

Elizabeth Brannen (SBN 226234)
ebrannen@stris.com
John Stokes (SBN 310847)
jstokes@stris.com
Lauren Martin (SBN 294367)
lmartin@stris.com
STRIS & MAHER LLP
17785 Center Court Dr N, Ste 600
Cerritos, CA 90703
T: (213) 995-6800
F: (213) 261-0299

Kyle Roche (*pro hac vice* forthcoming)
kroche@fnf.law
Devin (Velvel) Freedman (*pro hac vice*
forthcoming)
vel@fnf.law
Alex Potter (*pro hac vice* forthcoming)
apotter@fnf.law
**FREEDMAN NORMAND
FRIEDLAND LLP**
155 E. 44th Street, Ste 915
New York, NY 10017
T: (646) 494-2900

Counsel for Plaintiffs

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

R. O. KWON, ALEXANDER DREIHANN-
HOLENIA, STEPHANIE GRIEST,
ELIZABETH HINTON, BRIAN JACKSON,
JASON ALAN JANKOVSKY, BRANDEN W.
JOSEPH, KIMBERLY KAY HOANG,
GLENN KREISBERG, RICHARD
KREVOLIN, GERALD LAMB, CHANG-
RAE LEE, JEANNIE LEVINSON, STEVEN
LOWENTHAL, MACKENZIE BLUE LLC,
BRIAN MALEC, DONNA MATNEY, ILONA
MCCARTY, TIFFANY MCDANIEL, BRIAN
MERCHANT, ERIC L. MULLER, SONIA
NAZARIO, DANIEL JOSÉ OLDER,
LAURENCE RALPH, LILLIAM RIVERA,
SETH A. ROBERTS, AIDA SALAZAR,
ERIKA L. SANCHEZ, JEREMY N. SMITH,
SOUL WRITERS, LLC, RAMESH
SRINIVASAN, ROBERT TANNENBAUM,
KENNETH WARNER, and ROBERT
ZIMMERMAN,

Plaintiffs,

v.

ANTHROPIC PBC,

Defendant.

Civil Case No.:

COMPLAINT

DEMAND FOR JURY TRIAL

1 1. Plaintiffs R. O. Kwon, Alexander Dreihann-Holenia, Stephanie Griest, Elizabeth
2 Hinton, Brian Jackson, Jason Alan Jankovsky, Branden W. Joseph, Kimberly Kay Hoang, Glenn
3 Kreisberg, Richard Krevolin, Gerald Lamb, Chang-rae Lee, Jeannie Levinson, Steven Lowenthal,
4 Mackenzie Blue LLC, Brian Malec, Donna Matney, Ilona McCarty, Tiffany McDaniel, Brian
5 Merchant, Eric L. Muller, Sonia Nazario, Daniel José Older, Laurence Ralph, Lilliam Rivera, Seth
6 A. Roberts, Aida Salazar, Erika L. Sanchez, Jeremy N. Smith, Soul Writers, LLC, Ramesh
7 Srinivasan, Robert Tannenbaum, Kenneth Warner, and Robert Zimmerman (collectively
8 “Plaintiffs”), each proceeding in their individual capacity, bring this action against Anthropic PBC
9 (“Anthropic” or “Defendant”), and allege as follows:

10 **I. INTRODUCTION**

11 2. This case concerns Defendant Anthropic’s exploitation of Plaintiffs’ copyrighted
12 works to build its “Claude” family of large language models (“LLMs”). Defendant is in open and
13 active competition with its tech company rivals to win what many have deemed the “generative AI
14 arms race.” Participants in this race are under immense pressure to build more, better, and faster,
15 with an eye towards ensuring that *its* generative AI models are widely adopted. Broad adoption of
16 a company’s generative AI models will yield great profits, the thinking goes; on the converse, failure
17 to move quickly means missing the boat on this trillion-dollar industry. To remain competitive in
18 this race, Anthropic turned to notorious online “shadow libraries” and similar pirated datasets to
19 obtain vast quantities of copyrighted materials—books, articles, training materials, and the like—
20 which it desperately wanted to train and optimize its LLM models. Anthropic could have—and
21 *should have*—paid copyright owners, including Plaintiffs, for licenses to use their copyrighted
22 works in connection with training its Claude models. Instead, Anthropic fed from, and then back
23 into, the illegal market for digitally available copyrighted materials. Ultimately, Anthropic
24 downloaded, reproduced, distributed, and defaced these copyrighted works, including Plaintiffs’
25
26
27
28

1 works, to get ahead in the generative AI arms race and improve its bottom line. Plaintiffs bring this
2 action to hold Anthropic accountable for its brazen acts of copyright infringement.¹

3 3. LLMs are a form of “generative artificial intelligence” or “generative AI.” These
4 models are designed to process and emit natural language. Over the last several years, many of the
5 world’s biggest and wealthiest technology companies have entered the market to develop, distribute,
6 and commercialize generative AI models, believing these and other forms of artificial intelligence
7 have massive potential for growth in revenue and profits. Because artificial intelligence, including
8 generative AI, is viewed by industry leaders as the next foundational layer of the digital economy,
9 the pressure to build more, better, and faster has led to the aforementioned AI arms race.

10 4. Against this backdrop, LLM developers needed data upon which to train their
11 models. AI researchers quickly discovered that copyrighted published content—*e.g.*, books, articles,
12 and training materials—is really the gold standard when it comes to LLM training material. Unlike
13 Internet content and many other forms of text, published material typically is structured, long-form,
14 and highly polished. These materials embody the expressive output of their creators, are thoughtfully
15 planned, and take weeks, months, and years to develop and publish. In simple terms, if you want to
16 teach a machine how to “speak” or write like a human—capable of telling stories, using analogies,
17 and making jokes—feed it material that mimics how we express ourselves in our most articulate and
18 complete forms.

19 5. For its Claude training datasets, Anthropic needed gold standard training data. To get
20 its hands on published content, it bypassed the licensing market and downloaded copyrighted works
21 from these shadow libraries. Anthropic often torrented them. “Torrenting” works by breaking a file
22 into many small pieces and distributing those pieces across a network of participating computers. A
23 user downloads portions of a file from numerous other computers that already possess the file and
24 software reassembles those pieces into a completed whole on the user’s machine.

25
26
27 ¹ Unless otherwise indicated, references to Anthropic’s “Claude” refers to all versions of Claude in
28 any stage of their development or deployment.

1 6. But torrenting protocols not only facilitate downloading, they also facilitate and
2 encourage uploading. And through Anthropic’s torrenting of copyrighted material from illicit
3 shadow libraries, Anthropic became a distributor of unauthorized copies of protected works. Stated
4 somewhat differently, Anthropic used torrenting protocols that facilitated its reuploading of
5 copyrighted materials into peer-to-peer file-sharing networks. In Copyright Act terms, Anthropic
6 distributed copyrighted materials without consent or authorization. And through this distribution,
7 Anthropic also contributed to further acts of infringement, allowing other users on these peer-to-
8 peer networks to unlawfully reproduce and distribute copyrighted works.

9 7. Anthropic not only torrented copyrighted materials from online shadow libraries, it
10 also purchased and scanned millions of physical books—reproducing their contents—without
11 authorization. Its downloading, torrenting, and scanning activities were done both to build and train
12 its Claude models, but also to build and maintain a central library that it intended to retain
13 indefinitely. In connection with its training of the Claude models, Anthropic also embedded near-
14 verbatim copies of copyrighted works, including Plaintiffs’ works, in Claude’s model weights.

15 8. In addition to ripping pirated copies, reproducing, and distributing without
16 authorization Plaintiffs’ copyrighted materials, Anthropic also stripped these materials of copyright
17 management information, enabling, facilitating, and concealing the infringement of these works.
18 Anthropic did this to optimize its models’ performance.

