BARTZ_000436670. *See* ANT_BARTZ_000436670 (file named “code.libgen.tar.bz2” listing the three sets of torrents as “libgen_torrents.tar.gz” “libgfen_fiction_torrents.tar.gz” and “scimag_torrents.tar.gz” and commenting the code with “# Get torrent files”).

23. Anthropic admits that it used books from LibGen in LLM training. *See* Anthropic’s Supplemental Response to Request for Admission No. 11; Anthropic’s Supplemental Response to Interrogatory No. 4.

24. Anthropic also uses LibGen to monitor the continuous output of Claude to look for regurgitated copyrighted material. *See* Kaplan Decl. at ¶ 68; Kaplan Dep. Tr. at 93:20-94:6; *id.* at 186:11-14; *id.* at 247:3-7; *id.* at 249:6-13.

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

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER**C. Pirate Library Mirror (“PiLiMi”)**

25. I understand that “Pirate Library Mirror,” also known as “PiLiMi,” is a website offering—or which in the past offered—access to many pirated books. In July 2022, Ben Mann, co-founder of Anthropic, messaged his colleagues over slack the link to the Pirate Library Mirror website and appended the message “just in time!!” *See* ANT_BARTZ_000391318.

26. In July 2022, Pirate Library Mirror explained that the “focus of this project is illustrated by its name.” *See* “Pirate Library Mirror” homepage, archived by Wayback Machine from July 3, 2022 (available at <https://web.archive.org/web/20220703220153/http://pilimi.org/>). First, “Pirate” because “[w]e deliberately violate the copyright law in most countries.” *Id.* Second, “Library” because “[w]e focus primarily on written materials like books.” *Id.* And third, “Mirror” because “[w]e are strictly a mirror of existing libraries.” *Id.* Pirate Library Mirror first “mirrored” (i.e., provided access to the same books as) Z-Library, which Pirate Library Mirror described as “a popular (and illegal) library.” *Id.*

27. Anthropic has used at least two books datasets sourced from Pirated Library Mirror, called pilimi-complete-non-en and pilimi-complete-en in training LLMs. *See* Anthropic’s Supplemental Response to Interrogatory No. 4; Anthropic’s Supplemental Response to Request for Admission No. 12.

28. Anthropic has produced four Pirate Library Mirror datasets, which appear to be versions of the pilimi datasets Anthropic used in its training work: pilimi as ANT_BARTZ_000436663; pilimi_full as ANT_BARTZ_000436664; pilimi_full2 as ANT_BARTZ_000436665; and pilimi_complete as ANT_BARTZ_000436666. Based on my analysis of the datasets Anthropic produced, pilimi_complete appears to include all the unique file

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

ids present in the other pilimi datasets and would thus include all or nearly all of the books also present in the other datasets. I estimate pilimi_complete contains more than 2 million books.²

29. Anthropic acquired its Pirated Mirror Library datasets by torrenting them. *See* Anthropic’s Supplemental Response to Interrogatory No. 16. Torrenting is a decentralized method of peer-to-peer file sharing, commonly associated with the unauthorized downloading and distributing of pirated content over the BitTorrent protocol. *See supra* ¶ 22. It appears Anthropic has produced the source code it used to torrent the Pirate Library Mirror datasets as ANT_BARTZ_000436671. *See* ANT_BARTZ_000436671 (file names “code.pilimi.tar.bz2” listing e.g. “pilimi_torrent_files/pilimi-zlib-all.zip” as among the files to torrent).

D. Book Scanning

30. I understand that Anthropic also purchases and scans books for use in LLM training, a project it refers to as its “Book Scanning Project,” or “Project Panama.” *See* Sangani Dep. Tr. at 86:7–8 (“My understanding is that Claude 3.5 Haiku was trained on books from the book-scanning project.”); Ex. 3 to Sangani Dep.; ANT_BARTZ_000005492 (“Project Panama is our effort to destructively scan all the books in the world.”); Anthropic’s First Supplemental Response to Interrogatory No. 4 (noting that Anthropic’s “most recent commercially released and generally accessible Claude model (Sonnet 3.7) was trained on wombat_staging.panama_haiku_stem [and] wombat_staging.panama_haiku_non_stem.”). Based on discovery so far, I understand that Anthropic has used three training datasets of scanned books that it calls books.books.pdf_converted_text_exp, wombat_staging.panama_haiku_stem, and wombat_staging.panama_haiku_non_stem. *See* Anthropic’s First Supplemental Response to Interrogatory No. 4.

² This estimate is based on the number of unique files present in the dataset.

