

1 Justin A. Nelson (*pro hac vice*)  
Alejandra C. Salinas (*pro hac vice*)  
2 Collin Fredricks (*pro hac vice*)  
SUSMAN GODFREY L.L.P  
3 1000 Louisiana Street, Suite 5100  
Houston, TX 77002-5096  
4 Telephone: (713) 651-9366  
jnelson@susmangodfrey.com  
5 asalinas@susmangodfrey.com  
cfredricks@susmangodfrey.com

6 Rohit D. Nath (SBN 316062)  
SUSMAN GODFREY L.L.P  
7 1900 Avenue of the Stars, Suite 1400  
Los Angeles, CA 90067-2906  
8 Telephone: (310) 789-3100  
RNath@susmangodfrey.com

9 J. Craig Smyser (*pro hac vice*)  
SUSMAN GODFREY L.L.P  
10 One Manhattan West, 51<sup>st</sup> Floor,  
New York, NY 10019  
11 Telephone: (212) 336-8330  
Facsimile: (212) 336-8340  
12 csmysr@susmangodfrey.com

Jordan W. Connors (*pro hac vice*)  
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*)  
Wesley Dozier (*pro hac vice*)  
Anna Freymann (*pro hac vice*)  
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

13  
14 *Attorneys for Plaintiffs and the Proposed Class*  
15 *(Additional Counsel Listed on Signature Page)*

16  
17 **UNITED STATES DISTRICT COURT**  
18 **NORTHERN DISTRICT OF CALIFORNIA**  
19 **SAN FRANCISCO DIVISION**

20  
21 ANDREA BARTZ, ANDREA BARTZ, INC.,  
22 CHARLES GRAEBER, KIRK WALLACE  
JOHNSON, and MJ + KJ, INC., individually and  
23 on behalf of others similarly situated,

24 Plaintiffs,

25 v.

26 ANTHROPIC PBC,

27 Defendants.  
28

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

**FIRST AMENDED CLASS ACTION  
COMPLAINT**

**JURY TRIAL DEMANDED**

1 Plaintiffs Andrea Bartz, Andrea Bartz, Inc., Charles Graeber, Kirk Wallace Johnson, and  
2 MJ + KJ, Inc., on behalf of themselves and all other similarly situated (the “Class,” as defined  
3 below), for their complaint against Defendant Anthropic PBC (“Anthropic”), allege as follows:

4 **NATURE OF THE ACTION**

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

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

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

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

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

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

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

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

24 \_\_\_\_\_  
25 <sup>1</sup> “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).

26 <sup>2</sup> “Amazon Injects an Additional \$2.75B Vote of Confidence in Anthropic,” *Spice Works*, Mar. 28,  
27 2024, <https://www.spiceworks.com/tech/artificial-intelligence/news/amazon-invests-4-billion-anthropic/> (last visited Aug. 15, 2024).

28 <sup>3</sup> “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).

1 **JURISDICTION AND VENUE**

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

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

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

10 11. Venue is proper under 28 U.S.C. § 1400(a) because Anthropic or its agents reside  
11 or may be found in this District due to their infringing activities, along with their commercialization  
12 of their infringing activities, that occurred in this District. Venue is also proper under 28 U.S.C.  
13 § 1391(b)(2) because a substantial part of the events giving rise to Plaintiffs’ claims occurred in  
14 this District, including the largescale copyright infringement and commercialization of Anthropic’s  
15 LLMs.

16 **THE PARTIES**

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

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

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

1 15. Plaintiff Andrea Bartz, Inc. is a loan-out corporation wholly owned by Plaintiff  
2 Bartz.

3 16. Plaintiff MJ + KJ, Inc. is a loan-out corporation wholly owned by Plaintiff Johnson  
4 and his wife, Marie-Josée Cantin Johnson.