19 9. Adding insult to all of this injury, Anthropic’s use of Plaintiffs’ copyrighted materials
20 facilitated Anthropic’s creation of AI models capable of generating content that directly competes
21 with and will compete directly with Plaintiffs’ content. “Learning from” the creativity, research, and
22 expression embodied in thousands upon thousands of copyrighted works, including Plaintiffs’
23 works, Claude has the capability to flood the market with free and paid content that vies with
24 Plaintiffs for consumer attention.

1 10. The result of Anthropic’s use of copyrighted materials to train Claude? An estimated
2 \$1 trillion valuation as of April 2026.² The Copyright Act doesn’t allow Anthropic to get a free ride
3 on the backs of Plaintiffs and the other creators and publishers whose copyrighted works it exploited
4 to build its trillion-dollar generative AI enterprise. Plaintiffs bring this action to hold Anthropic
5 accountable for the infringement that enabled its rise in the generative-AI marketplace, and to
6 enforce the fundamental principle that creative expression cannot be taken, copied, or exploited
7 without permission or compensation.

8 11. To redress Anthropic’s repeated, unlawful, and *en masse* infringement of its works,
9 Plaintiffs seek (1) damages, (2) permanent injunctive relief barring Anthropic’s ongoing
10 infringement, and (3) any additional remedies the law provides.

11 12. Plaintiffs elect not to bring this case as a class action because the Copyright Act
12 entitles them to recover individualized statutory damages, determined by a jury, for Anthropic’s
13 infringement and related conduct. Plaintiffs desire to retain full control of their case and avoid
14 having their rights diluted by being swept into sprawling class-action settlements structured to
15 resolve claims for pennies on the dollar. Recent history has shown that certain class actions and
16 proposed settlement(s) seem to serve the tech conglomerate-infringers, not creators and publishers.
17 LLM companies should not be able to so easily extinguish thousands upon thousands of high-value
18 claims at bargain-basement rates, eliding what should be the true cost of their massive willful
19 infringement.

20 13. That is not how Plaintiffs plan to proceed. Under established Supreme Court
21 precedent, “the amount of statutory damages is a question for the jury.”³ The Copyright Act thus
22 vests authors with the right to have a jury evaluate the willfulness of infringement and assign a
23 damages amount tailored to Anthropic’s conduct.

24
25
26 ² Ben Bergman, *Anthropic has surged to a trillion-dollar valuation on secondary markets,*
27 *overtaking OpenAI*, Business Insider (Apr. 22, 2026, 5:29 PM), <https://perma.cc/Q5GH-HSKZ>.

28 ³ *Feltner v. Columbia Pictures Television, Inc.*, 523 U.S. 340, 353 (1998).

14. In sum, the Copyright Act’s statutory-damages and attorneys’ fee regime empowers individual authors and publishers to hold infringers accountable without the need for class action treatment. That is what Plaintiffs have chosen to do.

II. PARTIES

A. Plaintiffs

15. Plaintiffs are legal or beneficial owners of works that were included on the Works List compiled for the *Bartz v. Anthropic* Settlement.⁴

16. Plaintiffs have opted out of the *Bartz v. Anthropic* Settlement to pursue this action, each in their individual capacity.

17. A non-exhaustive list of registered copyrights owned by Plaintiffs is included as Exhibit A (herein, the “Plaintiffs’ Works”).⁵

18. Plaintiffs have never authorized Anthropic to reproduce, distribute, or otherwise use the Plaintiffs’ Works.

19. Anthropic illegally downloaded, reproduced, and distributed Plaintiffs’ Works as demonstrated by their inclusion on the Works List. Of course, as alleged below, Anthropic downloaded, reproduced, and distributed *millions* of works beyond what was included on the Works List in *Bartz*, and the full scope of Anthropic’s copyright infringement, including with respect to Plaintiffs, will be revealed in discovery. According to publicly available metadata—and as illustrated by their inclusion in the Works List—the Plaintiffs’ Works are contained in pirated online libraries and datasets including Books3 (and thus *The Pile*), Library Genesis or “LibGen,” Z-Library, Pirate Library Mirror (“PiLiMi”), and Anna’s Archive. As alleged below, Anthropic has directly or indirectly downloaded works contained in (at least) Books3, LibGen, and PiLiMi (and

⁴ In *Bartz v. Anthropic PBC*, No. 3:24-cv-05417 (N.D. Cal. June 23, 2025), the Class “includes legal or beneficial owners of the exclusive right to reproduce copies of a work that was included in the versions of the LibGen or PiLiMi datasets downloaded by Anthropic and is listed in the searchable database called the Works List.” *Anthropic Copyright Settlement Website, FAQ 5, What Is the Class?*, (last visited May 11, 2026), <https://perma.cc/Y62K-CJ5L>.

⁵ Even where other individuals or entities are listed as copyright claimants or authors on the relevant copyright registrations, Plaintiffs are owners of the copyrights listed in Exhibit A.

1 thus Z-Library) and there is accordingly a reasonable inference that Anthropic illegally downloaded
2 and reproduced the Plaintiffs' Works.

3 20. Plaintiff R. O. Kwon is the legal or beneficial owner of one or more of the Plaintiffs'
4 Works identified on Exhibit A.

5 21. Plaintiff Alexander Dreihann-Holenia is the legal or beneficial owner of one or more
6 of the Plaintiffs' Works identified on Exhibit A.

7 22. Plaintiff Stephanie Griest is the legal or beneficial owner of one or more of the
8 Plaintiffs' Works identified on Exhibit A.

9 23. Plaintiff Elizabeth Hinton is the legal or beneficial owner of one or more of the
10 Plaintiffs' Works identified on Exhibit A.

11 24. Plaintiff Brian Jackson is the legal or beneficial owner of one or more of the
12 Plaintiffs' Works identified on Exhibit A.

13 25. Plaintiff Jason Alan Jankovsky is the legal or beneficial owner of one or more of the
14 Plaintiffs' Works identified on Exhibit A.

15 26. Plaintiff Branden W. Joseph is the legal or beneficial owner of one or more of the
16 Plaintiffs' Works identified on Exhibit A.

17 27. Plaintiff Kimberly Kay Hoang is the legal or beneficial owner of one or more of the
18 Plaintiffs' Works identified on Exhibit A.

19 28. Plaintiff Glenn Kreisberg is the legal or beneficial owner of one or more of the
20 Plaintiffs' Works identified on Exhibit A.

21 29. Plaintiff Richard Krevolin is the legal or beneficial owner of one or more of the
22 Plaintiffs' Works identified on Exhibit A.

23 30. Plaintiff Gerald Lamb is the legal or beneficial owner of one or more of the Plaintiffs'
24 Works identified on Exhibit A.

25 31. Plaintiff Chang-rae Lee is the legal or beneficial owner of one or more of the
26 Plaintiffs' Works identified on Exhibit A.

27 32. Plaintiff Jeannie Levinson is the legal or beneficial owner of one or more of the
28 Plaintiffs' Works identified on Exhibit A.

1 33. Plaintiff Steven Lowenthal is the legal or beneficial owner of one or more of the
2 Plaintiffs' Works identified on Exhibit A.

3 34. Plaintiff Mackenzie Blue LLC, owned and controlled by author Tina Wells, is the
4 legal or beneficial owner of one or more of the Plaintiffs' Works identified on Exhibit A.

5 35. Plaintiff Brian Malec is the legal or beneficial owner of one or more of the Plaintiffs'
6 Works identified on Exhibit A.

7 36. Plaintiff Donna Matney is the legal or beneficial owner of one or more of the
8 Plaintiffs' Works identified on Exhibit A.

9 37. Plaintiff Ilona McCarty is the legal or beneficial owner of one or more of the
10 Plaintiffs' Works identified on Exhibit A.

