Justin A. Nelson *(pro hac vice forthcoming)*
Alejandra C. Salinas *(pro hac vice forthcoming)*
SUSMAN GODFREY L.L.P
1000 Louisiana Street, Suite 5100
Houston, TX 77002-5096
Telephone: (713) 651-9366
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (SBN 316062)
SUSMAN GODFREY L.L.P
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067-2906
Telephone: (310) 789-3100
RNath@susmangodfrey.com

J. Craig Smyser *(pro hac vice forthcoming)*
SUSMAN GODFREY L.L.P
One Manhattan West, 51st Floor,
New York, NY 10019
Telephone: (212) 336-8330
Facsimile: (212) 336-8340
csmyser@susmangodfrey.com

Jordan W. Connors *(pro hac vice forthcoming)*
SUSMAN GODFREY L.L.P
401 Union Street, Suite 3000
Seattle, WA 98101
Telephone: (206) 516-3880
jconnors@susmangodfrey.com

Rachel Geman *(pro hac vice forthcoming)*
Wesley Dozier *(pro hac vice forthcoming)*
Anna Freymann *(pro hac vice forthcoming)*
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, New York 10013-1413
Telephone: (212) 355-9500
rgeman@lchb.com
wdozier@lchb.com
afreymann@lchb.com

Reilly T. Stoler (SBN 310761)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111-3339
Telephone: (415) 956-1000
rstoler@lchb.com

Attorneys for Plaintiffs and the Proposed Class
(Additional Counsel Listed on Signature Page)

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

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

Plaintiffs,

v.

ANTHROPIC PBC,

Defendants.

Case No. _____

CLASS ACTION COMPLAINT

JURY TRIAL DEMANDED

Plaintiffs Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson on behalf of themselves and all other similarly situated (the "Class," as defined below), for their complaint against Defendant Anthropic PBC ("Anthropic"), allege as follows:

NATURE OF THE ACTION

1. Anthropic has built a multibillion-dollar business by stealing hundreds of thousands of copyrighted books. Rather than obtaining permission and paying a fair price for the creations it exploits, Anthropic pirated them. Authors spend years conceiving, writing, and pursuing publication of their copyrighted material. The United States Constitution recognizes the fundamental principle that creators deserve compensation for their work. Yet Anthropic ignored copyright protections. An essential component of Anthropic's business model—and its flagship "Claude" family of large language models (or "LLMs")—is the largescale theft of copyrighted works.

2. Plaintiffs are authors of an array of works of fiction and nonfiction. They bring this action under the Copyright Act to redress the harm caused by Anthropic's brazen infringement. Anthropic downloaded known pirated versions of Plaintiffs' works, made copies of them, and fed these pirated copies into its models. Anthropic took these drastic steps to help computer algorithms generate human-like text responses.

3. Anthropic has not even attempted to compensate Plaintiffs for the use of their material. In fact, Anthropic has taken multiple steps to hide the full extent of its copyright theft. Copyright law prohibits what Anthropic has done here: downloading and copying hundreds of thousands of copyrighted books taken from pirated and illegal websites.

4. Anthropic's Claude LLMs compromise authors' ability to make a living, in that the LLMs allow anyone to generate—automatically and freely (or very cheaply)—texts that writers would otherwise be paid to create and sell. Anthropic's LLMs, which dilute the commercial market for Plaintiffs' and the Class's works, were created without paying writers a cent.

5. Anthropic's immense success is a direct result of its copyright infringement. The quality of Claude, or any LLM, is a consequence of the quality of the data used to train it. The more high-quality, longform text on which an LLM is trained, the more adept an LLM will be in

generating lifelike, complex, and useful text responses to prompts. Without usurping the works of Plaintiffs and the members of the Class to train its LLMs to begin with, Anthropic would not have a commercial product with which to damage the market for authors' works. Anthropic has enjoyed enormous financial gain from its exploitation of copyrighted material. Anthropic projects it will generate more than $850 million of revenue in 2024.[1] After ten rounds of funding, Anthropic has raised $7.6 billion from tech giants like Amazon and Google. As December 2023, these investments valued the company in excess of $18 billion and is likely even higher today.[2]

6. Anthropic's commercial gain has come at the expense of creators and rightsholders, including Plaintiffs and members of the Class. Book readers typically purchase books. Anthropic did not even take that basic and insufficient step. Anthropic never sought—let alone paid for—a license to copy and exploit the protected expression contained in the copyrighted works fed into its models. Instead, Anthropic did what any teenager could tell you is illegal. It intentionally downloaded known pirated copies of books from the internet, made unlicensed copies of them, and then used those unlicensed copies to digest and analyze the copyrighted expression-all for its own commercial gain. The end result is a model built on the work of thousands of authors, meant to mimic the syntax, style, and themes of the copyrighted works on which it was trained.

7. Anthropic styles itself as a public benefit company, designed to improve humanity. In the words of its co-founder Dario Amodei, Anthropic is "a company that's focused on public benefit."[3] For holders of copyrighted works, however, Anthropic already has wrought mass destruction. It is not consistent with core human values or the public benefit to download hundreds of thousands of books from a known illegal source. Anthropic has attempted to steal the fire of Prometheus. It is no exaggeration to say that Anthropic's model seeks to profit from strip-mining the human expression and ingenuity behind each one of those works.