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

### 7 **FACTUAL ALLEGATIONS**

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

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

14 19. Anthropic released the first iteration of its flagship model Claude in March 2023,  
15 shortly after OpenAI's ChatGPT took the world by storm.<sup>5</sup> A few months later, in July 2023,  
16 Anthropic released the next iteration of the model, Claude 2, for public use.<sup>6</sup> Then, in March 2024,  
17 Anthropic released Claude 3.<sup>7</sup> Claude 3 was released with three levels, ranging from simplest to  
18 most advanced: Claude Sonnet, Claude Haiku, and Claude Opus.<sup>8</sup> In June of this year, Anthropic  
19 released Claude 3.5 Sonnet.<sup>9</sup> And most recently, Anthropic released Claude 3.5 Haiku in October  
20 2024.<sup>10</sup> Claude is available for use via web interface, on Android and iOS applications, and via an  
21 application programming interface, which allows developers to build custom generative AI tools

22 \_\_\_\_\_  
23 <sup>4</sup> LinkedIn Page for Dario Amodei, <https://www.linkedin.com/in/dario-amodei-3934934/> (last  
visited Aug. 15, 2024).

24 <sup>5</sup> 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).

25 <sup>6</sup> *Id.*

26 <sup>7</sup> See "Introducing the next generation of Claude," *Anthropic*, Mar. 4, 2024,  
<https://www.anthropic.com/news/claude-3-family> (last visited Aug. 15, 2024).

27 <sup>8</sup> *Id.*

28 <sup>9</sup> See "Claude 3.5 Sonnet," *Anthropic*, Jun. 20, 2024, <https://www.anthropic.com/news/claude-3-5-sonnet> (last visited Aug. 15, 2024).

<sup>10</sup> See "Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku," Oct. 22, 2024,  
<https://www.anthropic.com/news/3-5-models-and-computer-use> (last visited Dec. 4, 2024).

1 using Claude as the base.<sup>11</sup>

2 20. According to Anthropic, Claude outperforms other competing LLMs on the market,  
3 such as OpenAI's ChatGPT and Google's Gemini. Anthropic boasts that Claude can be used to  
4 "draft everything from a text message or email to a screenplay or a novel."<sup>12</sup>

5 21. Claude has been enormously successful, with both the general public and with  
6 enterprises. Anthropic currently offers access to its Claude 3.5 Sonnet model free, with increased  
7 usage caps and access to models like Claude 3 Opus or Haiku gated with subscription fees of  
8 between \$20 to \$30 per month.<sup>13</sup> Claude has garnered tens of millions of monthly active users and  
9 has been incorporated into Amazon Bedrock (Amazon Web Services' platform for cloud AI  
10 services).<sup>14</sup> Anthropic has been particularly successful in courting a corporate client base, which  
11 generates paid subscription revenue. Claude's customers include "leading enterprises and startups"  
12 such as Slack, Zoominfo, Asama, Bridgewater, LexisNexis, and Jane Street Capital.<sup>15</sup>

13 22. Throughout its existence, Anthropic has cloaked itself in the rhetoric of "AI safety"  
14 and "responsibility."<sup>16</sup> Its actions, however, have made a mockery of its lofty goals. Anthropic's  
15 immense success has been built, in large part, on its largescale copyright theft. As alleged in more  
16 detail below, Anthropic's models were trained on troves of pirated material. Moreover, Anthropic  
17 could not have built a model capable of digesting whole books and generating complex text outputs  
18 without its exploitation of these works.

19  
20  
21  
22 <sup>11</sup> See "What interfaces can I use to access Claude?," *Anthropic*,  
23 <https://support.anthropic.com/en/articles/8114487-what-interfaces-can-i-use-to-access-claude> (last  
visited Aug. 15, 2024).

24 <sup>12</sup> "What are some things I can use Claude for?," *Anthropic*,  
25 <https://support.anthropic.com/en/articles/7996845-what-are-some-things-i-can-use-claude-for>  
(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).