11 38. Plaintiff Tiffany McDaniel is the legal or beneficial owner of one or more of the
12 Plaintiffs' Works identified on Exhibit A.

13 39. Plaintiff Brian Merchant is the legal or beneficial owner of one or more of the
14 Plaintiffs' Works identified on Exhibit A.

15 40. Plaintiff Eric L. Muller is the legal or beneficial owner of one or more of the
16 Plaintiffs' Works identified on Exhibit A.

17 41. Plaintiff Sonia Nazario is the legal or beneficial owner of one or more of the
18 Plaintiffs' Works identified on Exhibit A.

19 42. Plaintiff Daniel José Older is the legal or beneficial owner of one or more of the
20 Plaintiffs' Works identified on Exhibit A.

21 43. Plaintiff Laurence Ralph is the legal or beneficial owner of one or more of the
22 Plaintiffs' Works identified on Exhibit A.

23 44. Plaintiff Lilliam Rivera is the legal or beneficial owner of one or more of the
24 Plaintiffs' Works identified on Exhibit A.

25 45. Plaintiff Seth A. Roberts is the legal or beneficial owner of one or more of the
26 Plaintiffs' Works identified on Exhibit A.

27 46. Plaintiff Aida Salazar is the legal or beneficial owner of one or more of the Plaintiffs'
28 Works identified on Exhibit A.

1 47. Plaintiff Erika L. Sanchez is the legal or beneficial owner of one or more of the
2 Plaintiffs' Works identified on Exhibit A.

3 48. Plaintiff Jeremy N. Smith is the legal or beneficial owner of one or more of the
4 Plaintiffs' Works identified on Exhibit A.

5 49. Plaintiff Soul Writers, LLC, owned and controlled by author Tony Gaskins Jr., is the
6 legal or beneficial owner of one or more of the Plaintiffs' Works identified on Exhibit A.

7 50. Plaintiff Ramesh Srinivasan is the legal or beneficial owner of one or more of the
8 Plaintiffs' Works identified on Exhibit A.

9 51. Plaintiff Robert Tannenbaum is the legal or beneficial owner of one or more of the
10 Plaintiffs' Works identified on Exhibit A.

11 52. Plaintiff Kenneth Warner is the legal or beneficial owner of one or more of the
12 Plaintiffs' Works identified on Exhibit A.

13 53. Plaintiff Robert Zimmerman is the legal or beneficial owner of one or more of the
14 Plaintiffs' Works identified on Exhibit A.

15 **B. Defendant**

16 54. Defendant Anthropic PBC is a Delaware public benefit corporation with its principal
17 place of business in San Francisco, California. Anthropic develops and commercializes large
18 language models (including the Claude series), which were trained using datasets sourced in part
19 from shadow libraries and other datasets containing pirated books.

20 **III. JURISDICTION AND VENUE**

21 55. This action arises under the Copyright Act of 1976, 17 U.S.C. § 101 *et seq.* This
22 Court has subject-matter jurisdiction under 28 U.S.C. §§ 1331 and 1338(a) because Plaintiffs assert
23 claims exclusively under federal copyright law.

24 56. This Court has personal jurisdiction over the Defendant. The Defendant has
25 purposefully availed itself of the privilege of conducting business in this District and the State of
26 California. The Defendant committed acts of copyright infringement in this District, directed
27 conduct toward this District, or knowingly caused harm that was suffered in this District. The
28 Defendant maintains substantial, continuous, and systematic contacts with this District.

57. Venue is proper in this District under 28 U.S.C. § 1400(a) because the Defendant or its agents resides or may be found in this District as a result of the infringing acts alleged herein. Venue is also proper under 28 U.S.C. § 1391(b)(2) because a substantial part of the events giving rise to Plaintiffs’ claims—including the acquisition of pirated copies of the Plaintiffs’ Works, the reproduction and ingestion of those copies into Defendant’s training pipelines, the training and fine-tuning of the relevant LLMs, and the commercialization of the resulting models—occurred in this District.

IV. FACTUAL ALLEGATIONS

A. LLMs and the Generative AI “Arms Race”

58. “Generative artificial intelligence” or “generative AI” refers to systems and models that create outputs—such as text or images—that simulate human expression, often in response to user prompts.

59. Large language models are a form of generative AI, which are designed to process—or “understand”—and generate natural language. At a high level, LLMs operate by studying and “learning” the statistical relationship between words and other text (such as punctuation marks); upon further refinement by their developers, the models then process complex mathematical sequences to “predict” sequences of text based on the statistical relationships they have learned and been trained to recognize.

60. Development of LLMs requires, among other things, “training” the model on data—*i.e.*, the model’s inputs. Training “requires identifying a formal measure or ‘objective’ for how well the model performs, and then repeatedly adjusting the model’s parameters based on that objective as the model is exposed to training data.”⁶ While not all LLM developers use the same terminology, developers commonly reference a “pre-training” process—“in which a massive amount of computing power and data is spent to teach the model the broad foundations of language, grammar, and reasoning”—and a “post-training” or “fine-tuning” process “where the pre-trained model is

⁶ U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training Pre-Publication*, at 17 (May 2025), <https://perma.cc/EY5U-EFUY>.

1 further trained on a (relative to pre-training) smaller amount of carefully curated data of specific
2 tasks.”⁷ For purposes of this Complaint, “training” refers to all stages of the LLM training process.

3 61. LLMs typically are trained, at least initially, by feeding them massive amounts of
4 text data from which they can “learn” statistical relationships between and among text. In the process
5 of compiling training data and using it during the training process, LLM developers as a matter of
6 course typically create multiple copies of “raw” datasets. So, for instance, if an LLM developer
7 downloads a copy of a digital book from some Internet-based source (*e.g.*, a shadow library), the
8 developer typically will create a new “cleaned” copy of the book (*e.g.*, stripped of certain
9 undesirable data), deduplicate that book against other potential copies, compile that book with other
10 data to create a new compilation (or compilations), and store that book in multiple locations. In
11 short, with respect to training data, the LLM development pipeline typically involves numerous
12 reproductions of any given piece of the dataset.

13 62. A model’s inputs—the training datasets—are directly tied to the model’s
14 performance. As one AI researcher has now famously said, a “model[’s] behavior is not determined
15 by architecture, hyperparameters, or optimizer choices. It’s determined by your dataset, nothing
16 else.”⁸

17 63. In determining the corpus of training data for an LLM, a model’s developer typically
18 considers “the quantity of data, its quality, and the ultimate purpose(s) of the model.”⁹

19 64. In terms of data quantity, the LLMs that have been developed and are being
20 developed by leading technology companies are trained on enormous datasets, typically on the scale
21 of terabytes. This is because AI researchers have found that “increasing the quantity of training data
22 typically increases a model’s ‘performance.’”¹⁰

24 ⁷ *Id.*

25 ⁸ James Betker, *The “it” in AI Models is the Dataset*, (June 10, 2023), <https://perma.cc/ZCH9-H53S>.

26 ⁹ U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training Pre-*
27 *Publication*, at 9.

28 ¹⁰ *Id.* at 10.

1 65. In terms of data quality, “[r]ecent research from major developers suggests that
2 quality may even be a more important consideration than quantity.”¹¹ “Garbage in, garbage out,”
3 the saying goes. AI researchers quickly discovered that published content—books in particular—
4 makes for some of the best training material for LLMs. Unlike Internet content and many other
5 forms of text, published material typically is structured, long-form, and highly polished. Published
6 materials provide formal, extended prose that teaches models narrative structure, complex syntax,
7 and coherent storytelling.