[1] "Anthropic Forecasts More Than $850 Mln In Annualized Revenue By 2024-end—Report," *Reuters,* Dec. 26, 2023, https://www.reuters.com/technology/anthropic-forecasts-more-than-850-mln-annualized-revenue-rate-by-2024-end-report-2023-12-26/ (last visited Aug. 15, 2024).

[2] "Amazon Injects an Additional $2.75B Vote of Confidence in Anthropic," *Spice Works,* Mar. 28, 2024, https://www.spiceworks.com/tech/artificial-intelligence/news/amazon-invests-4-billion-anthropic/ (last visited Aug. 15, 2024).

[3] "Anthropic CEO Dario Amodei on Being an Underdog, AI Safety, and Economic Inequality," *Time*, Jun. 23, 2024, https://time.com/6990386/anthropic-dario-amodei-interview/ (last visited Aug. 15, 2024).

JURISDICTION AND VENUE

8. The Court has subject matter jurisdiction under 28 U.S.C. §§ 1331 and 1338(a) because this action arises under the Copyright Act of 1976, 17 U.S.C. § 101, *et seq*.

9. The Court also has personal jurisdiction over Defendant because it has purposely availed itself of the privilege of conducting business in this District.

10. Anthropic is headquartered in this District and its copyright infringement—including the downloading and reproduction of the copyrighted works—occurred, in substantial part, in this District. Anthropic also marketed, sold, and distributed its LLMs from this District to citizens of California.

11. Venue is proper under 28 U.S.C. § 1400(a) because Anthropic or its agents reside or may be found in this District due to their infringing activities, along with their commercialization of their infringing activities, that occurred in this District. Venue is also proper under 28 U.S.C. § 1391(b)(2) because a substantial part of the events giving rise to Plaintiffs' claims occurred in this District, including the largescale copyright infringement and commercialization of Anthropic's LLMs.

THE PARTIES

12. Plaintiff Andrea Bartz is an author and journalist who resides in New York. She is the author of *The Lost Night: A Novel*, *The Herd*, *We Were Never Here*, and *The Spare Room*.

13. Plaintiff Charles Graeber is an author and journalist who resides in New York. He is the author of *The Good Nurse: A True Story of Medicine, Madness, and Murder* and *The Breakthrough: Immunotherapy and the Race to Cure Cancer*. Plaintiff Graeber has written numerous essays for *The New Yorker*, *The New York Times*, and *GQ*.

14. Plaintiff Kirk Wallace Johnson is an author and journalist who resides in Los Angeles, California. He is the author of *The Fisherman and the Dragon: Fear, Greed, and a Fight for Justice on the Gulf Coast*, *The Feather Thief: Beauty, Obsession, and the Natural History Heist of the Century*, and *To Be A Friend Fatal: The Fight to Save the Iraqis America Left Behind*. He is the founder of The List Project, a nonprofit that has helped resettle over 2,500 Iraqi refugees who worked for U.S.-affiliated organizations throughout the Iraq war.

15. Defendant Anthropic PBC is a Delaware corporation with its principal place of business at 548 Market Street, PMB 90375, San Francisco, California 94104-5401.

FACTUAL ALLEGATIONS

I. Anthropic's Founding and the Development and Commercialization of Claude

16. Anthropic was founded in January 2021 by seven former OpenAI employees, including current Chief Executive Officer Dario Amodei and President Daniela Amodei. Before founding Anthropic, Dario Amodei was OpenAI's vice president of research, where he was "one of two people who set[] overall research direction at OpenAI" and "led efforts to build [OpenAI's] GPT-2 and GPT-3" models.[4]

17. Anthropic released the first iteration of its flagship model Claude in March 2023, shortly after OpenAI's ChatGPT took the world by storm.[5] A few months later, in July 2023, Anthropic released the next iteration of the model, Claude 2, for public use.[6] Then, in March 2024, Anthropic released Claude 3.[7] Claude 3 was released with three levels, ranging from simplest to most advanced: Claude Sonnet, Claude Haiku, and Claude Opus.[8] Most recently, in June of this year, Anthropic released its latest Claude iteration, Claude 3.5 Sonnet.[9] Claude is available for use via web interface, on Android and iOS applications, and via an application programming interface, which allows developers to build custom generative AI tools using Claude as the base.[10]

18. According to Anthropic, Claude outperforms other competing LLMs on the market, such as OpenAI's ChatGPT and Google's Gemini. Anthropic boasts that Claude can be used to "draft everything from a text message or email to a screenplay or a novel."[11]

---

[4] LinkedIn Page for Dario Amodei, https://www.linkedin.com/in/dario-amodei-3934934/ (last visited Aug. 15, 2024).
[5] *See* "What Is Claude AI and Anthropic? A Closer Look at ChatGPT's Rival," *Tech.co*, https://tech.co/news/what-is-claude-ai-anthropic (last visited Aug. 15, 2024).
[6] *Id.*
[7] *See* "Introducing the next generation of Claude," *Anthropic*, Mar. 4, 2024, https://www.anthropic.com/news/claude-3-family (last visited Aug. 15, 2024).
[8] *Id.*
[9] *See* "Claude 3.5 Sonnet," *Anthropic*, Jun. 20, 2024, https://www.anthropic.com/news/claude-3-5-sonnet (last visited Aug. 15, 2024).
[10] *See* "What interfaces can I use to access Claude?," *Anthropic*, https://support.anthropic.com/en/articles/8114487-what-interfaces-can-i-use-to-access-claude (last visited Aug. 15, 2024).
[11] "What are some things I can use Claude for?," *Anthropic*, https://support.anthropic.com/en/articles/7996845-what-are-some-things-i-can-use-claude-for