26 <sup>13</sup> See "Pricing," *Anthropic*, <https://www.anthropic.com/pricing> (last visited Aug. 15, 2024).

27 <sup>14</sup> See "Anthropic's Claude in Amazon Bedrock," *Amazon Web Services*,  
<https://aws.amazon.com/bedrock/claude/> (last visited Aug. 15, 2024).

28 <sup>15</sup> "Customers," *Anthropic*, <https://www.anthropic.com/customers> (last visited Aug. 15, 2024).

<sup>16</sup> 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).

## 1 **II. Anthropic Engaged in Largescale Copyright Theft in Training Its LLMs**

### 2 **1. Large Language Models and the Training Process**

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

6 24. Claude is a complex web of mathematical functions comprised of a series of  
7 algorithms that break down input text into smaller pieces—words or portions of words, called  
8 “tokens”—then translate those pieces into “vectors,” or a sequence of numbers that is used to  
9 identify the token within the series of algorithms. Those vectors help place each token on a map,  
10 by identifying other tokens closely associated with the word. As described by Anthropic’s  
11 competitor OpenAI: “the process begins by breaking text down into roughly word-length ‘tokens,’  
12 which are converted to numbers. The model then calculates each token’s proximity to other tokens  
13 in the training data—essentially, how near one word appears in relation to any other word. These  
14 relationships between words reveal which words have similar meanings . . . and functions.”<sup>17</sup> As  
15 the model trains and digests more expression, the algorithms depicting the relationship between  
16 various tokens changes with it.

17 25. The model is trained on a massive corpus of text; without training, there is no LLM.  
18 As Anthropic has described it, “[l]arge language models such as Claude need to be ‘trained’ on text  
19 so that they can learn the patterns and connections between words. This training is important so  
20 that the model performs effectively and safely.”<sup>18</sup>

21 26. The model takes text inputs in the form of an incomplete phrase or passage, and  
22 attempts to complete the phrase, essentially a fill-in-the-blank quiz. The model compares its  
23 predicted phrase completion with the actual “correct” answer. The model then adjusts its internal  
24 algorithms to “learn” from its mistakes. In other words, it adjusts its algorithms to reduce the  
25

26 <sup>17</sup> See Comment of OpenAI “Re: Notice of Inquiry and Request for Comment [Docket No. 2023-  
27 06],” United States Copyright Office, Oct. 30, 2023, p. 5-6 (available at:  
[https://downloads.regulations.gov/COLC-2023-0006-8906/attachment\\_1.pdf](https://downloads.regulations.gov/COLC-2023-0006-8906/attachment_1.pdf)).

28 <sup>18</sup> “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](https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training) (last visited Aug. 15, 2024).

1 likelihood of making the same mistake again and thus minimizing the difference between any given  
2 text input and the “correct” text output.

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

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

18 29. For both the post- and pre-training processes in developing Claude, Anthropic  
19 created multiple, unlicensed copies of the training data. As the U.S. Patent and Trademark Office  
20 has observed, LLM “training” “almost by definition involve[s] the reproduction of entire works or  
21 substantial portions thereof.”<sup>19</sup>

22 30. The quality and quantity of the corpus is critical to the quality of the resulting model.  
23 With respect to LLM development, the phrase “garbage in, garbage out” carries weight. As one  
24 researcher put it: “[large language] model behavior is *not* determined by architecture,  
25 hyperparameters, or optimizer choices [i.e. technical features set during model training]. *It’s*  
26

27 <sup>19</sup> U.S. Patent & Trademark Office, *Public Views on Artificial Intelligence and Intellectual Property*  
28 *Policy 29* (2020), available at [https://www.uspto.gov/sites/default/files/documents/USPTO\\_AI-Report\\_2020-10-07.pdf](https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf) (last accessed Aug. 15, 2024).