8 66. That books make for particularly valuable training material for LLMs is evidenced
9 by the fact that several—if not most—of the major LLM developers either have paid for or expressly
10 contemplated paying (large sums) for licenses to use copyrighted works to train their LLMs. Indeed,
11 in an *amicus* brief filed in *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-3417 (N.D. Cal.), the
12 Association of American Publishers cited *dozens* of “AI licensing deals for textual works,” including
13 deals involving Amazon, Microsoft, OpenAI, and Perplexity.¹² And, as the Court found in that
14 litigation, Meta considered spending up to \$100 million to license copyrighted materials in
15 connection with LLM training.¹³

16 67. As alleged below, Anthropic’s employees reached the same conclusion as other AI
17 researchers—that published materials, and books in particular, were an essential ingredient for
18 developing a high-performing LLM.

19 68. As discussed in the introduction, technology companies have treated generative AI
20 as the next foundational layer of the digital economy. Industry leaders publicly describe an “AI arms
21 race,” in which they have redirected their corporate strategies to seize control of what they believe
22 will become a new infrastructure layer for commerce, communication, and knowledge work.¹⁴ For
23

24 ¹¹ *Id.* at 11.

25 ¹² ECF 535 at 12, Case No. 3:23-cv-3417-VC (N.D. Cal.) (Apr. 11, 2025).

26 ¹³ *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1040 (N.D. Cal. 2025).

27 ¹⁴ See Dr. Peter Asaro, *What is an ‘Artificial Intelligence Arms Race’ Anyway?*, 15 I/S: J.L. & Pol’y
28 for Info. Soc’y 45 (2019).

1 these companies, staying ahead of competitors is “code red.”¹⁵ Among other things, being too slow
 2 out of the blocks could mean ending up in last place; if by the time you publicly released *your* LLM
 3 model, the public writ large and private industry had already adopted *other companies’* models,
 4 there would be a real risk that your model would be left on the shelf.

5 69. Like other participants in this race, Anthropic risked falling behind and constantly
 6 needed to move fast to develop and roll out updates of Claude, its capstone LLM. And to do that,
 7 Anthropic needed to get its hands on high-quality training data.

8 **B. Shadow Libraries, *The Pile*, and Torrenting**

9 70. For years now, there have been illicit, online marketplaces for digital copies of books.
 10 These “shadow libraries,” as they are commonly known, systematically acquire, store, index, and
 11 disseminate full-fidelity digital copies of copyrighted books—typically in native formats such as
 12 EPUB, PDF, MOBI, or DJVU. These online repositories maintain searchable catalogs, metadata,
 13 mirrors, bulk-download mechanisms, and—critically—complete downloadable archives designed
 14 to allow third parties to replicate the entire collection.¹⁶ And they do all of this without authorization
 15 from authors or publishers.¹⁷ In short, shadow libraries facilitate the reproduction and distribution
 16 of unauthorized copies of copyrighted works at industrial scale.

17 71. These libraries are widely known within both piracy communities and the technology
 18 sector as illegal sources of copyrighted books. Many have been the subject of criminal prosecutions,
 19 civil injunctions, domain seizures, and formal designation as “notorious piracy markets” by United
 20

21 ¹⁵ See Sharon Goldman, *Sam Altman declares ‘Code Red’ as Google’s Gemini surges—three years*
 22 *after ChatGPT cause Google CEO Sundar Pichai to do the same*, FORTUNE (Dec. 2, 2025, 11:43
 AM), <https://perma.cc/J9MS-UQD2>.

23 ¹⁶ See Letter from Mary E. Rasenberger, CEO, Authors Guild, & Umair Kazi, Dir. of Pol’y &
 24 Advocacy, Authors Guild, to Daniel Lee, Assistant U.S. Trade Representative for Innovation &
 25 Intellectual Prop., Office of the U.S. Trade Representative (Oct. 7, 2022), <https://perma.cc/XM4R-NDN3>.

26 ¹⁷ See Riddhi Setty, *Rampant ‘Shadow Libraries’ Drive Calls for Anti-Piracy Action*, BLOOMBERG
 27 LAW (Oct. 19, 2022, 9:03 AM), <https://perma.cc/F5VH-3BA6>; Woodcock, *‘Shadow Libraries’ Are*
 28 *Moving Their Pirated Books to the Dark Web After Fed Crackdown*, VICE (Nov. 30, 2022, 11:38
 AM) <https://perma.cc/K9FA-VLPW>.

1 States trade authorities. As centralized shadow libraries increasingly faced enforcement actions,
 2 including the seizure of domains, third parties responded by creating full mirrored copies of those
 3 repositories for decentralized redistribution.¹⁸

4 72. In 2018, an OpenAI employee downloaded pirated copies of books from Library
 5 Genesis, or “LibGen”—a shadow library repeatedly enjoined by federal courts¹⁹—and used those
 6 books to create two internal datasets OpenAI called “LibGen1” and “LibGen2,” which OpenAI
 7 publicly referred to as “Books1” and “Books2.”²⁰ OpenAI used those pirated corpora to train GPT-
 8 3, which it released in June 2020 to widespread commercial acclaim. OpenAI’s use of a books
 9 corpora from a pirate library established the model for how to quickly and cheaply obtain published
 10 materials for use as LLM training data.

11 73. Within weeks of GPT-3’s release, an open research collective, EleutherAI, formed
 12 with the express goal of replicating GPT-3’s capabilities and democratizing access to large-scale
 13 language modeling. To do so, EleutherAI assembled and publicly released *The Pile*, a large (800+
 14 gigabyte) general-purpose training dataset expressly designed to be downloaded, used, and
 15 incorporated into LLMs by academic researchers, startups, and commercial AI developers.²¹

16 74. Because OpenAI did not disclose the precise makeup of its training datasets,
 17 members of EleutherAI constructed a pirated book corpus of their own: “Books3,” consisting of
 18
 19
 20

21
 22 ¹⁸ Woodcock, ‘Shadow Libraries’ Are Moving Their Pirated Books to the Dark Web After Fed
 Crackdowns.

23
 24 ¹⁹ See *Cengage Learning, Inc. et al. v. Does 1-50 d/b/a Library Genesis*, ECF No. 36, Case No. 23-
 cv-8136 (S.D.N.Y. Sept. 24, 2024).

25 ²⁰ *In re OpenAI, Inc., Copyright Infringement Litig.*, Case No. 1:2025-md-03143 (S.D.N.Y. 2025),
 26 ECF No. 846 at 9; Ashley Belanger, *OpenAI desperate to avoid explaining why it deleted pirated
 book datasets*, ARSTECHNICA (Dec. 1, 2025), <https://perma.cc/9M7K-8DA9>.

27 ²¹ Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, arXiv, 1
 28 (2020), available at <https://perma.cc/NHV6-R8YE>.

1 approximately 196,640 books.²² Books3 comprises about 12 percent (just over 100 gigabytes) of
 2 *The Pile*.²³

3 75. EleutherAI and the compiler of Books3, Shawn Presser, have confirmed the genesis
 4 of Books3 is the shadow library Bibliotik. Presser has publicly stated that Books3 represents “all of
 5 bibliotik,”²⁴ and an EleutherAI paper likewise confirms that Books3 “is a dataset of books derived
 6 from a copy of the contents of the Bibliotik private tracker.”²⁵

7 76. Bibliotik is a private, invitation-only torrent tracker that has long functioned as a
 8 centralized source of pirated e-books. The repository, which hosts and distributes hundreds of
 9 thousands of copyrighted books, is only accessible with the use of torrenting protocols, which are
 10 explained below.²⁶

11 77. Bibliotik has been widely recognized in piracy communities, academic literature, and
 12 AI-research documentation as a shadow library devoted to copyrighted books. Unsurprisingly, then,
 13 *The Pile*’s datasheet acknowledges that “Books3 is almost entirely comprised of copyrighted
 14 works.”²⁷ Bibliotik’s illicit nature is no secret—it has been openly discussed for years prior to major
 15 tech companies’ use of datasets derived directly from it.²⁸

16 78. In an interview with *The Atlantic*, Presser confirmed that the illuminating purpose
 17 behind Books3 was to ensure broad-based access to the tools necessary to create LLMs:
 18

19 ²² Stella Biderman et al., *Datasheet for the Pile*, arXiv, 8 (2020), <https://perma.cc/7KL2-LTLF>.