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

31. Anthropic has used all three datasets in training publicly available Claude models. It used the first dataset to train Claude 3.5 Haiku. *See* Anthropic's Response to Interrogatory No. 4. It used the second and third datasets to train the recently released model Claude 3.7. *See* Anthropic's Supplemental Response to Interrogatory No. 4.

32. Anthropic has produced a report of metadata identifying author, title, and ISBN for each book contained in the dataset used to train Claude 3.5 Haiku as ANT_BARTZ_0000002544 and ANT_BARTZ_000004894. Based on my analysis of the metadata, these scanned books datasets appear to contain approximately [REDACTED]

33. Anthropic produced similar metadata reports for the scanned books used to train Claude 3.7 Sonnet as ANT_BARTZ_000436674 and ANT_BARTZ_000436675. Turvey Decl. at ¶ 26. Based on my analysis of these datasets, they contain approximately [REDACTED]

34. Anthropic has produced records reflecting the books that Anthropic has purchased and scanned for use as training data at ANT_BARTZ_000028814. Turvey Decl. at ¶ 26. This scanned books dataset appears to contain approximately [REDACTED]

E. Plaintiffs' Books Are Present in Each of Anthropic's Datasets

35. According to the declaration of Mr. Tom Turvey, Anthropic's Vice President of Product Partnerships, "Anthropic has purchased and scanned all the specific books that Plaintiffs

³ I understand that ANT_BARTZ_000004894 may contain metadata for all the books included in ANT_BARTZ_0000002544 so have used the estimated number of books from ANT_BARTZ_000004894 as the total here. ANT_BARTZ_0000002544 contains an estimated [REDACTED]. The estimate is made using the number of metadata lines in the files, each of which appears to correspond to a book.

⁴ ANT_BARTZ_000436674 appears to list [REDACTED]. ANT_BARTZ_000436675 appears to list [REDACTED]. The estimate is made using the number of metadata lines in the files, each of which appears to correspond to a book.

⁵ 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 UNDER THE PROTECTIVE ORDER

contend Anthropic has infringed as listed on Exhibit A of the First Amended Complaint (ECF No. 70-1) (“Plaintiffs’ Books”).” *Id.* Also according to Mr. Turvey, “the following books were included in scanned books datasets used to train Claude 3.5 Haiku: *The Lost Night*, *The Herd*, *We Were Never Here*, and *The Spare Room* by Plaintiff Bartz; *The Good Nurse* by Plaintiff Graeber; and *The Feather Thief* and *To Be A Friend Is Fatal* by Plaintiff Johnson.” *Id.* And, as to Claude 3.7 Sonnet, “the following books were included in scanned books datasets used to train Claude 3.7 Sonnet: *The Lost Night*, *The Herd*, *We Were Never Here*, and *The Spare Room* by Plaintiff Bartz and *The Breakthrough* by Plaintiff Graeber.” *Id.*

36. Based on my investigations to date, Anthropic’s pirated books datasets—i.e. Books3, the two Libgen sets, and the “Pilimi” sets—contain multiple copies of works by the Plaintiffs. Specifically, these datasets collectively contain five copies of Andrea Bartz’s *The Herd*, three copies of Ms. Bartz’s *The Lost Night*, five copies of Charles Graeber’s *The Good Nurse*, four copies of Mr. Graeber’s *The Breakthrough*, and four copies of Kirk Wallace Johnson’s *The Feather Thief*. The copies identified are being produced alongside this report.

VI. ANTHROPIC FAILS TO OFFER A COMPREHENSIVE, ACCURATE, OR TESTABLE ACCOUNT OF THE LLM TRAINING PROCESS, AND THUS OBSCURES THE RELEVANT FACTS ABOUT THIS PROCESS

37. In its motion and supporting declaration Anthropic fails to provide a detailed description of the LLM training process that would permit testing or evaluation of the accuracy of Anthropic’s characterization. Anthropic describes the LLM training process alternatively as one akin to human “learning” and, at other points, as a process of “map[ping]” “inferences” gleaned—precisely how is never specified—from training data. *See* Anthropic’s Motion for Summary Judgment at 4. Anthropic’s motion references many processes and concepts but never explains step-by-step how those processes and concepts result in a trained LLM possessing the capabilities Anthropic claims. *See, e.g., id.*

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

38. To support its characterization of the training process, Anthropic principally relies on the declaration of Dr. Jared Kaplan, Anthropic's Chief Science Officer. *See id.* ("A detailed overview of the process to build and train Claude can be found in the Kaplan Declaration."). Following data acquisition, Dr. Kaplan lists the following steps for the LLM pre-training process: (1) data processing and tokenization, (2) vectorization and weights, (3) forward pass, and (4) loss calculation / backpropagation / gradient descent. Kaplan Decl. at ¶ 54-59. Anthropic has also described running evaluations on the model as part of the pre-training process. *See* Anthropic's Response to Plaintiffs' Interrogatory No. 3 (noting that Anthropic runs its models through suites of benchmarks—also called "evaluations"—to measure the model's pre-training progress). It is worth noting that the initial acquisition of the training data, processing and tokenizing the training data, and vectorizing the training data involves Anthropic creating and/or modifying copies of the training data.⁶

39. After the model finishes pre-training, Dr. Kaplan notes that the model may be fine-tuned and may also have certain guardrails—including guardrails which attempt to prevent the model from regurgitating copyrighted data it absorbed during pre-training—added prior to deployment to consumers. *Id.* at ¶ 60-68.