19. Claude has been enormously successful, with both the general public and with enterprises. Anthropic currently offers access to its Claude 3.5 Sonnet model free, with increased usage caps and access to models like Claude 3 Opus or Haiku gated with subscription fees of between $20 to $30 per month.[12] Claude has garnered tens of millions of monthly active users and has been incorporated into Amazon Bedrock (Amazon Web Services' platform for cloud AI services).[13] Anthropic has been particularly successful in courting a corporate client base, which generates paid subscription revenue. Claude's customers include "leading enterprises and startups" such as Slack, Zoominfo, Asama, Bridgewater, LexisNexis, and Jane Street Capital.[14]

20. Throughout its existence, Anthropic has cloaked itself in the rhetoric of "AI safety" and "responsibility."[15] Its actions, however, have made a mockery of its lofty goals. Anthropic's immense success has been built, in large part, on its largescale copyright theft. As alleged in more detail below, Anthropic's models were trained on troves of pirated material. Moreover, Anthropic could not have built a model capable of digesting whole books and generating complex text outputs without its exploitation of these works.

II. Anthropic Engaged in Largescale Copyright Theft in Training Its LLMs

1. Large Language Models and the Training Process

21. Claude is a type of large language model or "LLM." LLMs attempt to "understand" human language by processing input text, and are designed to mimic human use of language by generating output text on a predictive basis, *i.e.*, predicting what word follows what.

22. Claude is a complex web of mathematical functions comprised of a series of algorithms that break down input text into smaller pieces—words or portions of words, called "tokens"—then translate those pieces into "vectors," or a sequence of numbers that is used to identify the token within the series of algorithms. Those vectors help place each token on a map,

(last visited Aug. 15, 2024); *see* "Claude 3.5 Sonnet," *Anthropic*, Jun. 20, 2024, https://www.anthropic.com/news/claude-3-5-sonnet (last visited Aug. 15, 2024).

[12] *See* "Pricing," *Anthropic*, https://www.anthropic.com/pricing (last visited Aug. 15, 2024).

[13] *See* "Anthropic's Claude in Amazon Bedrock," *Amazon Web Services*, https://aws.amazon.com/bedrock/claude/ (last visited Aug. 15, 2024).

[14] "Customers," *Anthropic*, https://www.anthropic.com/customers (last visited Aug. 15, 2024).

[15] *See* "Inside the White-Hot Center of A.I. Doomerism," *New York Times*, Jul. 11, 2023, https://www.nytimes.com/2023/07/11/technology/anthropic-ai-claude-chatbot.html (last visited Aug. 15, 2024).

by identifying other tokens closely associated with the word. As described by Anthropic's competitor OpenAI: "the process begins by breaking text down into roughly word-length 'tokens,' which are converted to numbers. The model then calculates each token's proximity to other tokens in the training data—essentially, how near one word appears in relation to any other word. These relationships between words reveal which words have similar meanings . . . and functions."[16] As the model trains and digests more expression, the algorithms depicting the relationship between various tokens changes with it.

23. The model is trained on a massive corpus of text; without training, there is no LLM. As Anthropic has described it, "[l]arge language models such as Claude need to be 'trained' on text so that they can learn the patterns and connections between words. This training is important so that the model performs effectively and safely."[17]

24. The model takes text inputs in the form of an incomplete phrase or passage, and attempts to complete the phrase, essentially a fill-in-the-blank quiz. The model compares its predicted phrase completion with the actual "correct" answer. The model then adjusts its internal algorithms to "learn" from its mistakes. In other words, it adjusts its algorithms to reduce the likelihood of making the same mistake again and thus minimizing the difference between any given text input and the "correct" text output.

25. The model then repeats this same cycle millions, possibly billions, of times across the entire corpus, adjusting its algorithms each time to reflect the text input from the corpus. The pre-training process enables the model to process prompts and generate text output that mimics human language. It does so by exposing the model to a wide range of texts and using algorithms to predict the next word in the text. By repeating this process over and over, the model exhibits fluency in style, syntax, and expression of ideas, largely by digesting and processing the expression contained in the material used for training. In this way, the LLM effectively mines and feeds on the expression contained in the training corpus, adjusting its algorithms such that it can mirror and

[16] *See* Comment of OpenAI "Re: Notice of Inquiry and Request for Comment [Docket No. 2023-06]," United States Copyright Office, Oct. 30, 2023, p. 5-6 (available at: https://downloads.regulations.gov/COLC-2023-0006-8906/attachment_1.pdf).

[17] "How do you use personal data in model training?," *Anthropic*, https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training (last visited Aug. 15, 2024).

mimic the ordering of words, style, syntax, and presentation of facts, concepts, and themes.

26. After the pre-training process, the generative model must undergo a further post-training process. At this point, the model is capable of completing phrases and predicting the next word or words that come next after a particular text input, but cannot yet respond to questions, let alone with human-like responses. The post-training process is sometimes referred to as "fine-tuning." This stage typically involves more human supervision, and focuses on making adjustments to the model using comparatively smaller training datasets.