20 ²³ Biderman et al., *Datasheet for the Pile*, arXiv, 8.

21 ²⁴ Shawn Presser, X (Oct. 25, 2020, 1:32 AM), <https://perma.cc/7WRD-NHRX>.

22 ²⁵ Biderman et al., *Datasheet for the Pile*, arXiv, 8.

23 ²⁶ See Ruheni Mathenge, *The 12 Best Private Torrent Sites Still Working in 2026*, PRIVACYSAVVY
 24 (last accessed March 9, 2026), <https://perma.cc/4V8M-3ALY>.

25 ²⁷ Biderman et al., *Datasheet for the Pile*, arXiv, 8.

26 ²⁸ Kyle Barr, *Anti-Piracy Group Takes Massive AI Training Dataset ‘Books3’ Offline*, GIZMODO
 27 (Aug. 18, 2023, 8:50 AM), <https://perma.cc/5ZL9-RQCQ>; Peter Schoppert, *Whether you’re an*
 28 *undergraduate doing research, or a fan of the Nick Stone novels, or indeed a hungry AI*, SUBSTACK
 (Nov. 29, 2022), <https://perma.cc/8YD9-M4BD>.

1 [Presser] created Books3 in the hope that it would allow any developer to create
 2 generative-AI tools. “It would be better if it wasn’t necessary to have something like
 3 Books3,” he said. “But the alternative is that, without Books3, only OpenAI can do
 4 what they’re doing.”²⁹

5 79. EleutherAI’s compilation and distribution of *The Pile* and Books3 provided “off-the-
 6 shelf” access to a corpus of infringing works that any AI developer could download and immediately
 7 incorporate into a large-scale training pipeline. As a result, Books3’s presence within *The Pile*
 8 facilitated wide downstream distribution and adoption of a corpus derived from a pirate book
 9 library.³⁰

10 80. Bibliotik is just one of many Internet-based shadow libraries and, as alleged below,
 11 Anthropic relied on more than just *The Pile* in connection with its development of Claude.

12 81. As noted above, OpenAI’s developers trained its LLMs on a books corpus derived
 13 from LibGen, one of the largest and longest-running shadow libraries in the world. LibGen hosts
 14 millions of pirated books, academic texts, and scholarly articles.³¹ It operates as a centralized
 15 repository offering direct downloads of full-fidelity e-book files. It also distributes its entire database
 16 through bulk archives and torrent files, enabling third parties to download and locally host complete
 17 copies of its collection.³² LibGen has been repeatedly enjoined by federal courts for copyright
 18 infringement and has been designated a “notorious market” by the United States Trade
 19

20
 21 ²⁹ *Id.*

22 ³⁰ See Stella Biderman, *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*,
 23 ELEUTHERAI (Dec. 31, 2020), <https://perma.cc/JGT9-LLTK>.

24 ³¹ Office of the U.S. Trade Representative, REVIEW OF NOTORIOUS MARKETS FOR
 25 COUNTERFEITING AND PIRACY, 27 (2024), <https://perma.cc/22VN-9VD7>. (“Libgen ... hosts a
 26 large number of digital copies of books, manuals, journals, and other works, many of which are
 unauthorized copies of copyright protected content.”)

27 ³² See Letter from Mary E. Rasenberger, CEO, Authors Guild, & Umair Kazi, Dir. of Pol’y &
 28 Advocacy, Authors Guild, to Daniel Lee, Assistant U.S. Trade Representative for Innovation &
 Intellectual Prop., Office of the U.S. Trade Representative at n.5.

1 Representative.³³ Despite enforcement actions, LibGen has remained accessible through shifting
 2 domains, mirrors, and downloadable archives. Its persistence is a function of deliberate
 3 decentralization designed to evade shutdown.³⁴

4 82. Z-Library (also known as “B-ok”) is another well-known shadow library that
 5 emerged as an expanded and user-friendly derivative of LibGen. It incorporated large portions of
 6 LibGen’s catalog while adding additional titles, metadata, and interface features.³⁵ Z-Library
 7 offered premium features—including faster downloads and higher volume limits—in exchange for
 8 payment, operating in effect as a commercial piracy service.³⁶ In 2022, Z-Library’s domains were
 9 seized by law-enforcement authorities, and its operators were arrested and later indicted for criminal
 10 copyright infringement.³⁷ These actions confirmed what had long been publicly known: Z-Library
 11 was an illegal piracy operation. The seizure of Z-Library did not eliminate access to its content.
 12 Instead, third parties responded by creating full mirrors of its collection to ensure continued
 13 distribution.³⁸

14 83. Another major player in the shadow library ecosystem is PiLiMi, which is a complete
 15 mirrored archive of the Z-Library corpus, explicitly created to preserve and propagate Z-Library’s
 16
 17
 18

19 ³³ See Office of the U.S. Trade Representative, 2019 REVIEW OF NOTORIOUS MARKETS FOR
 20 COUNTERFEITING AND PIRACY, 27, <https://perma.cc/22VN-9VD7>.

21 ³⁴ See Andrew Albanese, *Textbook Publishers Sue Notorious ‘Shadow Library’ Libgen*,
 22 PUBLISHERS WEEKLY (Sep. 14, 2023), <https://perma.cc/3NPY-UJCX>.

23 ³⁵ Jordana Rosenfeld, *Z-Library*, ENCYCLOPEDIA BRITANNICA (last accessed March 9, 2026),
<https://perma.cc/4H26-GDCC>.

24 ³⁶ See Masood Farivar, *Two Russian Nationals Charged With Operating E-Book Piracy Site*, VOA
 25 (Nov. 16, 2022), <https://perma.cc/6MPD-QNKB>.

26 ³⁷ Press Release, U.S. Dep’t of Justice, U.S. Att’y’s Off., E. Dist. of N.Y., *Two Russian Nationals*
 27 *Charged with Running Massive E-Book Piracy Website* (Nov. 16, 2022), <https://perma.cc/CR4L-JLA3>.

28 ³⁸ See Woodcock, *‘Shadow Libraries’ Are Moving Their Pirated Books to The Dark Web*.

1 pirated collection after law-enforcement seizures, ensuring continuity of access despite shutdowns
2 of the original site.³⁹

3 84. PiLiMi is not merely a website or index. It is a full, downloadable dataset designed
4 to allow users to obtain and locally host millions of pirated books through torrent distribution.⁴⁰
5 Users who download PiLiMi do not passively receive data; they actively participate in copying and
6 redistributing copyrighted works through torrent “leeching” and “seeding.”⁴¹

7 85. PiLiMi later rebranded as “Anna’s Archive” and expanded to aggregate and host the
8 complete collections of LibGen, Z-Library, and other pirated sources.⁴² Like PiLiMi, Anna’s
9 Archive functions as a meta-library: it indexes, mirrors, and redistributes multiple shadow libraries
10 simultaneously, offering users unified access to millions of pirated books.⁴³

11 86. Anna’s Archive offers paid tiers that provide “high-speed” or priority access to its
12 pirated collections.⁴⁴ Through its downloadable archives and torrent-based distribution, Anna’s
13 Archive enables users to acquire and store local copies of millions of copyrighted books in bulk.⁴⁵
14
15

16 ³⁹ See Ernesto Van de Sar, “*Anna’s Archive*” *Opens the Door to Z-Library and Other Pirate*
17 *Libraries*, TORRENTFREAK (Nov. 19, 2022), <https://perma.cc/U88R-WTR4>.