40. Dr. Kaplan's declaration suffers from the same flaws as Anthropic's motion in that it fails to provide a detailed or comprehensive description of how the LLM training process actually occurs. In general, it relies on conclusory assertions about the training process without providing evidence or description sufficient to test such assertions.

⁶ Additionally, as noted below, the model training process itself involves creating a compressed copy of the training data.

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

41. For instance, Dr. Kaplan states that, during “vectorization,” tokens “are converted into ‘vectors’ which are mathematical representations of how words relate to other words.” *Id.* ¶ 56. He does not describe what function(s) perform this conversion, the size of the resulting vectors, their characteristics, or any information regarding how the vectors represent “how words relate to other words.” *Id.* These are all, according to Dr. Kaplan, crucial components of the training process and yet no account is given of how, precisely, they actually embody or encode the semantic relationships of various words or provide an account by which these vectors can reliably be translated into semantic or relational meaning. This makes impedes verification of his statements or that the vectors actually “represent[]” the relationship of “words . . . to other words.” *Id.*

42. Dr. Kaplan then notes that “[t]he model’s ‘weights’ and biases—numerical values that determine how the model understands vectors and their related concepts to be related—start off at random.” *Id.* Dr. Kaplan here states that the model “understands” vectors using “weights” but fails to provide any description of what “understanding” in this sense consists of or the underlying mathematical processes which instantiate or operationalize whatever it is he means by “understanding.” Similarly, Dr. Kaplan uses the term “related concepts” without specifying what the “relatedness” or even “concept” consists of in context of the models in question or how that corresponds in any way to vectors representing tokens.

43. Dr. Kaplan continues to provide cursory invocations of unexplained concepts throughout this paragraph, writing that:

“[T]he model adjusts the weights based on the vector inputs, as well as the vectors themselves, as it learns more about how humans use words and concepts in writing. The representation of that learning is a multidimensional matrix of weights that capture nuances in meaning; the relationship of words to one another; various grammatical properties; how words are modified in context with prefixes, suffixes, and the like; and essentially all other the way humans use language, including, e.g., semantic meaning (the meaning of the token in relation to other tokens); contextual relationships (how words relate to each other in context—“bank” as in financial

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

institution versus “bank” as in river edge); syntactic information (grammatical properties like parts of speech and word tense); morphological information (information about token structure, prefixes, suffixes, and other word forms); positional information (where tokens appear in a sequence); cultural and social connotations (subtle associations that words carry based on usage in society); and hierarchical relationships (information about whether concepts are broader or narrower than others), along with other information, all of which are nuanced details beyond the tokens’ plain definition.” *Id.*

44. Dr. Kaplan fails to explain how any of the above actually occurs or how weights are stored in or used by the model. As just a few examples, Dr. Kaplan provides no explanation of the math underlying his claims, no investigation of the source code underlying the model training, and no detailed demonstration or investigation of the model’s feature space. Instead, Dr. Kaplan simply invokes that the model achieves all of the above by “adjust[ing]” (how?) its “weights based on the vector inputs” (in what way?) “as well as the vectors themselves” (again, in what way?). *Id.* Dr. Kaplan asserts that through this unspecified process the model “learns” about “how humans use words and concepts in writing,” *id.*, but does not explain what he means by “learning” or what mathematical or functional process(es) inside the model are actually taking place. This invocation of “learning” obscures far more than it explains because, as explained below, statistical learning is—in the context of LLM training—entirely consistent with the understanding that the LLM training is compressing data into the model weights.

45. Dr. Kaplan goes on to assert that:

“[t]his iterative adjustment of the model’s internal weight parameters allows it to learn complex patterns and meaningful concepts regarding language, transforming the initially random vectors into rich, contextual representations of the way language works. The outcome of the process is a set of numerical values (the “weights” and “biases”) reflecting what the model has learned from the training corpus” *Id.*

46. As with his other descriptions he does not explain what he means by “learn,” or precisely how the model is “adjust[ed],” or what these “complex patterns and meaningful concepts”

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

are, or how the “random vectors” actually become “rich, contextual representations of the way language works,” or what constitutes a “rich, contextual representation,” or how it is that the referenced “numerical values . . . reflect[] what the model has learned from the training corpus.”

