

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

Jacqueline Sahlberg (*pro hac vice*
forthcoming)
jsahlberg@stris.com
STRIS & MAHER LLP
1717 K St NW Suite 900
Washington, DC 20006
T: (202) 800-5749

Counsel for Plaintiffs

Bridget Asay (*pro hac vice* forthcoming)
basay@stris.com
STRIS & MAHER LLP
15 East State Street, Suite 2
Montpelier, VT 05602
T: (802) 858-4285

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

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA

JOHN CARREYROU, LISA
BARRETTA, PHILIP SHISHKIN,
JANE ADAMS, MATTHEW
SACKS, and MICHAEL KOCHIN

Plaintiffs,

v.

ANTHROPIC PBC; GOOGLE LLC;
OPENAI, INC.; OPENAI OPCO LLC;
OPENAI GP LLC; OPENAI GLOBAL
LLC; OAI CORPORATION LLC;
OPENAI HOLDINGS LLC;
META PLATFORMS, INC.;
XAI CORPORATION; and
PERPLEXITY AI, INC.,

Defendants.

Civil Case No.:

COMPLAINT

DEMAND FOR JURY TRIAL

1 Plaintiffs John Carreyrou, Lisa Barretta, Philip Shishkin, Jane Adams,
2 Matthew Sacks, and Michael Kochin (collectively “Plaintiffs”), each proceeding in
3 their individual capacity, bring this action against Anthropic PBC (“Anthropic”),
4 Google LLC (“Google”), OpenAI, Inc. and its affiliated entities (“OpenAI”), Meta
5 Platforms, Inc. (“Meta”), xAI Corporation (“xAI”), and Perplexity AI, Inc.
6 (“Perplexity”) (collectively, “Defendants”), and allege as follows:

7 **I. INTRODUCTION**

8 1. This case concerns a straightforward and deliberate act of theft that
9 constitutes copyright infringement. Anthropic, Google, OpenAI, Meta, xAI, and
10 Perplexity, illegally copied vast quantities of copyrighted books without permission
11 and then used those stolen copies to build and train their commercial large language
12 models (“LLMs”) and/or optimize their product. Defendants helped themselves to
13 the copyrighted works of hundreds of authors—including bestselling writers,
14 Pulitzer Prize-winning journalists, and creators of widely read nonfiction and
15 fiction.

16 2. Rather than obtain licenses or pay for the use of these works, each
17 Defendant downloaded pirated copies of Plaintiffs’ books from shadow-library
18 websites such as LibGen, Z-Library, and OceanofPDF and then reproduced, parsed,
19 analyzed, re-copied, used, and embedded those works into their LLMs (and/or used
20 those works to optimize their product) to accelerate commercial development and
21 win the generative-AI race. The Copyright Act prohibits exactly this conduct.

22 3. Defendants targeted Plaintiffs’ works because they were of exceptional
23 value as training data. Defendants have acknowledged—internally and publicly—
24 that long-form, high-quality books are the “gold-standard” training material for
25 LLMs. Books teach models how narrative flows, how human expression is
26 structured, how syntax and rhythm operate, and how ideas are communicated
27 through creative choices. Instead of paying for that value, Defendants pilfered
28

1 illegal copies and used those copies to build systems now worth many hundreds of
2 billions of dollars.

3 4. The infringement here occurred at least twice for every work.

4 5. *First*, Defendants obtained Plaintiffs’ copyrighted books from illegal
5 shadow libraries.

6 6. *Second*, Defendants made additional unlicensed copies of the
7 unlawfully obtained books, including during ingestion, preprocessing, and model
8 training and/or retrieval-augmented generation. LLM training necessarily involves
9 making multiple copies of each work.

10 7. Defendants’ misconduct was willful. The libraries Defendants accessed
11 had, for years, been the subject of criminal prosecutions, civil lawsuits, and
12 widespread warnings within the technology industry. Defendants were repeatedly
13 told that using such datasets was unlawful, and employees across the industry raised
14 red flags about using them, including some calling them “illegal pirated websites”
15 and warning of liability for accessing them. But Defendants pressed forward
16 because copying pirated books allowed them to more cheaply build more-
17 sophisticated models faster and with higher performance. These choices gave
18 Defendants a competitive advantage—an advantage built on continuous and
19 unlawful reproduction of pirated works.

20 8. Anthropic’s Claude models were trained on datasets containing
21 hundreds of thousands of books obtained from piracy sources that included
22 Plaintiffs’ copyrighted books. Google’s Gemini and Imagen models rely on
23 datasets—including Z-Library and OceanofPDF—that incorporate large collections
24 of pirated works. OpenAI and Microsoft’s GPT-series models were trained on vast
25 pirated corpora, including LibGen, enabling products such as ChatGPT, Copilot,
26 GitHub Copilot, and a suite of AI-enhanced Microsoft applications. Meta’s Llama
27 models were trained on massive sets of books downloaded from shadow libraries,
28 including LibGen. xAI’s Grok models and Perplexity’s AI search systems likewise

1 relied on large-scale ingestion of pirated books. These models, all trained and/or
2 optimized on Plaintiffs' copyrighted books, now anchor multibillion-dollar product
3 ecosystems.