18 ⁴⁰ See Geoff Wheelright, *Will I get a piece of Anthropic’s \$1.5B settlement if my book was used to*
19 *train AI?*, GEEKWIRE (Sep. 18, 2025), <https://perma.cc/X5EW-TKY3>.

20 ⁴¹ See Robert Nogacki, *Anthropic’s Landmark Settlement: A \$1.5 Billion Copyright Precedent in*
21 *Artificial Intelligence Training Data*, LinkedIn (Sep. 7, 2025), <https://perma.cc/J3BD-P5JT>
(describing “leeching” and “seeding” as “processes characteristic of peer-to-peer networks where
22 users simultaneously download and distribute files”).

23 ⁴² See Ernesto Van de Sar, “*Anna’s Archive*” *Opens the Door to Z-Library and Other Pirate*
Libraries.

24 ⁴³ *Id.*

25 ⁴⁴ See *If you’re an LLM, please read this*, ANNA’S ARCHIVE (February 18, 2026),
26 <https://perma.cc/989X-GS3Q>.

27 ⁴⁵ See M. Luisa Simpson, *2024 Special 301 Out-of-Cycle Review of Notorious Markets: Request*
28 *for Comments*, ASSOCIATION OF AMERICAN PUBLISHERS (OCTOBER 2, 2024),
<https://perma.cc/P2E9-MYBY>.

1 87. Although individual domain names may change, Anna’s Archive and its underlying
2 datasets remain accessible through mirrors, torrents, and distributed storage systems.

3 88. According to Anna’s Archive, “virtually all major companies building LLMs
4 contacted us to train on our data. . . . We have given high-speed access to about 30 companies.”⁴⁶
5 Anna’s Archive blog stated as recently as February 18, 2026 that if an *LLM was reading its blog*
6 “you have likely been trained in part on our data.”⁴⁷

7 89. Many of these shadow libraries and datasets can be downloaded using “torrent,” a
8 file-sharing method used in peer-to-peer networks. Torrenting works by breaking a file into many
9 small pieces and distributing those pieces across a network of participating computers. A user who
10 torrents a shadow-library repository does not receive a single copy from a single source; rather, the
11 user downloads portions of the library from numerous other computers that already possess the
12 copyrighted books. Torrent software then reassembles those pieces into a complete library on the
13 user’s machine. Torrenting protocols, including BitTorrent, are often configured by default to
14 *reupload* pieces of the copyrighted files to others on the network both during download (“leeching”)
15 and after download is complete (“seeding”). This means that each participant in the torrent both
16 copies and redistributes the copyrighted works. By obtaining copyrighted materials through this
17 leech-and-seed process, a user may make multiple unauthorized reproductions of, and engage in
18 numerous distributions of, the copyrighted materials.

19 C. Anthropic’s Deliberate Infringement of Plaintiffs’ Copyrights

20 1. Anthropic’s Business and Bypassing of the Licensing Market

21 90. As alleged, Anthropic is the developer of the Claude family of LLMs. Its Claude
22 models are projected to generate billions of dollars in annual revenue and have buoyed Anthropic’s
23
24

25 _____
26 ⁴⁶ See *Copyright reform is necessary for national security*, ANNA’S ARCHIVE (Jan. 31, 2025),
<https://perma.cc/3RVZ-6G5B>.

27 ⁴⁷ See *If you’re an LLM, please read this*, ANNA’S ARCHIVE (Feb. 18, 2026), [https://perma.cc/MJ7Z-](https://perma.cc/MJ7Z-3ZCL)
28 [3ZCL](https://perma.cc/MJ7Z-3ZCL).

1 overall valuation.⁴⁸ Anthropic’s business model is built on the large-scale copying of published
 2 content. *First*, Anthropic developed and commercialized the Claude family of large language
 3 models by stealing up to seven million copyrighted books, including the Plaintiffs’ Works.⁴⁹ *Second*,
 4 Anthropic also had the explicit goal to “amass a central library of ‘all the books in the world’ to
 5 retain ‘forever.’”⁵⁰

6 91. To accomplish its goal, “it stole the works for its central library by downloading them
 7 from pirated libraries.”⁵¹ Anthropic did explore the licensing market but quickly decided that it was
 8 not “a practical approach”⁵² to satisfy its thirst for copyrighted works. Judge Alsup, presiding over
 9 the *Bartz v. Anthropic* litigation, stated simply: “From the start, Anthropic ‘ha[d] many places from
 10 which’ it could have purchased books, but it preferred to steal them to avoid ‘legal/practice/business
 11 slog.’”⁵³ In short, rather than paying for copyrighted materials, Anthropic downloaded pirated
 12 copies of protected works, reproduced them, fed them into its models, and otherwise retained them
 13 for its own “central library.”

14 2. Anthropic’s Acquisition of Pirated Works from Shadow Libraries

15 92. To train Claude, Anthropic downloaded Books3 in 2021, which co-founder Ben
 16 Mann “knew had been assembled from unauthorized copies of copyrighted books,” downloaded at
 17 least five million copies of books from LibGen, “which [Mann] knew had been pirated,” and
 18

19
 20 ⁴⁸ See Rashi Shrivastava, *Anthropic Is Cashing In On Claude Code’s Success*, Forbes (February 17,
 21 2026 5:19 PM), <https://www.forbes.com/sites/the-prompt/2026/02/17/anthropic-is-cashing-in-on-claude-codes-success/> (on file with author) (“Claude Code hit \$2.5 billion in run rate revenue and
 22 doubled since the start of the year...and [Anthropic’s] total annualized revenue has climbed to \$14 billion.”)

23 ⁴⁹ See *Bartz v. Anthropic PBC*, 791 F. Supp. 3d 1038, 1046 (N.D. Cal. 2025).

24 ⁵⁰ *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1014 (N.D. Cal. 2025).

25 ⁵¹ *Id.* at 1029.

26 ⁵² See *Bartz v. Anthropic PBC*, No. 3:24-cv-05417 (N.D. Cal. June 23, 2025), ECF No. 553, at 17-
 27 18.

28 ⁵³ *Bartz v. Anthropic PBC*, 787 F. Supp. 3d at 1015.

1 downloaded at least two million copies of books from PiLiMi, which “Anthropic knew had been
2 pirated.”⁵⁴

3 93. Anthropic’s decision to base its flagship models on pirated books was driven by
4 commercial advantage. As Jared Kaplan, Anthropic’s co-founder and Chief Science Officer, has
5 explained, “it is important to obtain vast amounts of books and also to have diverse types of books
6 in the training corpus to create a model with truly generative capabilities.”⁵⁵ As long-form content,
7 training LLMs on the “entire text” of published material—as Anthropic has admitted to doing—
8 offers great value.⁵⁶

9 94. At the time Anthropic downloaded Books3, LibGen, and PiLiMi, it knew or should
10 have known that they were repositories of unauthorized copies of copyrighted works.

11 3. Anthropic’s Torrenting and Distribution of Pirated Works

12 95. Anthropic did not merely obtain pirated books passively. It used BitTorrent to
13 acquire *and distribute* massive collections of infringing works.

14 96. In June 2021, Mann personally torrented approximately five million pirated books
15 from LibGen for Anthropic’s use.⁵⁷ Mann acted with the knowledge and approval of Anthropic’s
16 senior leadership. Before the LibGen torrenting, he discussed the plan with co-founders Dario
17 Amodei, Jared Kaplan, and other senior leaders.

18 97. Anthropic’s own Archive Team had described LibGen as a “blatant violation of
19 copyright,” and Amodei himself called it “sketchy.”⁵⁸ Yet Anthropic proceeded, choosing torrenting
20 over purchasing or licensing the copyrighted materials.