Id. There is thus no way to rigorously evaluate or test his claims.

47. Dr. Kaplan then asserts that “[t]he corpus itself is not stored in the model. And once trained, the model no longer uses the corpus.” *Id.* Dr. Kaplan does not provide any evidence for either of these sentences or contextualize what he means by corpus in this specific context (i.e. if he is referring to the Unicode text characters, or the tokenized characters, or some other form of corpus). As shown below and contrary to Dr. Kaplan’s claims, there is significant evidence—including from Anthropic itself—that the training corpus is compressed into the model during training and that the model is thus relying on the corpus when generating output.

48. In paragraph 58, Dr. Kaplan purports to provide an account of the measurement of the model’s performance during pretraining (“loss calculation”), the algorithms that isolate the cause of any errors (“backpropagation”), and adjust the model in order to modify that performance (“gradient descent”). *Id.* at ¶ 58. As I explain in further detail below, Dr. Kaplan’s characterization of the “loss calculation” is at best incomplete and misses that the optimization of the model’s “loss function” during training is actually equivalent to compressing the training data into the model.

49. At the end of this paragraph Dr. Kaplan again asserts that “[o]verall, through this process the model learns meaningful patterns rather than memorizing specific examples.” *Id.* As before, Dr. Kaplan does not explain what he means by “learns meaningful patterns,” or cite any evidence for his view. As explained below, there is significant evidence against Dr. Kaplan’s views both as concerns the training process and the resulting pretrained LLM.

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

50. Similar issues are also present in Dr. Kaplan’s descriptions regarding the earlier steps of the training process. For example, in characterizing the first step of the process—which Dr. Kaplan describes as “data processing and tokenization”—Dr. Kaplan states that sequences of tokens are “shuffled to make random batches for training which is meant to prevent the model from memorizing the corpus’ original arrangements, or learning false relationships that only exist because of how the original corpus was assembled.” *Id.* at ¶ 55. But as, explained below, Dr. Kaplan’s statement fails to grapple with the significant evidence of memorization in these models—including the significant efforts Anthropic implements to prevent regurgitation of copyrighted output—and, as explained above, fails to provide any explanation of what is meant by “learning” in this context.

51. These same problems appear in Dr. Kaplan’s characterization of the steps following pre-training: “fine-tuning” and “inference.” *Id.* at ¶ 60-65. As to fine-tuning and as one example, Dr. Kaplan states that the model “see[s] what a good response is to a given type of prompt,” *id.* at ¶ 62, without explaining what it would mean for a model to “see” or how, precisely, the model is adjusted in this process to provide different responses. As to “inference” Dr. Kaplan asserts that the model “is generative and not deterministic,” *id.* at ¶ 65, but again fails to grapple with or address the evidence—including Anthropic’s own internal research—suggesting that the models are determined by their training data and that the model training process itself is simply an effort to compress the training data in question. Further, Dr. Kaplan’s statement—to the extent he intends it to imply that the model possesses intelligence—posits a false dichotomy, as, for example, a random number generating function may be both generative and not deterministic without possessing any “intelligence.”

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

52. Finally, these issues are present in the other portions of Dr. Kaplan’s declaration which characterize the training process. For instance, in paragraphs 36 and 37, Dr. Kaplan again invokes phrases like “learn language patterns and the relationship between the words and phrases in the training corpus,” and “map of how humans use language,” and “generalized, functional understanding of how language works.” *Id.* at ¶ 36-37; *see, e.g., id.* at ¶ 40 (stating that the high volume of training data “allows the LLM to gain a generalized understanding of how language works, so that it can interpret prompts and produce original output”). As should be clear from the preceding discussions, Dr. Kaplan’s account here is heavy on assertions and invocations of loaded concepts like “learning” and “understanding” and “map[ping],” but devoid of any specifics or evidence which would permit a meaningful assessment of his claims or the extent to which they are true.

53. Anthropic’s—and Dr. Kaplan’s—cursory description of the training process thus fail to accurately explain what is happening during the training process. More fundamentally, however, these descriptions fail to grapple with alternative—and incompatible—understandings of LLM training that have significant factual support in both the research literature and the factual record of this case. I provide one such understanding here based on my understanding of the LLM training process, the research literature, and the evidence produced in this matter so far.