4 9. Defendants' unauthorized copying of Plaintiffs' books has inflicted
5 immediate and ongoing harm. Plaintiffs spent years creating the works at issue;
6 Defendants spent seconds copying them. By embedding Plaintiffs' creative
7 expression into their model parameters and/or optimization, Defendants have
8 appropriated—and continue to monetize—the fruits of Plaintiffs' copyrighted labor
9 across cloud platforms, consumer products, enterprise tools, advertising systems,
10 and subscription services.

11 10. While Defendants' conduct constitutes classic copyright infringement,
12 their conduct is unique in that they have willfully infringed Plaintiffs' copyrights at
13 an unprecedented scale for massive commercial gain.

14 11. To redress Defendants' repeated, unlawful, and massive infringement
15 of their work, each Plaintiff individually seeks (1) damages, (2) permanent
16 injunctive relief barring Defendants' ongoing infringement, and (3) any additional
17 remedies the law provides.

18 12. Plaintiffs bring this action to hold Defendants accountable for the
19 infringement that enabled their rise in the generative-AI marketplace, and to enforce
20 the fundamental principle that creative expression cannot be taken, copied, or
21 exploited without permission or compensation.

22 13. Plaintiffs elect not to bring this case as a class action because the
23 Copyright Act entitles them to recover individualized statutory damages, determined
24 by a jury, for each Defendant's infringement of their work. Plaintiffs desire to retain
25 full control of their case and avoid having their rights diluted by being swept into
26 sprawling class-action settlements structured to resolve claims for pennies on the
27 dollar.

14. The danger is not hypothetical. In the class action against Anthropic pending in the Northern District of California, the court has recently preliminarily approved a settlement framework where each work will only receive approximately \$3,000 less attorneys' fees and costs—a tiny fraction (just 2%) of the Copyright Act's statutory ceiling of \$150,000 in addition to attorneys' fees per willfully infringed work.

15. These pending class actions and proposed settlement(s) seem to serve Defendants, not creators. LLM companies should not be able to so easily extinguish thousands upon thousands of high-value claims at bargain-basement rates, eliding what should be the true cost of their massive willful infringement.

16. That is not how Plaintiffs plan to proceed. Under established Supreme Court precedent, "the amount of statutory damages is a question for the jury."¹ The Copyright Act thus vests authors with the right to have a jury evaluate the willfulness of infringement and assign a damages amount tailored to the Defendant's conduct.

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

II. PARTIES

A. Plaintiffs

18. Plaintiff John Carreyrou is an author and journalist who resides in New York. He is the author of *Bad Blood: Secrets and Lies in a Silicon Valley Startup*. His work is contained in pirated online libraries such as the LibGen and Z-Library shadow libraries. Defendants have directly or indirectly downloaded books illegally contained in LibGen and Z-Library, and there is accordingly a reasonable inference that Defendants illegally downloaded Carreyrou's work.

¹ *Feltner v. Columbia Pictures Television, Inc.*, 523 U.S. 340, 353 (1998).

1 19. Plaintiff Lisa Barretta is an author who resides in Pennsylvania. She is
2 the author of *The Street-Smart Psychic's Guide to Getting a Good Reading*. Her
3 work is contained in the LibGen and Z-Library shadow libraries. Defendants have
4 directly or indirectly downloaded books illegally contained in LibGen and Z-
5 Library, and there is accordingly a reasonable inference that Defendants illegally
6 downloaded Barretta's work.

7 20. Plaintiff Philip Shishkin is an author and journalist who resides in
8 Washington D.C. He is the author of *Restless Valley: Revolution, Murder and*
9 *Intrigue in the Heart of Central Asia*. His work is contained in the LibGen and Z-
10 Library shadow libraries. Defendants have directly or indirectly downloaded books
11 illegally contained in LibGen and Z-Library, and there is accordingly a reasonable
12 inference that Defendants illegally downloaded Shishkin's work.

13 21. Plaintiff Jane Adams is an author and journalist who resides in
14 Washington. She is the author of *Boundary Issues: Using Boundary Intelligence to*
15 *Get the Intimacy You Want and the Independence You Need in Life, Love, and*
16 *Work and How to Sell What You Write*. Her works are contained in the LibGen and
17 Z-Library shadow libraries. Defendants have directly or indirectly downloaded
18 books illegally contained in LibGen and Z-Library, and there is accordingly a
19 reasonable inference that Defendants illegally downloaded Adams's works.

20 22. Plaintiff Matthew Sacks is an author and journalist who resides in
21 California. He is the author of *Pro Website Development and Operations*. His work
22 is contained in the LibGen and Z-Library shadow libraries. Defendants have directly
23 or indirectly downloaded books illegally contained in LibGen and Z-Library, and
24 there is accordingly a reasonable inference that Defendants illegally downloaded
25 Sacks's work.

26 23. Plaintiff Michael Kochin is an author and journalist who resides in
27 Israel. He is the author of *Five Chapters on Rhetoric: Character, Action, Things,*
28 *Nothing & Art*. His work is contained in the LibGen and Z-Library shadow libraries.

1 Defendants have directly or indirectly downloaded books illegally contained in
2 LibGen and Z-Library, and there is accordingly a reasonable inference that
3 Defendants illegally downloaded Kochin’s work.