1 *determined by your dataset, nothing else. Everything else is a means to an end in efficiently*  
2 *deliver[ing] compute to approximating that dataset.”<sup>20</sup>*

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

8 32. In this way, books are especially valuable training material. As one commentator  
9 put it, “[b]ooks offer formal and lengthy texts which help LLMs understand complex language  
10 structures, grasp long-term context, and produce coherent narratives.”<sup>21</sup>

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

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

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

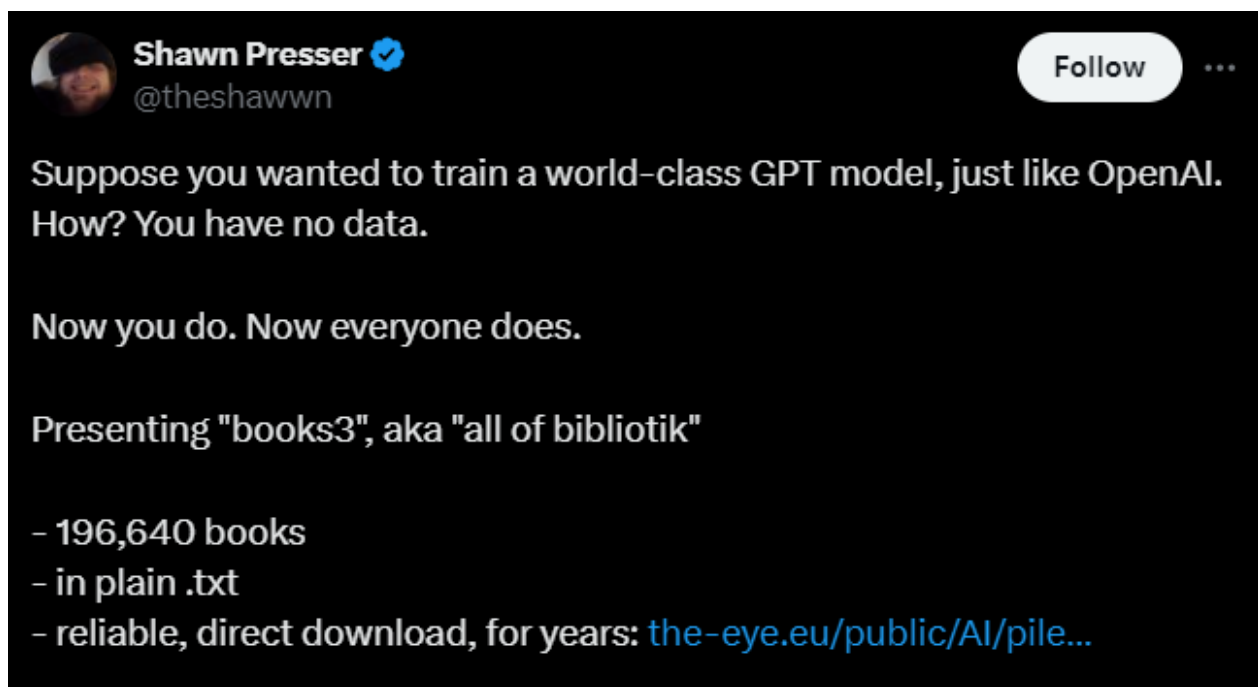
20 35. One of The Pile’s architects is an independent developer named Shawn Presser.  
21 Presser created a dataset included in The Pile called “Books3,” which is a trove of pirated books.

22  
23  
24  
25  
26 <sup>20</sup> 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).

27 <sup>21</sup> “Pretraining of Large Language Models,” Ritwik Raha, *Github.com*  
<https://gist.github.com/ritwikraha/77e79990992043f60a9588610b2781c5> (last visited Aug. 15, 2024).

28 <sup>22</sup> “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).