21 98. Anthropic repeated the same conduct in 2022 with PiLiMi. As U.S. law enforcement
22 was working to shut down existing pirate libraries, a group online copied LibGen and built upon it

23 ⁵⁴ *Id.*

24 ⁵⁵ See Kaplan Decl. ¶ 47, *Bartz*, ECF No. 128.

25 ⁵⁶ *Id.* at ¶¶ 43, 47.

26 ⁵⁷ B. Mann Dep. Tr. at 89:6-8, (Aug. 15 and 18, 2025), *Bartz*, ECF No. 337-1.

27 ⁵⁸ *Id.* at 144:4-13, 396:3-13.

1 to create Z-Library. The FBI later shut down Z-Library as well.⁵⁹ However, by that point, Z-Library
2 had itself been fully copied, or “mirrored,” into another repository: PiLiMi.⁶⁰ Mann circulated the
3 PiLiMi source to colleagues, and Anthropic employees torrented approximately two million
4 additional pirated books not already captured from LibGen.⁶¹ Those files were not abstract “data”
5 but full-text digital books in .epub, .pdf, and .txt formats,⁶² including the Plaintiffs’ Works.

6 99. In torrenting from these files, Anthropic knew or should have known that it was
7 participating in a *peer-to-peer network that traded in unauthorized copies of copyrighted material*.
8 Stated differently, Anthropic knew or should have known that it was facilitating further copyright
9 infringement by making available to others on the peer-to-peer network unauthorized copies of
10 copyrighted materials.

11 4. Anthropic’s Stripping of Copyright Management Information

12 100. Anthropic not only reproduced and distributed the Plaintiffs’ Works without
13 authorization, it also deliberately stripped those materials of copyright management information
14 (“CMI”) before using them to train Claude.

15 101. Anthropic knew that major AI-training datasets, including *The Pile*, WebText,
16 WebText2, and Common Crawl, contained copyrighted works whose copyright notices, ownership
17 information, and other identifying material had been removed through extraction tools such as
18 Newspaper, Dragnet, Readability, and jusText. Anthropic’s founders and senior employees were
19 familiar with those tools before and after Anthropic’s founding. Dario Amodei, Benjamin Mann,
20 Jared Kaplan, and other future Anthropic personnel had used or developed datasets at OpenAI that
21 extracted text from scraped webpages while omitting surrounding material, including footers where
22
23

24 ⁵⁹ See *Bartz*, 791 F. Supp. 3d at 1046.

25 ⁶⁰ *Id.*

26 ⁶¹ *Id.*

27 ⁶² *Id.*

1 copyright notices typically appear.⁶³ Yet Anthropic trained Claude on these same stripped datasets,
 2 despite knowing that they included unauthorized copies of copyrighted works stripped of their CMI.

3 102. Anthropic also affirmatively chose tools that removed CMI more effectively,
 4 including from Plaintiffs' copyrighted materials. In 2021, Mann, Kaplan, and other Anthropic
 5 employees compared extraction tools for filtering training data and then used these extraction
 6 methods to clean its training data, including the Plaintiffs' Works. The purpose and effect of this
 7 process was to train Claude on Plaintiffs' expressive content while suppressing the ownership
 8 information attached to it. Anthropic wanted Claude to reproduce protected expression, not the
 9 copyright notices that identify the rights holders. By stripping that information from training data
 10 and outputs, Anthropic concealed the source of its infringement from users, Plaintiffs, and other
 11 copyright owners.

12 **5. Anthropic's Unauthorized Scanning of Copyrighted Works**

13 103. Since 2024, Anthropic has purchased physical books at scale, often in batches of tens
 14 of thousands, and scanned them into digital files for AI training.⁶⁴ To date, Anthropic has spent
 15 millions of dollars buying and scanning millions of physical books.⁶⁵ This scanning project was not
 16 authorized by Plaintiffs or other copyright owners. As with the millions of pirated books Anthropic
 17 torrented from LibGen and PiLiMi, Anthropic converted these works into training material without
 18 permission, payment, or license.

19 **6. Anthropic's "Everything Forever" Library of Pirated Works**

20 104. Defendant did not merely make temporary copies of copyrighted works for a discrete
 21 training process (not that that would be permissible, in any event). Instead, from all those
 22 infringements, Anthropic created a general "research library" or "generalized data area."⁶⁶ That is,

23 _____
 24 ⁶³ Tom B. Brown et al., *Language Models Are Few-Shot Learners*, 8–9, arXiv (July 22, 2020),
 25 <https://perma.cc/9WAX-F9HG>; Jared Kaplan et al., *Scaling Laws for Neural Language Models*, 7,
 arXiv (Jan. 23, 2020), <https://perma.cc/VJU8-59FH>.

26 ⁶⁴ *See Bartz*, 791 F. Supp. 3d at 1047.

27 ⁶⁵ *Bartz*, ECF No. 553, at 17-18.

28 ⁶⁶ *Bartz*, 787 F. Supp. 3d at 1016.

1 Anthropic created a permanent central library of pirated published materials, including the
 2 Plaintiffs’ Works, to exploit for a range of unspecified *other* purposes, such as “research.”⁶⁷
 3 Anthropic continued this conduct, for which it “lacked any entitlement to hold copies of the books
 4 at all” and “retain[ed] them even after deciding it would not make further copies from them for
 5 training.”⁶⁸ The plan was clear: “ke[eping] in the original version of the underlying book files
 6 Anthropic had obtained or created, that is, pirated or scanned” and “stor[ing] everything forever.”⁶⁹
 7 Anthropic saw “no compelling reason to delete a book—even if not used for training LLMs.”⁷⁰

8 105. Whether or not a particular work was ultimately selected for training, Anthropic’s
 9 initial copying and continued retention of the work as part of its “forever” pirated library constituted
 10 an unauthorized reproduction. That infringement is independent of, and not excused by, any later
 11 use Anthropic may claim to have made in connection with AI training.

12 7. Anthropic’s Embedding of Near-Verbatim Copies in Claude

13 106. Anthropic’s copying did not end when it acquired the Plaintiffs’ Works. Scientific
 14 research confirms that training large language models on copyrighted books can embed persistent,
 15 near-verbatim copies of those works inside the models’ internal parameters, allowing the models to
 16 reproduce substantial portions of the original text when prompted.

17 107. A 2025 study by researchers from Cornell, Stanford, and West Virginia University
 18 tested leading LLM models, including Anthropic’s Claude 3.7 Sonnet, and extracted memorized
 19 copyrighted books from them.⁷¹ Using repeated prompting and continuation techniques, the
 20 researchers were able to recover lengthy, near-verbatim passages. The results for Claude were
 21 severe. From Claude 3.7 Sonnet alone, researchers extracted 97.5% of *The Great Gatsby*, 95.5% of

22
 23 ⁶⁷ *Id.*

24 ⁶⁸ *Id.* at 1031.

25 ⁶⁹ *Id.* at 1016 (internal quotation marks omitted).

26 ⁷⁰ *Id.* (internal quotation marks omitted).

27 ⁷¹ A. Feder Cooper et al., *Extracting Copyrighted Long-Form Text from Production Language*
 28 *Models*, arXiv, 1-2 (2025), <https://perma.cc/F9YQ-NKV7>.

1 1984, 94.3% of *Frankenstein*, 92.3% of *Harry Potter and the Sorcerer's Stone*, and 70.2% of *The*
 2 *Hobbit*.⁷² These results show that Claude does not merely learn abstract patterns from copyrighted
 3 materials; it stores recoverable copies of the works at extraordinary scale.

4 108. A separate 2026 study reached the same conclusion through a different method.
 5 Researchers from Stony Brook University, Carnegie Mellon University, and Columbia Law School
 6 found that finetuned frontier models could reproduce up to 85–90% of copyrighted books, with
 7 single verbatim spans exceeding 460 words.⁷³ The models reproduced these passages from semantic
 8 prompts alone, without being given the book text in the prompt.