VII. LLM TRAINING CAN BE UNDERSTOOD AS ENCODING TRAINING DATA THROUGH COMPRESSION

54. Anthropic and Dr. Kaplan’s description of the training process as one in which the LLM exclusively develops “generalized, functional understandings” of how language works is inconsistent with my understanding of how LLMs function and the LLM training process. At this time, I have focused my analysis on demonstrating how Anthropic’s analysis is insufficient and fails to grapple with significant evidence to the contrary, however I reserve the right to supplement

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

my analysis and provide additional views on LLMs and their training process as this case progresses.

55. As one example, Anthropic and Dr. Kaplan’s accounts fail to grapple with the fact that LLMs memorize huge portions of their training data. Whatever Anthropic’s goals, and as one Anthropic researcher put it, “Large LMs memorize A LOT, like A LOT.” ANT_BARTZ_000029063. Dr. Kaplan simply mentions memorization as something to be avoided but fails to offer any detailed explanation of how or why it results from the LLM training process. *See, e.g.,* Kaplan Decl. at ¶ 20, 30-32, 55, 58.

56. Significant theoretical and practical evidence—including from Anthropic’s own internal research and documents—demonstrates that the LLM training process may be accurately characterized as a process of approximately compressing the data used to train the LLM, and thus the memorization in the model is simply a result of the LLM training process.

57. This compression view is supported by well-understood principles of statistical learning, research from other leading AI labs, Anthropic’s own model training source code, Anthropic’s internal statements that models memorize huge portions of their training data, and Anthropic’s own actions in using large troves of pirated books from LibGen as part of their copyrighted-output-prevention system (known internally as “Copy Fairy” or “Content Compare System”). *See* Kaplan Decl. at ¶ 68; Kaplan Dep. Tr. at 93:20-94:6; *id.* at 186:11-14; *id.* at 247:3-7; *id.* at 249:6-13.

58. Research in statistical learning has recognized and verified the close relationship between data compression and the type of statistical modeling which underlies LLM training. Results in statistical learning theory have formally established that “learnability is equivalent to . . . compression,” and “any learning algorithm can be transformed into a compression algorithm,

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

compressing a sample of size m into a sub-sample of size roughly $\log(m)$, and that such a compression algorithm implies learning.” Ofir David, et al, “On statistical learning via the lens of compression,” p. 2, Proceedings of the 30th conference on Neural Information Processing Systems (NIPS 2016) (available at: <https://dl.acm.org/doi/pdf/10.5555/3157382.3157410>). This conclusion applies with equal force to LLMs. As a recent paper by Google AI researchers puts it: “Language Modeling is Compression.” G. Deletang et al, “Language Modeling is Compression,” 12th International Conference on Learning Representations (ICLR 2024) (available at: <https://openreview.net/pdf?id=jznbginyus>). These findings have the significant implication that LLM training can effectively be characterized as a lossy compression process by which the model becomes a compressed version of its training data.

59. To begin, *lossless* compression algorithms are deterministic algorithms that take an input and replace the original content with shorter / smaller / more efficient bit representations of that content. Lossless compression algorithms ideally perform this process in a way that can be deterministically reversed, i.e. uncompressed or decompressed, to recover the original content. For example, the Lempel–Ziv–Welch (“LZW”) family of compression algorithms works by iteratively building a table of sequences of characters and replacing them with short identifiers. When content is run through these algorithms, sequences in the content are replaced with short identifiers as specified in the tables. This compression is deterministic (each run of the algorithm produces an identical result) and lossless, in that the content can always be uncompressed to recover the exact original content.

60. Compression algorithms generally have the property that the more times some piece of data appears in the input, the easier those pieces of data can be compressed. The more unique a piece of data is, i.e. the higher its “entropy,” the harder that data will be to compress losslessly.

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

61. In a lossy compression algorithm, by contrast, the compression algorithm sacrifices some loss of information in order to compress the data further than a lossless algorithm could achieve. Unlike lossless compression algorithms, compressions produced by lossy compression algorithms may not be fully reversed, i.e. used to fully recover the entire and exact original input before it was compressed. However, it is generally possible to recover a close approximation of the original, simply by replacing the lost components with random or deterministically generated data.

62. As an example of a lossy compression algorithm, consider a simple function that takes a real number and rounds it off to produce an integer. This trivial algorithm would take the number “12.5” as input and produce “12” as output. “12” will take fewer bits to store and thus compression has occurred. But decompressing this lossy stored value presents a problem of recovering the digit lost during compression. One solution would be to simply take “12” and append a random decimal component. If we know that the inputs have at most 1 decimal value, then the recovery algorithm has roughly a 1 in 10 chance of fully recovering the original input (“12.5”) simply by guessing that digit.

63. As Dr. Kaplan notes, LLMs are a type of neural network. *See* Kaplan Decl. at ¶ 35. At the start of the training process, the strength of the connections between the network’s neurons—also called the “weights”—are randomized. *See id.* at 56.