27. For both the post- and pre-training processes in developing Claude, Anthropic created multiple, unlicensed copies of the training data. As the U.S. Patent and Trademark Office has observed, LLM "training" "almost by definition involve[s] the reproduction of entire works or substantial portions thereof."[18]

28. The quality and quantity of the corpus is critical to the quality of the resulting model. With respect to LLM development, the phrase "garbage in, garbage out" carries weight. As one researcher put it: "[large language] model behavior is *not* determined by architecture, hyperparameters, or optimizer choices [i.e. technical features set during model training]. *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.*"[19]

29. Claude, for example, has shown the capability of coherently processing an entire book at once—up to 75,000 words—and generating coherent, clearly-written passages in response. These responses mimic an understanding not just of the proper ordering of words and syntax, but also higher-level themes and ideas. Claude could only develop this capability from training on high-quality prose and complex, longer pieces.

30. In this way, books are especially valuable training material. As one commentator put it, "[b]ooks offer formal and lengthy texts which help LLMs understand complex language

[18] U.S. Patent & Trademark Office, *Public Views on Artificial Intelligence and Intellectual Property Policy* 29 (2020), *available at* https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf (last accessed Aug. 15, 2024).

[19] *See* "The 'it' in AI models is the data set," James Betker, https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/ (June 10, 2023) (emphasis added).

structures, grasp long-term context, and produce coherent narratives."[20]

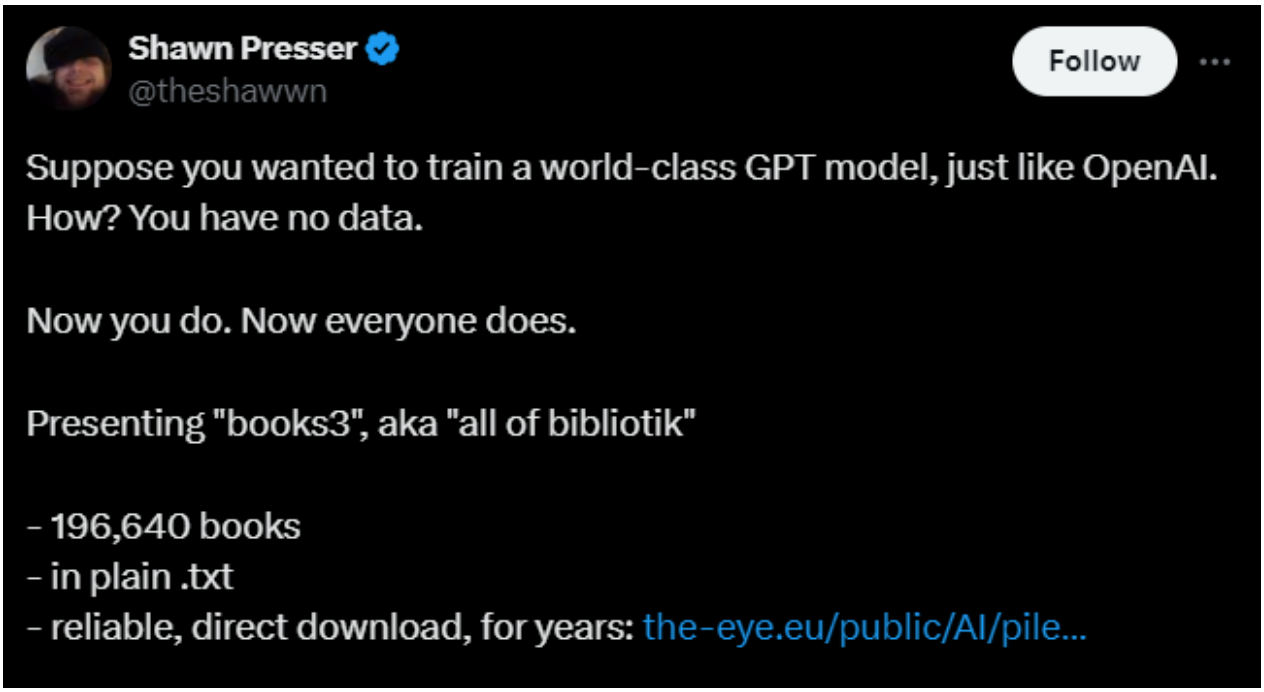
2. Anthropic Copied A Massive Trove of Pirated Books To Train Claude

31. Though Anthropic has been particularly secretive about the sources of its training corpus for Claude, Anthropic has admitted to using a dataset called The Pile.

32. The Pile is an 800 GB+ open-source dataset created for large language model training. The Pile was hosted and made publicly available online by a nonprofit called EleutherAI. As described by its creators, "The Pile is constructed from 22 diverse high-quality subsets . . . many of which derive from academic and professional sources. . . . [M]odels trained on the Pile improve significantly over both Raw CC and CC-100 on all components of the Pile, while improving performance on downstream evaluations."[21]

33. One of The Pile's architects is an independent developer named Shawn Presser. Presser created a dataset included in The Pile called "Books3," which is a trove of pirated books.

34. Presser described how he created Books3 in a Twitter thread from October 2020:[22]



[20] "Pretraining of Large Language Models," Ritwik Raha, *Github.com* https://gist.github.com/ritwikraha/77e79990992043f60a9588610b2781c5 (last visited Aug. 15, 2024).

[21] "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," Gao et al, Abstract, https://arxiv.org/pdf/2101.00027 (last visited Aug. 15, 2024).