9 109. The 2026 study also traced the likely source of the memorized content. The authors
 10 searched the extracted verbatim passages against more than eight trillion tokens of public web data
 11 and found that many of the longest passages did not appear in those web collections. Yet 80 of the
 12 81 tested books appeared in Books3 or LibGen. That finding confirms that the memorized text came
 13 not from ordinary web exposure, but from datasets and pirate libraries of the kind Anthropic used
 14 to train Claude.

15 110. Together, these studies also show that safety filters do not remove the underlying
 16 copyrighted material from the model. The works remain embedded in model weights and can be
 17 extracted despite protective layers.⁷⁴ In practical effect, Claude functions as a repository of pirated
 18 copyrighted works: Anthropic copied published materials to train the model, and the model retained
 19 near-verbatim reproductions of them.

20 8. Anthropic's Harm to the Market for Plaintiffs' Copyrighted Materials

21 111. Anthropic's conduct has a detrimental effect on the potential market for and value of
 22 Plaintiffs' works, including by, among other things, developing products that create and are capable
 23 of creating content which serves as a direct substitute for the Plaintiffs' Works, developing products
 24

25 ⁷² *Id.* at 12, Fig. 5.

26 ⁷³ A. Liu et al., *Alignment Whack-a-Mole : Finetuning Activates Verbatim Recall of Copyrighted*
 27 *Books in Large Language Models*, arXiv, 2, (2026), <https://perma.cc/EC6X-Z38M>.

28 ⁷⁴ *Id.* at 9–11.

1 that create content and are capable of creating content which serves as indirect substitutes for the
 2 Plaintiffs' Works, and undermining Plaintiffs' ability to participate in and profit from the market for
 3 licensing their works for the purpose of training LLMs.

4 112. *First*, Anthropic's decision to download and use unauthorized copies of the
 5 Plaintiffs' Works from shadow libraries deprived Plaintiffs of revenue in the form of licensing fees
 6 that they would have otherwise earned. As alleged herein, there is an existing market for licensing
 7 copyrighted materials such as Plaintiffs', including for use in the development of LLMs. Anthropic
 8 bypassed that market, and in doing so, deprived Plaintiffs of licensing revenue they would have
 9 earned. Anthropic now allows users to opt out of having their data used to train its AI models, but
 10 it deprived Plaintiffs and many others of that choice. Anthropic's misconduct undermined Plaintiffs
 11 and many others, eliminating the bargaining power they should have had, and otherwise would have
 12 had, with respect to licensing terms for the use Anthropic made of their works.

13 113. *Second*, Claude, trained on the Plaintiffs' Works (and the protected works of others),
 14 is capable of generating outputs that compete directly with, and risk serving as replacements for, the
 15 Plaintiffs' Works. As the U.S. Copyright Office has warned, "the speed and scale at which AI
 16 systems generate content pose a serious risk of diluting markets for works of the same kind as in
 17 their training data."⁷⁵

18 114. *Third*, even if Claude models are restricted from outputting extended portions of
 19 verbatim text from copyrighted works, they are nevertheless capable of producing nearly
 20 indistinguishable "versions" of copyrighted works such that a consumer would use the AI-generated
 21 version of the material rather than pay for a copy of the actual copyrighted work.

22 **V. CLAIMS FOR RELIEF**

23 **COUNT I**

24 **Direct Copyright Infringement (17 U.S.C. § 501)**

25 115. Plaintiffs incorporate the allegations above.

27 ⁷⁵ U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training Pre-*
 28 *Publication* at 65.

1 116. Plaintiffs are the legal or beneficial owners of the copyrighted works listed in Exhibit
2 A (referred to herein as the Plaintiffs' Works).

3 117. The Defendant, without authorization from Plaintiffs, copied, downloaded,
4 reproduced, ingested, parsed, embedded, and used pirated copies of the Plaintiffs' Works in the
5 development, training, fine-tuning, and deployment of its commercial large language models. These
6 acts violated Plaintiffs' exclusive rights under 17 U.S.C. § 106.

7 118. Defendant's infringement occurred repeatedly throughout the lifecycle of its AI-
8 model development. As alleged above, Defendant:

- 9 • acquired through torrenting and direct downloading the Plaintiffs' Works from
10 shadow-library repositories and datasets containing pirated works from shadow
11 libraries;
- 12 • distributed the Plaintiffs' Works through the use of torrenting software, programs, or
13 protocols;
- 14 • reproduced additional copies during ingestion, preprocessing, storage, deduplication,
15 formatting, and/or tokenization; and
- 16 • while training its models made even more copies of the text—because every training
17 pass (each epoch and each step of gradient descent) automatically requires creating
18 and working with fresh versions of that text.

19 119. Defendant's reproductions and distributions of Plaintiffs' copyrighted works were
20 made without permission, license, or consent and violated Plaintiffs' exclusive rights under the
21 Copyright Act.

22 120. Defendant's infringement was **willful**. As alleged above, Defendant knowingly
23 trained its models on and/or optimized its product with datasets saturated with pirated books,
24 including the Plaintiffs' Works; relied on shadow-library corpora it knew to be illegal; ignored
25 internal and external warnings; attempted to conceal the composition of its training datasets; and
26 continued copying after public reports, lawsuits, law-enforcement seizures, cease-and-desist
27 notices, and industry-wide alerts made the illegality unmistakable.
28

PRAYER FOR RELIEF

WHEREFORE, Plaintiffs request that the Court enter judgment on their behalf by ordering:

- a. Judgment in favor of Plaintiffs against the Defendant;
- b. A declaration that the Defendant has infringed Plaintiffs' exclusive copyrights under the Copyright Act;
- c. A declaration that such infringement is willful;
- d. A declaration that Defendant violated 17 U.S.C. § 1202(b) through its removal of copyright management information;
- e. A permanent injunction enjoining the Defendant and all those acting in concert with it from engaging in the infringing conduct alleged herein;
- f. That the Defendant be directed to account to Plaintiffs for all gains, profits, and advantages derived from its unlawful acts;
- g. An award of statutory damages under the Copyright Act;
- h. An award of statutory or actual damages under 17 U.S.C. § 1203(c);
- i. An award of restitution, disgorgement, costs, expenses, and attorneys' fees as permitted by law (including those allowable under 17 U.S.C. § 505 and/or 17 U.S.C. § 1203(b)(4)–(5));
- j. Pre- and post-judgment interest on the damages awarded to Plaintiffs; and
- k. Further relief for Plaintiffs as the Court may deem just and proper.

JURY TRIAL DEMANDED

Under Federal Rule of Civil Procedure 38(b), Plaintiffs demand a trial by jury.

1 Dated: May 15, 2026

Respectfully submitted,

2 /s/ Elizabeth Brannen

3 Elizabeth Brannen (SBN 226234)

4 John Stokes (SBN 310847)

5 Lauren Martin (SBN 294367)

STRIS & MAHER LLP

17785 Center Court Dr N, Ste 600

6 Cerritos, CA 90703

7 T: (213) 995-6800

8 F: (213) 261-0299

ebrannen@stris.com

jstokes@stris.com

9 lmartin@stris.com

10 Kyle Roche (*pro hac vice* forthcoming)

11 Devin (Velvel) Freedman (*pro hac vice*
forthcoming)

12 Alex Potter (*pro hac vice* forthcoming)

**FREEDMAN NORMAND FRIEDLAND
LLP**

13 155 E. 44th Street, Ste 915

14 New York, NY 10017

15 T: (646) 494-2900

kroche@fnf.law

16 vel@fnf.law

17 apotter@fnf.law

Counsel for Plaintiffs