1 36. Presser described how he created Books3 in a Twitter thread from October 2020:<sup>23</sup>



13

14 37. Presser went on. He said he created Books3 in response to “OpenAI’s papers on

15 GPT-2 and 3,” which “refer[] to datasets named ‘books1’ and ‘books2,’” the latter of which Presser

16 suspects “might be ‘all of libgen.’”<sup>24</sup> LibGen refers to “Library Genesis,” a website offering pirated

17 books that was ordered shut down for copyright infringement in 2015. *See Elsevier, Inc. et al v.*

18 *www.Sci-hub.org et al*, 15-cv-2482-RWS, Dkt. 53 (Oct. 30, 2015). To create a pirated-book dataset

19 comparable to what he suspected OpenAI created for itself, Presser announced that Books3 was

20 also a direct download of all books from a *different pirated website*—a compilation of “196,640

21 books,” which comprises “all of bibliotik.”<sup>25</sup>

22 38. Bibliotik is a “notorious pirated collection” of “pirated books.”<sup>26</sup> For years prior to

23

24 <sup>23</sup> See Tweet by Shawn Presser, Oct. 25, 2020, <https://x.com/theshawwn/status/1320282149329784833?lang=en> (last visited Aug. 15, 2024).

25 <sup>24</sup> See Tweet by Shawn Presser, Oct. 25, 2020, <https://x.com/theshawwn/status/1320282152689336320> (last visited Aug. 15, 2024).

26 <sup>25</sup> See Tweet by Shawn Presser, Oct. 25, 2020, <https://x.com/theshawwn/status/1320282149329784833?lang=en> (last accessed Aug. 15, 2024).

27 <sup>26</sup> See Schoppert, “Whether you’re an undergraduate doing research, or a fan of the Nick Stone

28 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,

1 its use as “Books3,” Bibliotik was frequently included in roundups of the best—and most popular—  
2 sources for pirated material.<sup>27</sup>

3 39. Books3 was a critical part of The Pile. In EleutherAI’s paper on The Pile, it touted  
4 the key value of Books3 as training material: “Books3 is a dataset of books derived from a copy of  
5 the contents of the Bibliotik private tracker . . . Bibliotik consists of a mix of fiction and nonfiction  
6 books and is almost an order of magnitude larger than our next largest book dataset  
7 (BookCorpus2).” The paper then summarized its key point for why The Pile included this known  
8 source of illegal copyright material: “*We included Bibliotik because books are invaluable for  
9 long-range context modeling research and coherent storytelling.*”<sup>28</sup>

10 40. At the same time, Presser and EleutherAI repeatedly and publicly acknowledged  
11 that, with The Pile and Books3, they were making available a cache of pirated material.  
12 EleutherAI’s paper on The Pile noted that “there is little acknowledgment of the fact that the  
13 processing and distribution of data owned by others may also be a violation of copyright law.”<sup>29</sup>  
14 Furthermore, The Pile’s datasheet notes that “Books3 is almost entirely comprised of copyrighted

15 \_\_\_\_\_  
16 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  
17 in the past 20 years.”); “Revealed: The Authors Whose Pirated Books are Powering Generative  
18 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,  
19 that circulate via the BitTorrent file-sharing network.”); “Are ChatGPT, Bard and Dolly 2.0 Trained  
20 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.”).

21 <sup>27</sup>See Commit History of “Awesome Piracy,” *Github.com*, Oct. 13, 2018, [https://github.com/aviranzeroniac/awesome-piracy/commit/61928765e3ee0b4f3dbe3c0724b196e5f0f17e59?short\\_path=5a831ea#diff-5a831ea67cf5cf8703b0de46901ab25bd191f56b320053be9332d9a3b0d01d15](https://github.com/aviranzeroniac/awesome-piracy/commit/61928765e3ee0b4f3dbe3c0724b196e5f0f17e59?short_path=5a831ea#diff-5a831ea67cf5cf8703b0de46901ab25bd191f56b320053be9332d9a3b0d01d15) (October 13, 2018  
22 commit to “awesome piracy” repo listing “Bibliotik Popular ebooks/audiobooks private  
23 tracker”); “Reddit Piracy Megathread” repo., *Github.com*, Mar. 21, 2019  
24 <https://github.com/magicoflolis/Reddit-Piracy-Megathread/blob/master/data/findingtextbooks.md> (March 21, 2019 guide from “r/piracy” on how  
25 to source textbooks listing “Bibliotik”); “List of free eBook download sites,” *Pirates-forum.org*,  
26 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”).