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

6 **B. Defendants**

7 25. Defendant Anthropic PBC (“Anthropic”) is a Delaware public benefit
8 corporation with its principal place of business in San Francisco, California.
9 Anthropic develops and commercializes large language models (including the
10 Claude series). Anthropic directed, authorized, and profited from the acts of
11 copyright infringement alleged in this Complaint, including the acquisition of
12 pirated copies of Plaintiffs’ copyrighted books from shadow-library websites and the
13 reproduction, ingestion, and use of those works in the training, development, and
14 deployment of its LLMs. Anthropic conducts substantial business in this District and
15 throughout the United States.

16 26. Defendant Google LLC (“Google”) is a Delaware limited liability
17 company with its principal place of business in Mountain View, California. Google
18 develops, trains, and commercializes generative AI models, including Gemini, Bard
19 (formerly), and Imagen, which were trained using datasets containing large volumes
20 of pirated books. Google copied, reproduced, and embedded Plaintiffs’ copyrighted
21 works into its models without permission or license and continues to profit from
22 those infringements across its commercial product ecosystem, including Google
23 Cloud, Google Search, and various AI-powered enterprise tools.

24 27. Defendant OpenAI, Inc., and its affiliated entities OpenAI OpCo LLC,
25 OpenAI GP LLC, OpenAI Global LLC, OAI Corporation LLC, OpenAI Holdings
26 LLC, OpenAI Startup Fund I LP, OpenAI Startup Fund GP I LLC, and OpenAI
27 Startup Fund Management LLC (collectively, “OpenAI”) are entities organized
28 under the laws of Delaware with principal places of business in San Francisco,

1 California. OpenAI develops and commercializes the GPT family of models
2 (including GPT-3, GPT-3.5, GPT-4, GPT-4o, and their derivatives), which were
3 trained on datasets containing illegal copies of Plaintiffs’ copyrighted books.
4 OpenAI reproduced Plaintiffs’ works multiple times during data collection,
5 preprocessing, and training, and continues to exploit those works commercially
6 through ChatGPT, ChatGPT Enterprise, the OpenAI API, and other products.

7 28. Defendant Meta Platforms, Inc. (“Meta”) is a Delaware corporation
8 with its principal place of business in Menlo Park, California. Meta develops and
9 distributes the Llama series of LLMs, including Llama-1, Llama-2, and Llama-3,
10 which were trained using datasets sourced in part from shadow libraries such as
11 LibGen containing pirated books. Meta also acts as a distributor of such datasets
12 within its research ecosystem. Meta copied Plaintiffs’ copyrighted works without
13 license and monetizes those infringements through its integration of Llama models
14 into Facebook, Instagram, WhatsApp, Ray-Ban Meta Glasses, enterprise APIs, and
15 other products.

16 29. Defendant xAI Corporation (“xAI”) is a Nevada corporation with its
17 principal place of business in Palo Alto, California. xAI develops the Grok series of
18 LLMs, which were trained on large-scale text corpora that include illegally obtained
19 books and datasets containing Plaintiffs’ copyrighted works. xAI copied,
20 reproduced, and embedded Plaintiffs’ works into its models for use in Grok and its
21 associated commercial services, including products offered through X Corp.
22 (formerly Twitter).

23 30. Defendant Perplexity AI, Inc. (“Perplexity”) is a Delaware corporation
24 with its principal place of business in San Francisco, California. Perplexity builds
25 and deploys AI search and text-generation systems that rely on the unauthorized use
26 of copyrighted works to optimize its product through its retrieval-augmented
27 generation or “RAG” process. On information and belief, Perplexity’s RAG process
28 relies on pirated copies of Plaintiffs’ books. On information and belief, Perplexity

1 reproduced and exploited Plaintiffs’ copyrighted works without authorization in its
2 AI search systems.

3 **III. JURISDICTION AND VENUE**

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

7 32. This Court has personal jurisdiction over each Defendant. Each
8 Defendant has purposefully availed itself of the privilege of conducting business in
9 this District and the State of California. Each Defendant committed acts of copyright
10 infringement in this District, directed conduct toward this District, or knowingly
11 caused harm that was suffered in this District. Each Defendant maintains substantial,
12 continuous, and systematic contacts with this District.

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

21 **IV. FACTUAL ALLEGATIONS**

22 **A. The Generative AI Arms Race.**

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

26 35. Over the last several years, technology companies have treated
27 generative AI as the next foundational layer of the digital economy. Industry leaders
28 publicly describe an “AI arms race,” in which they have redirected their corporate

1 strategies to seize control of what they believe will become a new infrastructure
2 layer for commerce, communication, and knowledge work.²

3 36. For these companies, staying ahead of competitors is “code red.”³
4 Google itself responded by consolidating its AI research divisions, devoting
5 unprecedented resources to generative AI, and rapidly integrating AI features across
6 its product suite.⁴

7 37. OpenAI, for its part, built a sequence of large language models—
8 beginning with GPT-1 and GPT-2 and continuing through GPT-3, GPT-3.5, GPT-4,
9 GPT-4o, and their derivatives—that power products such as ChatGPT, the OpenAI
10 API, and Microsoft’s GPT-based offerings including Bing Chat and Copilot. These
11 models sit at the center of an enterprise now valued in the hundreds of billions of
12 dollars.