[22] *See* Tweet by Shawn Presser, Oct. 25, 2020, https://x.com/theshawwn/status/1320282149329784833?lang=en (last visited Aug. 15, 2024).

35. Presser went on. He said he created Books3 in response to "OpenAI's papers on GPT-2 and 3," which "refer[] to datasets named 'books1' and 'books2,'" the latter of which Presser suspects "might be 'all of libgen.'"[23] LibGen refers to "Library Genesis," a website offering pirated books that was ordered shut down for copyright infringement in 2015. *See Elsevier, Inc. et al v. www.Sci-hub.org et al*, 15-cv-2482-RWS, Dkt. 53 (Oct. 30, 2015). To create a pirated-book dataset comparable to what he suspected OpenAI created for itself, Presser announced that Books3 was also a direct download of all books from a *different pirated website*—a compilation of "196,640 books," which comprises "all of bibliotik."[24]

36. Bibliotik is a "notorious pirated collection" of "pirated books."[25] For years prior to its use as "Books3," Bibliotik was frequently included in roundups of the best—and most popular—sources for pirated material.[26]

37. Books3 was a critical part of The Pile. In EleutherAI's paper on The Pile, it touted

[23] *See* Tweet by Shawn Presser, Oct. 25, 2020, https://x.com/theshawwn/status/1320282152689336320 (last visited Aug. 15, 2024).

[24] *See* Tweet by Shawn Presser, Oct. 25, 2020, https://x.com/theshawwn/status/1320282149329784833?lang=en (last accessed Aug. 15, 2024).

[25] *See* Schoppert, "Whether you're an undergraduate doing research, or a fan of the Nick Stone novels, or indeed a hungry AI...," Nov. 29, 2022, https://aicopyright.substack.com/p/whether-youre-an-undergraduate-doing ("What is Bibliotik? A notorious pirated collection."); "What I Found in a Database Meta Uses to Train Generative AI," Alex Reisner, *The Atlantic*, Sept. 25, 2023, https://www.theatlantic.com/technology/archive/2023/09/books3-ai-training-meta-copyright-infringement-lawsuit/675411/ ("a collection of pirated ebooks, most of them published in the past 20 years."); "Revealed: The Authors Whose Pirated Books are Powering Generative AI," Alex Reisner, *The Atlantic*, Aug. 19, 2023, https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/ ("collections of pirated books, such as Library Genesis, Z-Library, and Bibliotik, that circulate via the BitTorrent file-sharing network."); "Are ChatGPT, Bard and Dolly 2.0 Trained On Pirated Content?," Roger Monti, *Search Engine Journal*, April 20, 2023, https://www.searchenginejournal.com/are-chatgpt-bard-and-dolly-2-0-trained-on-pirated-content/485089/ ("The Books3 dataset contains the text of books that were pirated and hosted at a pirate site called, bibliotik.").

[26] *See* Commit History of "Awesome Piracy," *Github.com,* Oct. 13, 2018, https://github.com/aviranzerioniac/awesome-piracy/commit/61928765e3ee0b4f3dbe3c0724b196e5f0f17e59?short_path=5a831ea#diff-5a831ea67cf5cf8703b0de46901ab25bd191f56b320053be9332d9a3b0d01d15 (October 13, 2018 commit to "awesome piracy" repo listing "Bibliotik Popular ebooks/audiobooks private tracker"); "Reddit Piracy Megathread" repo., *Github.com*, Mar. 21, 2019 https://github.com/magicoflolis/Reddit-Piracy-Megathread/blob/master/data/findingtextbooks.md (March 21, 2019 guide from "r/piracy" on how to source textbooks listing "Bibliotik"); "List of free eBook download sites," *Pirates-forum.org,* Mar. 06 2014, https://pirates-forum.org/Thread-List-of-free-eBook-download-sites?highlight=bibliotik (March 31, 2014 post from "pirates forum" thread entitled "List of free eBook download sites" listing "bibliotik").

the key value of Books3 as training material: "Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker . . . Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude larger than our next largest book dataset (BookCorpus2)." The paper then summarized its key point for why The Pile included this known source of illegal copyright material: "***We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling***."[27]

38. At the same time, Presser and EleutherAI repeatedly and publicly acknowledged that, with The Pile and Books3, they were making available a cache of pirated material. EleutherAI's paper on The Pile noted that "there is little acknowledgment of the fact that the processing and distribution of data owned by others may also be a violation of copyright law."[28] Furthermore, The Pile's datasheet notes that "Books3 is almost entirely comprised of copyrighted works . . . ."[29] Presser, for his part, has admitting to releasing Books3 despite "fear of copyright backlash."[30]

39. In August 2023, Books3 was removed from the "most official" copy of The Pile hosted by "The Eye" due to copyright complaints. Despite this takedown, the original version appears otherwise available as part of The Pile from other sources.

40. Though Anthropic has gone to great lengths to conceal the contents of its training datasets, what is known about the training data indicates that Anthropic's Claude models were trained on a mass of copyrighted books and other copyrighted material. It is apparent that Anthropic downloaded and reproduced copies of The Pile and Books3, knowing that these datasets were comprised of a trove of copyrighted content sourced from pirate websites like Bibiliotik.