27 <sup>28</sup> “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 <sup>29</sup> *Id.* at 14-15.

1 works . . . .”<sup>30</sup> Presser, for his part, has admitting to releasing Books3 despite “fear of copyright  
2 backlash.”<sup>31</sup>

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

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

11 43. In a December 2021 research paper on large language model training, Anthropic  
12 described creating a dataset “most of which we sourced from the Pile” and which included “32%  
13 internet books,” a code word in the industry for pirated copies of books available on the internet.<sup>32</sup>

14 44. More recently, in July 2024, Anthropic has publicly acknowledged that it used The  
15 Pile to train its Claude models. As reported by Proof News, company spokesperson Jennifer  
16 Martinez “confirm[ed] use of the Pile in Anthropic’s generative AI assistant Claude.”<sup>33</sup> Anthropic  
17 confirmed the same to Vox News.<sup>34</sup> Independent researchers have tested Claude to shed light on  
18 the composition of its training set, and their work has confirmed a high likelihood that Claude was  
19 trained on copyrighted books.<sup>35</sup>

20  
21 <sup>30</sup> “Datasheet for the Pile,” Gao et al, Jan. 20, 2022, p. 15, <https://arxiv.org/pdf/2201.07311> (last  
visited Aug. 15, 2024).

22 <sup>31</sup> Comment of “sillysaurusx,” *Hacker News*, Jul. 11, 2023,  
<https://news.ycombinator.com/item?id=36685115> (last visited Aug. 15, 2024).

23 <sup>32</sup> “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).

24 <sup>33</sup> See “Apple, Nvidia, Anthropic Used Thousands of Swiped YouTube Videos to Train AI,”  
25 *ProofNews*, Jul. 16, 2024, [https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-  
swiped-youtube-videos-to-train-ai](https://www.proofnews.org/apple-nvidia-anthropic-used-thousands-of-swiped-youtube-videos-to-train-ai) (last visited Aug. 15, 2024).

26 <sup>34</sup> 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](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,  
27 using the Pile . . .”).

28 <sup>35</sup> 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).

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

6           46. Anthropic’s commercial copying of Plaintiffs’ work and works owned by the  
7 proposed Class was manifestly unfair use, for several reasons. Anthropic has suggested that it uses  
8 the training data to “learn the patterns and connections between words,” much in the way a human  
9 learns.<sup>36</sup> While Anthropic’s self-serving anthropomorphizing of its models is clearly misplaced, at  
10 a minimum, humans who learn from books buy lawful copies of them, or borrow them from  
11 libraries that buy them, providing at least some measure of compensation to authors and creators.  
12 Anthropic does not, and it has usurped authors’ content for the purpose of creating a machine built  
13 to generate the very type of content for which authors would usually be paid.

14           47. Anthropic, in taking authors’ works without compensation, has deprived authors of  
15 books sales and licensing revenues. There has long been an established market for the sale of books  
16 and e-books, yet Anthropic ignored it and chose to scrape a massive corpus of copyrighted books  
17 from the internet, without even paying for an initial copy.

18           48. Anthropic has also usurped a licensing market for copyright owners. In the last two  
19 years, a thriving licensing market for copyrighted training data has developed. A number of AI  
20 companies, including OpenAI, Google, and Meta, have paid hundreds of millions of dollars to  
21 obtain licenses to reproduce copyrighted material for LLM training. These include deals with Axel  
22 Springer, News Corporation, the Associated Press, and others. Furthermore, absent Anthropic’s  
23 largescale copyright infringement, blanket licensing practices would be possible through  
24 clearinghouses, like the Copyright Clearance Center, which recently launched a collective licensing  
25 mechanism that is available on the market today.<sup>37</sup>

26 \_\_\_\_\_  
27 <sup>36</sup> “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](https://support.anthropic.com/en/articles/7996885-how-do-you-use-personal-data-in-model-training) (last visited Aug. 15, 2024).