13 38. Google’s Gemini family of models and its Imagen text-to-image
14 systems have likewise been woven into core Google products—including Search,
15 Cloud, Workspace, and other AI-powered products—which Google attributes with
16 driving billions of dollars in new revenue and record quarterly results.⁵

17
18 ² See Dr. Peter Asaro, *What is an ‘Artificial Intelligence Arms Race’ Anyway?*, 15 I/S: J.L. & Pol’y
for Info. Soc’y 45 (2019).

19 ³ See Sharon Goldman, *Sam Altman declares ‘Code Red’ as Google’s Gemini surges—three years*
20 *after ChatGPT cause Google CEO Sundar Pichai to do the same*, FORTUNE (Dec. 2, 2025, 11:43
21 AM), <https://fortune.com/2025/12/02/sam-altman-declares-code-red-google-gemini-ceo-sundar-pichai/>.

22 ⁴ See, e.g., Sundar Pichai, *Building for our AI future*, Google (Apr. 18, 2024),
23 <https://blog.google/inside-google/company-announcements/building-ai-future-april-2024/>; Tom
24 Jowitt, *Google Consolidates DeepMind and AI Research Teams*, SILICON (Apr. 19, 2024, 9:35
25 PM), [https://www.silicon.co.uk/e-innovation/artificial-intelligence/google-consolidates-deepmind-
and-ai-research-teams-559660#:~:text=Alphabet's%20Google%20division%20is%20once,in%202014%20for%20\\$500m](https://www.silicon.co.uk/e-innovation/artificial-intelligence/google-consolidates-deepmind-and-ai-research-teams-559660#:~:text=Alphabet's%20Google%20division%20is%20once,in%202014%20for%20$500m)
26 (discussing consolidation).

27 ⁵ See Kyle Wiggers and Maxwell Zeff, *Google Gemini: Everything you need to know about the*
28 *generative AI apps and models*, TECHCRUNCH (Feb. 26, 2025, 6:09 PM),
<https://techcrunch.com/2025/02/26/what-is-google-gemini-ai/> (“The Gemini apps aren’t the only

1 39. Anthropic has taken the same path. Its Claude models—trained to
2 write, summarize, and analyze text at book-length scale—are projected to generate
3 hundreds of millions of dollars in annual revenue and have supported valuations in
4 the hundreds of billions of dollars, funded by major technology investors such as
5 Amazon and Google.

6 40. Meta, which had fallen behind in the AI race, repositioned itself by
7 pouring billions of dollars into its “Llama” series of large language models. Meta
8 has integrated Llama into its core products, including Facebook, Instagram, and
9 WhatsApp, and views its generative-AI investments as central to its future
10 competitive advantage.

11 41. In this race, access to high-quality training data is a decisive
12 competitive weapon. For large language models in particular, companies have
13 repeatedly acknowledged that “books are actually more important than web data”:
14 they provide formal, extended prose that teaches models narrative structure,
15 complex syntax, and coherent storytelling.⁶

16 42. The Defendants did not obtain that gold-standard material lawfully.
17 Instead, in order to win the generative-AI arms race cheaply and quickly, each
18 Defendant turned to the same piracy repositories—shadow-library websites like
19 LibGen, Z-Library, Bibliotik, Books3, and similar datasets—and copied Plaintiffs’
20 books without permission, without licenses, and without compensation.

21
22
23 means of recruiting Gemini models’ assistance with tasks. Slowly but surely, Gemini-imbued
24 features are making their way into staple Google apps and services like Gmail and Google
25 Docs.”); Jennifer Elias, *Google Cloud chief details how search giant is making billions*
26 *monetizing its AI products*, CNBC (Sep. 9, 2025 3:58 PM),
<https://www.cnbc.com/2025/09/09/google-cloud-chief-details-how-tech-company-is-monetizing-ai.html> (quoting Google Cloud CEO Thomas Kurian: “We’ve made billions using AI already.” (cleaned up)).

27 ⁶ See Alex Reisner, *The Unbelievable Scale of AI’s Pirated-Books Problem*, THE ATLANTIC (Mar.
28 20, 2025), <https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/>.

43. Defendants did not rely on a single source of illicit books, but instead assembled their training corpora through multiple, distinct pirated datasets, each differing in origin, structure, and method of distribution. Among those used by Defendants were (a) *Books3*, a curated dataset of approximately 200,000 pirated books derived from the Bibliotik shadow library and distributed as extracted text paired with filenames; (b) *Library Genesis* (“*LibGen*”), a centralized shadow library hosting millions of full-fidelity ebook files in native formats such as .epub and .pdf; and (c) *Z-Library*, an expanded and refined derivative of LibGen that incorporated overlapping content as well as additional titles, metadata, and organizational features. Each dataset constituted a separate repository of copyrighted works, and Defendants’ acquisition of each involved independent acts of unauthorized reproduction.⁷

44. As centralized shadow libraries increasingly faced enforcement actions, including the seizure of Z-Library’s domains, third parties responded by creating full mirrored copies of those repositories for decentralized redistribution. One such mirror—known as the *Pirate Library Mirror* or “PiLiMi”—consists of replicas of the Z-Library corpus (itself derived in substantial part from LibGen), designed to ensure continued access to pirated books even after the original sites were shut down. PiLiMi is not merely a website or index, but a complete, downloadable dataset intentionally created to perpetuate mass infringement through peer-to-peer copying.