64. As a general matter during the training process, a portion of training data is sent as input into the network and the network—following the strength of its weights—produces a predicted output. That output is compared to the “correct” output, i.e. the value contained in training data. For example, the input might be “the quick brown fox jumps over the lazy,” the “correct” output distribution (per the training data) might be “dog” (100%), but the model’s

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

predicted possible outputs distribution might be “boy” (50%), “dog” (12.5%), and a long tail of less-likely potential answers taking up the remaining 12.5%. The meat of the training process—and the equivalence of the LLM training process and compression—comes from how the algorithms measure and correct the neural network in response to this incorrect guess.

65. The central feature of this process is the “loss function.” In LLM training, the loss function measures how much error there is between the predicted output made by current model parameters and the true output in the training data. The loss function calculates a “loss” which represents how far the current model’s output is to the ideal, and is then used to compute updates to the model parameters.

66. Loss functions simply represent the error in how well the current model parameters match a compressed version of the training data. This is because any given loss function is generally calculated by testing the model’s ability to complete text based on partial inputs. That is equivalent to testing how completely a compressed file corresponds to the original data by decompressing it and measuring the difference between the decompressed version and the original data. As such, the lower the measured loss, the closer the compressed version stored in the model is to an ideal compression of the training data.

67. An example of a loss function is cross-entropy loss. As with other loss functions, cross-entropy loss measures the difference between the ideal or true output and the model’s predicted output.⁷ A lower cross-entropy loss indicates that the predicted probabilities are closer to the true or ideal output. An example of a cross-entropy loss calculation looks like:

⁷ In cross-entropy loss generally the calculation is performed relative to “distributions,” but in LLM training, where output comes discretely in token-form, a comparison between distributions can effectively operate as a comparisons of outputs. As noted below, it appears Anthropic’s cross-entropy loss operated in this manner.

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

$$\mathcal{L} = - \sum_{t=1}^T \log P_{\theta}(y_t | y_{<t})$$

68. In this formula the loss equals the negative sum of the log probabilities of the model's prediction given the previous sequence of tokens.⁸

69. Anthropic's model training source code demonstrates that it uses a simple form of cross-entropy loss which computes the loss by comparing a predicted output to the true output present in the training data. *See* ANT_BARTZ_000534168, ANT_BARTZ_000534169 (“ContextShardedCrossEntropyLoss” code implementing loss function).⁹

70. During the LLM training process, the model's parameters are adjusted continuously by measuring how closely the parameters reproduce the training data as shown by loss functions and optimizing the model to reduce the calculated loss. That process of compression, decompression, comparison, and adjustment is repeated over and over until the model approximates a compression of the training data as measured by the loss function. The model training process is thus functionally equivalent to a lossy compression of the training data.

71. To be sure, LLMs have a fixed size or capacity, which is defined by the architecture and the model parameters. That size will not change regardless of the training data used, whether it is 5 sentences or 5 million books. A large LLM that is trained on a small dataset may have a large amount of redundancy in its model weights. That same LLM trained on much more training

⁸ In this formula, T is the length of the sequence; y_t is the token of this sequence at index t; $y_{<t}$ represents the set of tokens in the sequence before t; and $P_{\theta}(y_t \text{ given } y_{<t})$ represents the model (which is defined by parameter θ) predicted probability y_t given the previous tokens $y_{<t}$.

⁹ Anthropic has also confirmed the use of cross-entropy loss in its Claude 2 model card. *See* “Model Card and Evaluations for Claude Models,” Anthropic, p. 9, Nov. 2023, <https://www-cdn.anthropic.com/5c49cc247484cecf107c699baf29250302e5da70/claude-2-model-card.pdf> (measuring Claude 2 performance using “autoregressive cross-entropy loss”).

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

data will still store large portions of the training dataset, but may discard portions that do not fit into the size of the LLM model.¹⁰

72. In addition to Anthropic’s **loss function source code**, the third-party research described above, and my understanding of the LLM training process, the compression understanding of the LLM training process described in this declaration is further supported by (1) Anthropic’s **internal statements and research regarding model memorization** and (2) Anthropic’s **attempts to prevent regurgitation of copyrighted training data by using huge volumes of pirated books**.

73. Anthropic’s internal documents **concede that models memorize and, in fact, that “[b]igger models memorize faster.”** ANT_BARTZ_000004936 at 46. This is entirely consistent with the compression account of model training as compression entails memorization and bigger models have more space to store compressed data. Even just making one pass over training data, **Anthropic’s models are capable of huge quantities of memorization.** See ANT_BARTZ_000004957 at 62-63 (noting that a model had memorized “a ~third of common crawl, and like a fifth of Github” even in a “one epoch regime” i.e. seeing the source once).