41. In a December 2021 research paper on large language model training, Anthropic described creating a dataset "most of which we sourced from the Pile" and which included "32%

[27] "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," Gao et al, p. 3-4, https://arxiv.org/pdf/2101.00027 (last visited Aug. 15, 2024).
[28] *Id.* at 14-15.
[29] "Datasheet for the Pile," Gao et al, Jan. 20, 2022, p. 15, https://arxiv.org/pdf/2201.07311 (last visited Aug. 15, 2024).
[30] Comment of "sillysaurusx," *Hacker News*, Jul. 11, 2023, https://news.ycombinator.com/item?id=36685115 (last visited Aug. 15, 2024).

internet books," a code word in the industry for pirated copies of books available on the internet.[31]

42. More recently, in July 2024, Anthropic has publicly acknowledged that it used The Pile to train its Claude models. As reported by Proof News, company spokesperson Jennifer Martinez "confirm[ed] use of the Pile in Anthropic's generative AI assistant Claude."[32] Anthropic confirmed the same to Vox News.[33] Independent researchers have tested Claude to shed light on the composition of its training set, and their work has confirmed a high likelihood that Claude was trained on copyrighted books.[34]

43. Anthropic thus copied and exploited a trove of copyrighted books—including but not limited to the books contained in Books3—knowing that it was violating copyright laws. Instead of sourcing training material from pirated troves of copyrighted books from this modern-day Napster, Anthropic could have sought and obtained a license to make copies of them. It instead made the deliberate decision to cut corners and rely on stolen materials to train their models.

44. Anthropic's commercial copying of Plaintiffs' work and works owned by the proposed Class was manifestly unfair use, for several reasons. Anthropic has suggested that it uses the training data to "learn the patterns and connections between words," much in the way a human learns.[35] While Anthropic's self-serving anthropomorphizing of its models is clearly misplaced, at a minimum, humans who learn from books buy lawful copies of them, or borrow them from libraries that buy them, providing at least some measure of compensation to authors and creators. Anthropic does not, and it has usurped authors' content for the purpose of creating a machine built to generate the very type of content for which authors would usually be paid.

45. Anthropic, in taking authors' works without compensation, has deprived authors of

[31] "A General Language Assistant as a Laboratory for Alignment," Amodei et al, Dec. 9, 2021, p. 27-28, https://arxiv.org/pdf/2112.00861 (last visited Aug. 15, 2024).

[32] *See* "Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI," *ProofNews*, Jul. 16, 2024, https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-swiped-youtube-videos-to-train-ai (last visited Aug. 15, 2024).

[33] *See* "It's practically impossible to run a big AI company ethically," *Vox.com,* https://www.vox.com/future-perfect/364384/its-practically-impossible-to-run-a-big-ai-company-ethically (last updated Aug. 5, 2024) ("Anthropic acknowledges that it trained its chatbot, Claude, using the Pile . . .").

[34] *See* "De-Cop Detecting Copyrighted Content in Language Models Training Data," Duarte et al, Feb. 15, 2024, https://arxiv.org/html/2402.09910v1#S1 (last visited Aug. 15, 2024).

[35] "How do you use personal data in model training?," *Anthropic*, https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training (last visited Aug. 15, 2024).

books sales and licensing revenues. There has long been an established market for the sale of books and e-books, yet Anthropic ignored it and chose to scrape a massive corpus of copyrighted books from the internet, without even paying for an initial copy.

46. Anthropic has also usurped a licensing market for copyright owners. In the last two years, a thriving licensing market for copyrighted training data has developed. A number of AI companies, including OpenAI, Google, and Meta, have paid hundreds of millions of dollars to obtain licenses to reproduce copyrighted material for LLM training. These include deals with Axel Springer, News Corporation, the Associated Press, and others. Furthermore, absent Anthropic's largescale copyright infringement, blanket licensing practices would be possible through clearinghouses, like the Copyright Clearance Center, which recently launched a collective licensing mechanism that is available on the market today.[36]

47. Anthropic, however, has chosen to use Plaintiffs works and the works owned by the Class free of charge, and in doing so has harmed the market for the copyrighted works by depriving them of book sales and licensing revenue.

III. Anthropic Has Profited From Its Unlicensed Exploitation of Copyrighted Material At the Expense of Authors

48. Anthropic's Claude and other LLMs like it seriously threaten the livelihood of the very authors—including Plaintiffs here, as discussed specifically below—on whose works they were "trained."

49. Goldman Sach's estimates that generative AI could replace 300 million full-time jobs in the near future, or one-fourth of labor currently performed in the United States and Europe.

50. Already, writers report losing income from copywriting, journalism, and online content writing, which are important sources of income for book authors. The Authors Guild, the oldest professional organization representing writers and authors, recently published an earnings study that shows a median writing-related income for full-time authors of just over $20,000, and

[36] Copyright Clearance Center, The Intersection of AI & Copyright, https://www.copyright.com/resource-library/insights/intersection-ai-copyright/ (last visited Aug. 15 2024).

that full-time traditional authors earn only half of that from their books.[37] The rest comes from activities like content writing—work that is starting to dry up as a result of generative AI systems trained on those writers' works, without compensation, to begin with.