45. At least some Defendants knowingly treated PiLiMi as a distinct and supplemental pirated dataset rather than a redundant copy of materials already obtained. Before downloading PiLiMi, some Defendants compared its catalog

⁷ See *AI Watchdog: Books3*, THE ATLANTIC (Sep. 10, 2025), <https://www.theatlantic.com/technology/archive/2025/09/dataset-books3/683662/>; Claire Woodcock, ‘Shadow Libraries’ Are Moving Their Pirated Books to the Dark Web After Fed Crackdowns, VICE (Nov. 30, 2022, 11:38 AM), <https://www.vice.com/en/article/shadow-libraries-are-moving-their-pirated-books-to-the-dark-web-after-fed-crackdowns/>.

1 against their existing LibGen holdings, identified which titles were not already in
2 their possession, and deliberately downloaded only those additional works. Through
3 this process, some Defendants expanded their illicit libraries by millions of unique
4 copyrighted books obtained after the shutdown of Z-Library, while retaining earlier
5 pirated copies from Books3 and LibGen in centralized storage. These actions reflect
6 intentional sourcing, selection, and accumulation of multiple pirated book datasets at
7 different times, through different mechanisms, and in conscious disregard of
8 copyright law.

9 46. This ecosystem exists for one purpose: making copyrighted works
10 available for unauthorized download. Defendants exploited these shadow libraries
11 and datasets—along with others not named here—to train their LLM models on
12 Plaintiffs’ books without permission or compensation.

13 47. Many of these shadow libraries and datasets can be downloaded using
14 “torrent,” a file-sharing method. Torrenting works by breaking a file into thousands
15 of small pieces and distributing those pieces across a network of participating
16 computers. A user who torrents a shadow-library repository does not receive a
17 single copy from a single source; rather, the user downloads portions of the library
18 from numerous other computers that already possess the copyrighted books. Torrent
19 software then reassembles those pieces into a complete library on the user’s
20 machine. Certain torrenting protocols are configured by default to reupload pieces of
21 the copyrighted files to others on the network both during download (“leeching”) and
22 after download is complete (“seeding”). This means that each participant in the
23 torrent both copies and redistributes the copyrighted works without permission. By
24 obtaining Plaintiffs’ books through this leech-and-seed process, a user may make
25 multiple unauthorized reproductions of Plaintiffs’ works.

26 48. Defendants’ unlawful conduct did not end with the unauthorized
27 downloads of Plaintiffs’ works. In addition to making unauthorized copies when
28 torrenting shadow libraries, Defendants reproduced Plaintiffs’ copyrighted books

1 without permission numerous—potentially countless—other times, including in
 2 preprocessing and deduplicating the data and in iteratively training and fine-tuning
 3 their LLMs. Defendants’ businesses and products would not exist in their current
 4 forms without these repeated violations of the Copyright Act.

5 **B. Anthropic Trained Its LLM Models On Copyrighted Works That**
 6 **Were Pirated.**

7 49. Anthropic’s business model is built on the large-scale copying of
 8 books. Anthropic has developed and commercialized the “Claude” family of large
 9 language models by stealing millions of copyrighted books, including Plaintiffs’
 10 works. Rather than pay for the creative expression it exploits, Anthropic
 11 downloaded pirated copies of books, reproduced them, and fed them into its models.

12 50. Anthropic’s own public statements and technical papers confirm that
 13 books are central to Claude’s capabilities. Anthropic has described a training corpus
 14 “most of which we sourced from The Pile,”⁸ an 800-gigabyte dataset assembled for
 15 large-language-model training that includes a books subset known as “Books3.”

16 51. The Pile’s architects have explained that Books3 is composed of books
 17 scraped from Bibliotik, a private torrent tracker long identified in piracy
 18 communities as a source of illegal ebooks.⁹

19 52. Anthropic has admitted that it used The Pile (which includes Books3)
 20 to train its Claude models and that roughly one-third of one core Claude training
 21 dataset consisted of “internet books.”¹⁰ By downloading these datasets and ingesting
 22 them into Claude, Anthropic necessarily made multiple unlicensed copies of
 23 Plaintiffs’ works: once when obtaining them from pirate sources, again during

24 _____
 25 ⁸ See Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment*, arXiv,
 26 27 (2021), <https://arxiv.org/pdf/2112.00861>.

26 ⁹ Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, arXiv, 3
 27 (2020), <https://arxiv.org/pdf/2101.00027>.

28 ¹⁰ See Askell et al., *A General Language Assistant as a Laboratory for Alignment* at 27.