74. Further buttressing the compression understanding of the training process and the equivalence of compression and memorization, **Anthropic measured model memorization in terms of *loss*, i.e. the same parameter optimized during the training process and which, as explained**

¹⁰ This is consistent with the view that the LLM model size effectively set a ceiling on how powerful a model could be, that model size should increase as the amount of training data increases, and helps explain why AI developers generally continue to push for bigger and bigger models and correspondingly larger datasets. See, e.g., Kaplan Decl. at ¶ 38 n.25 (quoting Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models* (Mar. 29, 2022) at 1, <https://arxiv.org/pdf/2203.15556> as saying that “Kaplan et al. (2020) showed that there is a power law relationship between the number of parameters in an autoregressive language model (LM) and its performance. As a result, the field has been training larger and larger models, expecting performance improvements.”).

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

elsewhere, can be understood a measure of compression. As one researcher put it: “Large LMs memorize A LOT, like A LOT” and of a “few hundred thousand tokens” checked a given model “had memorized a sixth of them down to losses of . . . 0.000001 (i.e. a confidence of 99.9999%).” ANT_BARTZ_000029063 at 109.

75. Anthropic’s deployment of its LLMs to consumers and businesses reflects its understanding of the nature of the training process and the scale of training data memorization. In particular, the system Anthropic uses to detect and prevent its models from outputting copyrighted text to users—a system which Anthropic has called “Copy Fairy” or its “Content Compare System”—relies on a huge repository of pirated books from Libgen. *See* Kaplan Dep. Tr. at 93:20-94:6 (“My recollection is that Anthropic later used [the Libgen] dataset as a part of a CCS system”); *id.* at 186:11-14 (“Libgen is being used in the CCS tool”); *id.* at 247:3-7 (“What I understand is that a Copyfairy was an informal term used earlier in Anthropic’s history for the system that I’ve referred to as CCS.”); *id.* at 249:6-13 (“[M]y understanding is the Copyfairy system and CCS are the same thing. The Copyfairy was an informal term used earlier in Anthropic’s history for the system now designated as CCS. And so there isn’t any distinction between Copyfairy and CCS.”).¹¹

76. This system—touted by Dr. Kaplan as one of the model’s “guardrails”—works by checking model outputs against hashed versions of the pirated works in the Libgen set and prevents the display of the output if there is a match. Kaplan Decl. at ¶ 68; *see* ANT_BARTZ_000533421-22 (source code implementing Copyfairy hash check). This check required optimization to perform well. *See* ANT_BARTZ_000533421 (github commits re: reducing queries to hashed works).

¹¹ Based on Github information produced at ANT_BARTZ_000336625, it appears Anthropic may have also used Books3 for this purpose as well.

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

Anthropic even created a failsafe which would bypass the filter when it took too long to run. *See* ANT_BARTZ_000533421 (defining “validation_timeout”); ANT_BARTZ_000533425 (implements time limit on validation); ANT_BARTZ_000533438 (code indicating validation timeout does not stop model generation).

77. Based on my experience, if Anthropic were confident that its models were not compressing large fractions of the millions of books in its training data during the training process—or if Anthropic truly believed that this training compression could not result in the model decompressing training data as output—it would not make sense to check all model outputs against this large corpus of pirated books.

VIII. ANTHROPIC HIGHLY VALUED BOOKS AS TRAINING DATA FOR ITS LLMs

78. Anthropic’s use of books as training data resulted in significant benefits to its LLMs as shown by Anthropic’s own statements and third-party research. Those statements and that research confirms that Anthropic uses books in training its LLMs because of their high-quality long-form expressive content.

79. To begin with, Dr. Kaplan himself admits that it is “important to train [LLMs] on long-form content like books.” Kaplan Decl. at ¶ 43. In numerous internal documents and transcripts of internal meetings, Anthropic has repeatedly acknowledged that books are generally very valuable in training its LLMs. *See, e.g.*, ANT_BARTZ_000003492 (“Hard to quantify, but overall we believe that books represent very high quality pretraining data.”); ANT_BARTZ_000340761 at 66 (explaining “if we can get the [REDACTED] of books that arrive in May into the model ASAP, I think it will be very valuable. We should make sure we are ready to move fast here.”); ANT_BARTZ_000004587 (noting that books “align[] with our core content source requirements” because they “offer a vast quantity of high-quality, unique content”); ANT_BARTZ_000003439

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

(describing “books” as “higher quality data” and that “We expect a large one-time boost if we are able to train on a large number of books.”); ANT_BARTZ_000005224 at 25 (recognizing the data “quality” of “scanning [REDACTED] books” to be “quite high”); ANT_BARTZ_000003433 (identifying “books” as the first category of “high quality data”); ANT_BARTZ_000533972 at 75-76 (discussing the need for more training data, stating “libgen books are highest quality data,” and so torrenting Pilimi “could theoretically solve our data issue with mostly high quality data”).