51. Since the explosion of LLM use in 2023, which coincided with the release of Claude, there has been an explosion of AI-generated books. When journalist Kara Swisher released her memoir *Burn Book* earlier this year, Amazon was flooded AI generated copycats. This was not an isolated incident. In another instance, author Jane Friedman discovered "a cache of garbage books" written under her name for sale on Amazon.[38] As LLMs have become more advanced—and enabled to train on more and more copyrighted material—they are able to generate more content and more sophisticated content. The result is that it is easier than ever to generate rip-offs of copyrighted books that compete with the original, or at a minimum dilute the market for the original copyrighted work.

52. Claude in particular has been used to generate cheap book content. For example, in May 2023, it was reported that a man named Tim Boucher had "written" 97 books using Anthropic's Claude (as well as OpenAI's ChatGPT) in less than year, and sold them at prices from $1.99 to $5.99.[39] Each book took a mere "six to eight hours" to "write" from beginning to end.[40] Claude could not generate this kind of long-form content if it were not trained on a large quantity of books, books for which Anthropic paid authors nothing.

53. In short, the success and profitability of Anthropic is predicated on mass copyright infringement without a word of permission from or a nickel of compensation to copyright owners, including Plaintiffs here.

---

[37] Authors Guild, "Top Takeaways from the 2023 Author Income Survey (2023), available at https://authorsguild.org/news/key-takeaways-from-2023-author-income-survey/#:~:text=Though%20overall%20author%20incomes%20are,coming%20in%20a%20close%20second (last visited Aug. 15, 2024).

[38] Tweet by Jane Friedman, Aug. 6, 2023, https://x.com/JaneFriedman/status/1688180228206530560 (last visited Aug. 15, 2024).

[39] *See* "AI Author? Man Writes Nearly 100 Books Using ChatGPT and Claude, Earns Roughly $2,000 Through Them," *Science Times*, May 18, 2023, https://www.sciencetimes.com/articles/43848/20230518/ai-author-man-writes-nearly-100-books-using-chatgpt-claude.htm (last visited Aug. 15, 2024).

[40] *Id. See also* "I'm Making Thousands Using AI to Write Books," Tim Boucher, https://www.newsweek.com/ai-books-art-money-artificial-intelligence-1799923 (last visited Aug. 15, 2024).

IV. Anthropic Exploited Each of Plaintiffs' Copyrighted Works

54. Each author, both Plaintiffs and Class Members, has a distinct voice, style, and creative mode expression. But all Plaintiffs and Class Members have suffered identical harms from Anthropic's infringement.

55. The contents of the datasets that Anthropic used to "train" its LLMs are peculiarly within its knowledge, such that Plaintiffs are unable to discern those contents with perfect accuracy. But Anthropic has admitted to using The Pile to train Claude, which included Books3 during the relevant time period, and the contents of Books3 is widely reported. Plaintiffs make specific allegations of infringement below based on what is known about Anthropic's training practices; what is known about the contents, uses, and availability of pirated book repositories that it is suspected Anthropic used, like Bibliotik; and the results of Plaintiffs' testing of Claude.

56. ***Plaintiff Bartz***. Plaintiff Bartz is the author of a number of books, including *The Lost Night: A Novel*. This novel was included in the Books3 dataset, based on public reporting about the dataset. Pirated copies of her work are available online through websites like LibGen and Bibliotek. Bartz is the author and owner of the registered copyright listed in Exhibit A, under the name of her wholly owned S Corporation, Andrea Bartz Inc.

57. ***Plaintiff Graeber***. Plaintiff Graeber is the author of a number of books, including *The Good Nurse* and *The Breakthrough*. Both books are part of the Books3 dataset, based on public reporting about that dataset. Pirated copies of all of Plaintiff Graeber's works are available online through websites like LibGen and Bibliotik. Graeber is the author and owner of the registered copyrights listed under his name in Exhibit A.

58. ***Plaintiff Johnson***. Plaintiff Johnson is the author of a number of books, including *To Be a Friend Is Fatal*. This book is part of the Books3 dataset, based on public reporting about that dataset. Pirated copies of all of Plaintiff Johnson's works are available online through websites like LibGen and Bibliotik. Johnson is the author and legal and/or beneficial owner of the registered copyright listed under his name in Exhibit A.

CLASS ALLEGATIONS

59. This action is brought by Plaintiffs individually and on behalf of the Class, as

defined below, pursuant to Rule 23(a), (b)(3) and 23(b)(2), (c)(4), and (g) of the Federal Rules of Civil Procedure:

> All natural persons, estates, or literary trusts that are legal or beneficial owners of copyrighted works that: (a) are registered with the United States Copyright Office; (b) were or are used by Defendant in LLM training, research, or development, including but not limited to training Defendant's Claude family of models; and (c) are books. The Class excludes Defendant, its officers and directors, members of their immediate families, their co-conspirators, aiders and abettors, and the heirs, successors or assigns of any of the foregoing.

60. The Class consists of at least thousands of authors and copyright holders and thus is so numerous that joinder of all members is impractical. The identities of members of the Class can be readily ascertained from business records maintained by Defendant and at a minimum from the content of the Books3 database that Anthropic illegally downloaded.

61. The claims asserted by Plaintiff are typical of the claims of the Class, all of whose works were also copied as part of the LLM training process.

62. The Plaintiff will fairly and adequately protect the interests of the Class and does not have any interests antagonistic to those of other members of the Class.

63. The Plaintiff has retained attorneys who are knowledgeable and experienced in copyright and class action matters, as well as complex litigation.