1 preprocessing and storage, and repeatedly during training and fine-tuning. As the
 2 U.S. Patent and Trademark Office has explained, LLM training “almost by
 3 definition involve[s] the reproduction of entire works or substantial portions” of
 4 them.¹¹

5 53. Anthropic selected books precisely because they are especially valuable
 6 training material. Anthropic touts Claude’s ability to process entire books (up to
 7 roughly 75,000 words) and generate coherent long-form responses that reflect not
 8 only word ordering and syntax, but also themes, narrative structure, and high-level
 9 ideas—capabilities that could be developed only by training on a large corpus of
 10 long-form prose.¹²

11 **C. Anthropic’s Infringement Was Willful.**

12 54. Anthropic’s infringement was not inadvertent. It knowingly relied on
 13 datasets that the industry and its own researchers understood to be saturated with
 14 pirated books.

15 55. The Pile’s own documentation states that the Books3 subset was
 16 created from a copy of Bibliotik,¹³ a “shadow library” whose existence and illicit
 17 nature had been publicly discussed for years in piracy forums, GitHub repositories,
 18 and arXiv papers. The EleutherAI paper on The Pile explains that Bibliotik was
 19 included because books are “invaluable” for long-range context modeling and
 20
 21

22 ¹¹ U.S. Patent & Trademark Office, *Public Views on Artificial Intelligence and Intellectual*
 23 *Property Policy* 24 (2020), [https://www.uspto.gov/sites/default/files/documents/USPTO_AI-](https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf)
 24 [Report_2020-10-07.pdf](https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf).

25 ¹² Anthropic, *Introducing 100K Context Windows*, [https://www.anthropic.com/news/100k-context-](https://www.anthropic.com/news/100k-context-windows)
 26 [windows](https://www.anthropic.com/news/100k-context-windows) (last visited Dec. 22, 2025) (“We’ve expanded Claude’s context window from 9K to
 27 100K tokens, corresponding to around 75,000 words!”); Anthropic, *Claude 2*,
<https://www.anthropic.com/news/claude-2> (last visited Dec. 22, 2025) (“Claude can work over
 hundreds of pages of technical documentation or even a book.”).

28 ¹³ Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* at 3.

1 “coherent storytelling”—precisely the qualities that make Plaintiffs’ works
2 valuable.¹⁴

3 56. Public commentary and enforcement actions have long identified
4 Bibliotik, LibGen, Z-Library, and similar sites as notorious hubs of copyright
5 infringement. These sites have been targeted in criminal cases, civil suits by
6 publishers, and “notorious markets” reports by United States trade authorities.¹⁵

7 57. Despite this, Anthropic chose to source its training data from The Pile
8 and Books3, and then attempted to conceal the precise composition of its training
9 corpus. Anthropic has endeavored to keep its training data secret even as outside
10 researchers and Anthropic’s own prior work revealed heavy reliance on The Pile and
11 internet-book datasets.

12 58. Anthropic’s decision to base its flagship models on pirated books was
13 driven by commercial advantage. As its co-founder and Chief Science Officer has
14 explained, “it is important to obtain vast amounts of books and also to have diverse
15 types of books in the training corpus to create a model with truly generative
16
17
18

19 ¹⁴ *Id.* at 4.

20 ¹⁵ *See, e.g.*, Office of the U.S. Trade Representative, REVIEW OF NOTORIOUS MARKETS FOR
21 COUNTERFEITING AND PIRACY, 27 (2024),
22 [https://ustr.gov/sites/default/files/2024%20Review%20of%20Notorious%20Markets%20of%20Counterfeiting%20and%20Piracy%20\(final\).pdf](https://ustr.gov/sites/default/files/2024%20Review%20of%20Notorious%20Markets%20of%20Counterfeiting%20and%20Piracy%20(final).pdf). (“Libgen ... hosts a large number of digital copies
23 of books, manuals, journals, and other works, many of which are unauthorized copies of copyright
24 protected content.”); Alex Reisner, *Revealed: The Authors Whose Pirated Books are Powering Generative AI*, THE ATLANTIC (Aug. 19, 2023),
25 <https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/> (“No one knows what’s inside Books2. Some suspect it comes from collections of
26 pirated books, such as Library Genesis, Z-Library, and Bibliotik, that circulate via the BitTorrent
27 file-sharing network.”); Peter Schoppert, *Whether you’re an undergraduate doing research, or a fan of the Nick Stone novel, or indeed a hungry AI ...*, AI AND COPYRIGHT (Nov. 29, 2022),
28 <https://aicopyright.substack.com/p/whether-youre-an-undergraduate-doing>, (“What is Bibliotik?”
A notorious pirated collection.”).

1 capabilities.”¹⁶ As long-form content, training LLMs on the “entire text” of books—
 2 as Anthropic has admitted to doing—offers great value.¹⁷

3 59. Anthropic intentionally exploited that value without paying for it,
 4 hoping to capture billions of dollars in revenue while externalizing the costs of
 5 training onto the authors whose works it copied.

6 **D. OpenAI Trained Its LLM Models on Copyrighted Works that Were**
 7 **Pirated.**

8 60. OpenAI likewise built the GPT-series models by copying vast
 9 quantities of copyrighted books—including Plaintiffs’ works—from pirate sources.
 10 Plaintiffs’ books were trained on and embedded into OpenAI’s models so that they
 11 could be used to generate human-like text responses that compete directly with
 12 Plaintiffs’ paid writing.