80. Indeed, Anthropic has further acknowledged that “having *every* book is more valuable to us than a smaller volume of the most critically-acclaimed literary works when it comes to training our models,” ANT_BARTZ_000004936 at 37, and that “Just Volume - *any* books is valuable,” *id.* at 46. *See* ANT_BARTZ_000168435 at 41 (“[W]e care most about sheer volume of books rather than specialization at this time.”). Anthropic has its “strongest conviction” about books as training data because “[a]cross all categories of written data, books are both more useful and available in larger quantities.” ANT_BARTZ_000168435; *see also, id.* at 36 (“The highest volume of useful data is available to us via published books.”); *id.* (listing books as the only data with “high to very high” “conviction in usefulness and “very high” in “availability in large quantities”).

81. And using books as training data unlocked large compute and time savings for Anthropic. *See id.* at 40 (“book data would make our models [REDACTED] more efficient, which effectively means we can train future models for [REDACTED] less compute (less long) and get the same result as without the book data.”).

82. As described below, the performance of Anthropic’s LLMs greatly benefit from training on a large corpus of books for at least three reasons.

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

83. First, a large corpus of books can be crucial to an LLM’s ability to produce high-quality expressive text “rather than low quality internet speak.” ANT_BARTZ_000003433. This is what Anthropic describes as “The Good Writing Hypothesis,” which is “Models just need good writing.” *Id.* Books are an excellent source of a large volume of “good writing” because they are comprised of “content an editor would approve of.” *Id.* As Anthropic recognized, then, books are generally superior to data that is typically derived from online sources, such as message board, social media, or other webscraped material, *id.*, which can often be unstructured and require extensive preprocessing and filtration before it can be used for training. *See also* ANT_BARTZ_000279483 at 95 (noting “book tokens dramatically outperform web-crawl content in multiple dimensions”).

84. Anthropic’s conclusions align with those of 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” and have described 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).¹² 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

¹² 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 UNDER THE PROTECTIVE ORDER

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 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, books are generally a great source of well-written long-form text that enable LLMs to both process coherent long-form text as an input and produce coherent long-form text as

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

an output. As Anthropic explains, “models need data with coherent long-contexts so that they can learn to read and think over large segments of text.” ANT_BARTZ_000003433.

86. This is further supported by academic research. 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>).

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, a large corpus of books can enable LLMs to produce expressively varied text in a wide array of genres, topics, and styles. See ANT_BARTZ_000279483 at 95 (explaining that Anthropic “want[s] more volume and diversity of books”); ANT_BARTZ_000004587 (recognizing books “offer a vast quantity of high-quality, unique content”). This again is supported by academic research and Anthropic’s own documents. As a 2024 research paper titled *Understanding LLMs: A Comprehensive Overview from Training to Inference* explains, large books datasets that “include a wide range of literary genres, including novels, essays, poetry,

CONTAINS MATERIAL DESIGNATED UNDER THE PROTECTIVE ORDER

history, science, philosophy, and more ... contribute to the models' training by exposing them to a diverse array of textual genres and subject matter, fostering a more comprehensive understanding of language across various domains." Liu, et. al, Jan. 2024, p. 7 (available at: <https://arxiv.org/pdf/2401.02038>). Dr. Kaplan and third-party researchers agree that LLMs must be trained with a diverse set of training data. *See* Kaplan Decl. at ¶ 47; "What Language Model to Train if You Have One Million GPU Hours?," Scao et. al., p. 4, Nov. 2022 (available at: <https://arxiv.org/pdf/2210.15424>) ("Diverse cross-domain pretraining data combining web crawls with curated high-quality sources improves zero-shot generalization over pretraining datasets constructed from Common Crawl only.").

89. A review of Anthropic's own description of its book datasets demonstrates that Anthropic prioritized this variety. For example, in a spreadsheet summarizing recent book acquisitions, Anthropic stated that it had ordered "children's books", "poetry, classics", "literature & fiction", "animal[]" books, "Christmas" books, "Thriller & Suspense", "Cookbooks, Food & Wine", "Historical" books, "Self-Help, Spiritual Growth" books, "Humor" books, and "Contemporary" books, just to name a few. *See* ANT_BARTZ_000263673.

90. In sum, based on my research, third-party research, and Anthropic's own documents, books were extremely valuable as training data for Anthropic's LLMs. These same factors regarding the quality and value of books as a source of training data underlay Anthropic's decision to engage in large-scale acquisition of books through pirated websites and book scanning.