28 <sup>37</sup> Copyright Clearance Center, *The Intersection of AI & Copyright*,

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

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

6           50. Anthropic’s Claude and other LLMs like it seriously threaten the livelihood of the  
7 very authors—including Plaintiffs here, as discussed specifically below—on whose works they  
8 were “trained.”

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

11           52. Already, writers report losing income from copywriting, journalism, and online  
12 content writing, which are important sources of income for book authors. The Authors Guild, the  
13 oldest professional organization representing writers and authors, recently published an earnings  
14 study that shows a median writing-related income for full-time authors of just over \$20,000, and  
15 that full-time traditional authors earn only half of that from their books.<sup>38</sup> The rest comes from  
16 activities like content writing—work that is starting to dry up as a result of generative AI systems  
17 trained on those writers’ works, without compensation, to begin with.

18           53. Since the explosion of LLM use in 2023, which coincided with the release of Claude,  
19 there has been an explosion of AI-generated books. When journalist Kara Swisher released her  
20 memoir *Burn Book* earlier this year, Amazon was flooded AI generated copycats. This was not an  
21 isolated incident. In another instance, author Jane Friedman discovered “a cache of garbage books”  
22 written under her name for sale on Amazon.<sup>39</sup> As LLMs have become more advanced—and enabled  
23 to train on more and more copyrighted material—they are able to generate more content and more

24  
25 <https://www.copyright.com/resource-library/insights/intersection-ai-copyright/> (last visited Aug.  
15 2024).

26 <sup>38</sup> Authors Guild, “Top Takeaways from the 2023 Author Income Survey (2023), available at  
27 <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).

28 <sup>39</sup> Tweet by Jane Friedman, Aug. 6, 2023,  
<https://x.com/JaneFriedman/status/1688180228206530560> (last visited Aug. 15, 2024).

1 sophisticated content. The result is that it is easier than ever to generate rip-offs of copyrighted  
2 books that compete with the original, or at a minimum dilute the market for the original copyrighted  
3 work.

4 54. Claude in particular has been used to generate cheap book content. For example, in  
5 May 2023, it was reported that a man named Tim Boucher had “written” 97 books using  
6 Anthropic’s Claude (as well as OpenAI’s ChatGPT) in less than year, and sold them at prices from  
7 \$1.99 to \$5.99.<sup>40</sup> Each book took a mere “six to eight hours” to “write” from beginning to end.<sup>41</sup>  
8 Claude could not generate this kind of long-form content if it were not trained on a large quantity  
9 of books, books for which Anthropic paid authors nothing.

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

#### 13 **IV. Anthropic Exploited Each of Plaintiffs’ Copyrighted Works**

14 56. Each author, both Plaintiffs and Class Members, has a distinct voice, style, and  
15 creative mode expression. But all Plaintiffs and Class Members have suffered identical harms from  
16 Anthropic’s infringement.

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

24 58. ***Plaintiff Bartz.*** Plaintiff Bartz is the author of a number of books, including *The*

25 \_\_\_\_\_  
26 <sup>40</sup> See “AI Author? Man Writes Nearly 100 Books Using ChatGPT and Claude, Earns Roughly  
27 \$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).

28 <sup>41</sup> *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).

1 *Lost Night: A Novel*. This novel was included in the Books3 dataset, based on public reporting  
2 about the dataset. Plaintiff Bartz is also the author of *The Herd, We Were Never Here*, and *The*  
3 *Spare Room*. Pirated copies of Plaintiff Bartz's works are available online through websites like  
4 LibGen, Internet Archive, and Bibliotik. Bartz is the author and legal and/or beneficial owner of  
5 the registered copyrights listed in Exhibit A, under the name of her wholly owned S Corporation,  
6 Andrea Bartz Inc.