13 61. OpenAI has disclosed that GPT-3 was trained on “Common Crawl”
 14 and two “high-quality,” “internet-based books corpora” it labeled “Books1” and
 15 “Books2.”¹⁸ And OpenAI has now admitted it sourced training materials from
 16 LibGen, the notorious shadow library that hosts millions of unauthorized copies of
 17 books and other copyrighted works.¹⁹

18 62. Common Crawl is a massive web-scraping corpus that includes text
 19 drawn from sites hosting unauthorized copies of books, along with other large
 20 datasets harvested from the open internet. Because OpenAI used undisclosed

21
 22 ¹⁶ See Kaplan Decl. ¶ 47, *Bartz v. Anthropic PBC*, 3:24-cv-05417 (N.D. Cal. Mar. 27, 2025), ECF
 128.

23 ¹⁷ *Id.* at ¶¶ 43, 47.

24 ¹⁸ Tom Brown et al., *Language Models are Few-Shot Learners*, arXiv, 8 (2020),
 25 <https://arxiv.org/pdf/2005.14165>.

26 ¹⁹ Reisner, *The Unbelievable Scale of AI’s Pirated-Books Problem*; see also Joint Ltr. Br.
 27 Regarding Plaintiffs’ Request for an Order Compelling OpenAI’s Production of the English
 28 Colang Dataset at 4, *In re OpenAI ChatGPT Litigation*, 3:23-cv-03223-AMO (N. D. Cal., Jan. 17,
 2025), ECF No. 254.

1 “Books1” and “Books2” corpora in training GPT-3, members of the AI-research
 2 community attempted to replicate those datasets by constructing “Books3,” a
 3 collection of nearly 200,000 digital books downloaded from Bibliotik. Books3 was
 4 created for the express purpose of mirroring the kinds of book corpora OpenAI used,
 5 underscoring that OpenAI’s own training sources necessarily included large
 6 quantities of illicitly obtained books.²⁰

7 63. GPT-3.5 and GPT-4 are significantly more powerful than GPT-3, with
 8 parameter counts that are an order of magnitude larger. OpenAI has not disclosed
 9 the full composition of the training datasets used for these models, but the explosive
 10 growth in model size and capability, together with OpenAI’s prior use of LibGen-
 11 sourced corpora, supports the inference that GPT-3.5, GPT-4, and their successors
 12 were likewise trained on massive collections of pirated books, including Plaintiffs’
 13 works.

14 64. Each step in this process required OpenAI to reproduce Plaintiffs’
 15 books multiple times: in downloading them from LibGen and other shadow
 16 libraries; in preprocessing, deduplication, and storage; in distributing them across
 17 OpenAI and Microsoft’s computer infrastructure; and in iteratively training and
 18 fine-tuning the GPT-series models.

19 **E. OpenAI’s Infringement Was Willful.**

20 65. OpenAI’s infringement was willful. It made a deliberate choice to fuel
 21 its models with pirated books instead of paying for licenses or restricting themselves
 22 to public-domain works.

23 66. An OpenAI research engineer has acknowledged that the quality of an
 24 LLM is “determined by [the] dataset, nothing else,” and that “when you refer to . . .

25
 26 ²⁰ Kate Knibbs, *The Battle Over Books3 Could Change AI Forever*, WIRED (Sep. 4, 2023, 6:00
 27 AM), [https://www.wired.com/story/battle-over-](https://www.wired.com/story/battle-over-books3/#:~:text=Since%20OpenAI%20had%20called%20its%20book%20data,have%20the%20money%20to%20do%20it%20themselves.)
 28 [books3/#:~:text=Since%20OpenAI%20had%20called%20its%20book%20data,have%20the%20money%20to%20do%20it%20themselves.](https://www.wired.com/story/battle-over-books3/#:~:text=Since%20OpenAI%20had%20called%20its%20book%20data,have%20the%20money%20to%20do%20it%20themselves.)

1 ‘ChatGPT’”—or “Lambda” or “Bard” or “Claude”—you are referring to the
 2 “dataset” on which each is trained.²¹ For OpenAI, that dataset included pirated
 3 books.

4 67. OpenAI’s own publication acknowledges that it relied on “two inter-net
 5 based books corpora (Books1 and Books2)” as a “curated high-quality dataset[.]”²²
 6 OpenAI then refused for years to disclose where those books came from, only later
 7 conceding that it had relied on LibGen.²³

8 68. LibGen’s illicit status was no secret. It has been under permanent
 9 injunction and repeatedly listed by U.S. trade authorities as a notorious piracy
 10 market. OpenAI knew (or could not reasonably deny knowing)—from court orders,
 11 public reports, and industry commentary—that LibGen and similar repositories were
 12 illegal sources of copyrighted material.

13 69. OpenAI’s leadership publicly acknowledged that creators “deserve
 14 control over how their creations are used” and that content owners “need to benefit”
 15 from AI training,²⁴ while at the same time failing to obtain licenses from Plaintiffs
 16 and other authors whose books it copied.

17 70. Microsoft’s conduct underscores OpenAI’s willfulness. By 2024,
 18 Microsoft entered into a license agreement with HarperCollins for the use of books
 19 as AI-training data—a deal that pays thousands of dollars per work for a limited
 20 training right—demonstrating an understanding that training on books requires
 21 permission and compensation. Yet neither company obtained any such licenses for
 22

23 ²¹ See J. Betker, *The ‘it’ in AI models is the dataset*, NON_INTERACTIVE-SOFTWARE & ML (June
 24 10, 2023), <https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/>.