64. This action is appropriate as a class action pursuant to Rule 23(b)(3) of the Federal Rules of Civil Procedure because common questions of law and fact affecting the Class predominate over those questions affecting only individual members. The law is uniform. And, the common factual questions giving rise to common answers that move this litigation forward include:

a. Whether Anthropic's reproduction of the Class's copyrighted work constituted copyright infringement;

b. Whether Anthropic's reproduction of the Class's copyrighted work in the course of training their generative AI models was fair use;

c. Whether Anthropic's reproduction of the Class's copyrighted work harmed

Class member and whether Class member is entitled to damages, including statutory damages and the amount of statutory damages; and

d. Whether Anthropic's infringement was willful.

65. In addition, the class device is the superior mechanism for handling this action, and a class trial is eminently manageable.

66. This action is also appropriate as a class action pursuant to Rule 23(b)(2) of the Federal Rules of Civil Procedure because Anthropic's decision to train its models on a large trove of the Class's books affects all class members in the same way, and any injunctive relief awarded will affect the Class as a whole.

67. Finally, at the very minimum, there are multiple common issues relating to Anthropic's uniform context, such as (but not limited to) their ingestion, reproduction, and willfulness.

CLAIM FOR RELIEF: Copyright Infringement (17 U.S.C. § 501)

Against Defendant Anthopic PBC

68. Plaintiffs incorporate by reference the allegations in Paragraphs 1 to 65 as though fully set forth herein.

69. Plaintiffs and members of the proposed Class have created literary works that are original and fixed in a tangible medium of expression, and they own the registered copyrights in the works that Anthropic reproduced and appropriated to train their artificial intelligence models.

70. Plaintiff and members of the proposed Class therefore hold the exclusive rights, including the rights of reproduction and distribution, to those works under 17 U.S.C. § 106.

71. Anthropic infringed on the exclusive rights, under 17 U.S.C. § 106, of Plaintiff and members of the proposed Class by, among other things, reproducing the works owed by Plaintiff and the proposed Class in datasets used to train their artificial intelligence models.

72. On information and belief, Anthropic's infringing conduct alleged herein was and continues to be willful. Anthropic infringed on the exclusive rights of Plaintiffs and members of the proposed Class knowing that they were profiting from mass copyright infringement.

73. Plaintiffs and members of the proposed Class are entitled to, at their election,

statutory damages or, actual damages, disgorgement, attorneys' fees and costs, and other remedies available under the Copyright Act.

74. Plaintiffs and members of the proposed Class have been and continue to be irreparably injured due to Anthropic's conduct, for which there is no adequate remedy at law. Anthropic will continue to infringe on the exclusive right of Plaintiffs and the proposed class unless their infringing activity is enjoined by this Court. Plaintiffs are therefore entitled to permanent injunctive relief barring Anthropic's ongoing infringement.

PRAYER FOR RELIEF

WHEREFORE, Plaintiffs seek that this matter be certified as a class action, and that their attorneys be appointed Class Counsel and that they be appointed Class Representatives, and Plaintiffs demand judgment against Defendant as follows:

1. Awarding Plaintiffs and the proposed Class statutory damages or compensatory damages at their election, restitution, disgorgement, attorneys' fees and costs, and any other relief that may be permitted by law or equity pursuant to the claim(s) for relief;

2. Permanently enjoining Anthropic from engaging in the infringing conduct alleged herein;

3. Awarding Plaintiffs and the proposed Class pre-judgment and post-judgment interest pursuant to the claim(s) for relief;

4. Awarding Plaintiffs and the proposed Class costs, expenses, and attorneys' fees as permitted by law; and

5. Awarding Plaintiffs and the proposed Class further relief as the Court may deem just and proper under the circumstances.

DEMAND FOR JURY TRIAL

Pursuant to Rule 38 of the Federal Rules of Civil Procedure, Plaintiffs hereby demand a jury trial for all claims so triable.

Dated: August 19, 2024

By: /s/ *Rohit D. Nath*

Justin A. Nelson*
Alejandra C. Salinas*
SUSMAN GODFREY L.L.P
1000 Louisiana Street, Suite 5100
Houston, TX 77002-5096
Telephone: (713) 651-9366
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (SBN 316062)
SUSMAN GODFREY L.L.P
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067-2906
Telephone: (310) 789-3100
RNath@susmangodfrey.com

Jordan W. Connors*
SUSMAN GODFREY L.L.P
401 Union Street, Suite 3000
Seattle, WA 98101
Telephone: (206) 516-3880
jconnors@susmangodfrey.com

J. Craig Smyser*
SUSMAN GODFREY L.L.P
One Manhattan West, 51st Floor,
New York, NY 10019
Telephone: (212) 336-8330
csmyser@susmangodfrey.com

Rachel Geman*
Wesley Dozier*
Anna Freymann*
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, New York 10013-1413
Telephone: (212) 355-9500
rgeman@lchb.com
wdozier@lchb.com
afreymann@lchb.com

Reilly T. Stoler (SBN 310761)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111-3339
Telephone: (415) 956-1000
rstoler@lchb.com

Scott J. Shoulder*
CeCe M. Cole*
COWAN DEBAETS ABRAHAMS & SHEPPARD LLP
60 Broad Street, 30th Floor
New York, New York 10010
Telephone: (212) 974-7474
sshoulder@cdas.com
ccole@cdas.com

Attorneys for Plaintiffs and the Proposed Class

**(Pro Hac Vice forthcoming)*