7 59. **Plaintiff Graeber**. Plaintiff Graeber is the author of a number of books, including  
8 *The Good Nurse: A True Story of Medicine, Madness, and Murder* and *The Breakthrough:*  
9 *Immunotherapy and the Race to Cure Cancer*. Both books are part of the Books3 dataset, based on  
10 public reporting about that dataset. Pirated copies of Plaintiff Graeber's works are available online  
11 through websites like LibGen, Internet Archive, and Bibliotik. Graeber is the author and owner of  
12 the registered copyrights listed under his name in Exhibit A.

13 60. **Plaintiff Johnson**. Plaintiff Johnson is the author of a number of books, including  
14 *To Be a Friend Is Fatal*. This book is part of the Books3 dataset, based on public reporting about  
15 that dataset. Plaintiff Johnson is also the author of *The Fishermen and the Dragon: Fear, Greed,*  
16 *and a Fight for Justice on the Gulf Coast*, and *The Feather Thief: Beauty, Obsession, and the*  
17 *Natural History Heist of the Century*. Pirated copies of Plaintiff Johnson's works are available  
18 online through websites like LibGen, Internet Archive, and Bibliotik. Johnson is the author and  
19 legal and/or beneficial owner of the registered copyrights listed in Exhibit A. *To Be a Friend is*  
20 *Fatal* is registered under Johnson's name; *The Fishermen and the Dragon: Fear, Greed, and a*  
21 *Fight for Justice on the Gulf Coast*, and *The Feather Thief: Beauty, Obsession, and the Natural*  
22 *History Heist of the Century* are registered under the name of the S Corporation, MJ + KJ, Inc.,  
23 which he and his wife, Marie-Josée Cantin Johnson, wholly own.

24 61. **Plaintiff Andrea Bartz, Inc.** Plaintiff Andrea Bartz, Inc. is a loan-out corporation  
25 wholly owned by Plaintiff Bartz. Plaintiff Andrea Bartz, Inc. is the legal and/or beneficial owner  
26 of the registered copyrights authored by Andrea Bartz listed in Exhibit A.

27 62. **Plaintiff MJ + KJ, Inc.** Plaintiff MJ + KJ, Inc. is a loan-out corporation wholly  
28 owned by Plaintiff Johnson and his wife, Marie-Josée Cantin Johnson. Plaintiff MJ + KJ, Inc. is



1 the legal and/or beneficial owner of the registered copyrights authored by Kirk Wallace Johnson  
2 listed in Exhibit A.

3 **CLASS ALLEGATIONS**

4 63. This action is brought by Plaintiffs individually and on behalf of the Class, as  
5 defined below, pursuant to Rule 23(a), (b)(3) and 23(b)(2), (c)(4), and (g) of the Federal Rules of  
6 Civil Procedure:

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

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

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

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

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

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

28 a. Whether Anthropic's reproduction of the Class's copyrighted work

- 1 constituted copyright infringement;
- 2 b. Whether Anthropic’s reproduction of the Class’s copyrighted work in the
- 3 course of training their generative AI models was fair use;
- 4 c. Whether Anthropic’s reproduction of the Class’s copyrighted work harmed
- 5 Class member and whether Class member is entitled to damages, including
- 6 statutory damages and the amount of statutory damages; and
- 7 d. Whether Anthropic’s infringement was willful.

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

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

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

17 **CLAIM FOR RELIEF: Copyright Infringement (17 U.S.C. § 501)**  
18 **Against Defendant Anthropic PBC**

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

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

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

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



1 jury trial for all claims so triable.

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

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

Dated: December 4, 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, 51<sup>st</sup> 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

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

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

*Attorneys for Plaintiffs and the Proposed Class*