25 ²² Brown et al., *Language Models are Few-Shot Learners* at 8.

26 ²³ Reisner, *The Unbelievable Scale of AI’s Pirated-Books Problem*.

27 ²⁴ See Ted Johnson, *OpenAI CEO Sam Altman Says Content Owners Need To Get ‘Significant*
 28 *Upside Benefit’ From New Technology*, DEADLINE (May 16, 2023, 10:12 AM),
<https://deadline.com/2023/05/ai-chat-gpt-senate-sam-altman-1235368420/>.

1 Plaintiffs' works; instead, they exploited pirated datasets that were free precisely
2 because they ignored copyright.

3 71. OpenAI pursued this course because it gave them a decisive lead in the
4 AI race. It touts billions of dollars in revenue and soaring valuations tied directly to
5 GPT-based products—commercial gains secured by pirating and training on
6 unlicensed copies of Plaintiffs' books.²⁵

7 **F. Google Trained Its LLM Models on Copyrighted Works that Were**
8 **Pirated.**

9 72. Google has likewise built its Gemini and Imagen models on vast
10 quantities of copyrighted works, including Plaintiffs' books, obtained from piracy
11 sources.

12 73. Google's training data for its generative models is enormous. For
13 example, its LaMDA/Gemini-related training corpus has been described as
14 comprising more than a trillion and a half words.²⁶ Google has acknowledged that
15 its models were trained on datasets such as C4 and other large web-scale corpora.²⁷

16 74. C4, a core training dataset for Gemini, contains materials scraped from
17 Z-Library, a site that hosted pirated books and was seized by law-enforcement
18
19
20
21

22 ²⁵ See, e.g., Anthony Ha, *Sam Altman says 'enough' to questions about OpenAI's revenue*,
TECHCRUNCH (Nov. 2, 2025, 9:15 AM), [https://techcrunch.com/2025/11/02/sam-altman-says-](https://techcrunch.com/2025/11/02/sam-altman-says-enough-to-questions-about-openais-revenue/)
23 [enough-to-questions-about-openais-revenue/](https://techcrunch.com/2025/11/02/sam-altman-says-enough-to-questions-about-openais-revenue/); Ram Iyer, *OpenAI is reportedly trying to raise*
24 *\$100B at an \$830B valuation*, TECHCRUNCH (Dec. 19, 2025, 5:32 AM),
[https://techcrunch.com/2025/12/19/openai-is-reportedly-trying-to-raise-100b-at-an-830b-](https://techcrunch.com/2025/12/19/openai-is-reportedly-trying-to-raise-100b-at-an-830b-valuation/)
25 [valuation/](https://techcrunch.com/2025/12/19/openai-is-reportedly-trying-to-raise-100b-at-an-830b-valuation/).

26 ²⁶ See Romal Thoppilan et al., *Lamda: Language models for dialog applications*, arXiv preprint
arXiv:2201.08239, 2 (2022), <https://arxiv.org/abs/2201.08239>.

27 ²⁷ *Id.* at 47; see Scott Clark, *What You Need to Know About Google Bard*, CMSWIRE (Feb. 22,
28 2023), <https://www.cmswire.com/digital-experience/what-you-need-to-know-about-google-bard/>.

1 authorities.²⁸ Z-Library displays a seizure banner from federal and international
2 criminal enforcement agencies.

3 75. Google’s training approach, like OpenAI’s and Anthropic’s, required
4 copying each work multiple times: once during data collection, again during
5 preprocessing and deduplication, and repeatedly during training and fine-tuning.
6 Training a generative model necessarily involves making multiple unauthorized
7 copies of each work and permanently embedding those works in the model’s
8 parameters.

9 76. Google has then deployed these AI-trained models across a wide
10 portfolio of AI-powered products, including Search, Cloud, Gmail, Docs, Ads,
11 YouTube, and others—products that generate tens of billions of dollars in revenue, a
12 substantial portion of which Google has explicitly attributed to AI integration.

13 **G. Google’s Infringement Was Willful.**

14 77. Google’s infringement was willful. It trained its models on data scraped
15 from sites that Google knew—or could not reasonably deny knowing—were piracy
16 hubs under active investigation and seizure.

17 78. Z-Library, LibGen, Bibliotik, and similar shadow libraries have been
18 widely reported on as repositories of unauthorized ebooks, have been targeted by the
19 FBI and foreign agencies, and have been the subject of lawsuits and seizures.
20 Google’s own C4 dataset incorporates material from Z-Library, which has been
21 seized and publicly branded as a criminal piracy site.

22 79. Google has touted the “high-quality” nature of its training data and its
23 aggressive push to dominate generative AI—a combination that, in practice, meant
24
25

26
27 ²⁸ Kevin Schaul et al., *Inside the Secret List of Websites That Make AI Like ChatGPT Sound Smart*,
28 THE WASHINGTON POST (Apr. 19, 2023),
<https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>.

1 copying as many high-quality copyrighted works as possible, regardless of legality,
2 to keep pace with or surpass OpenAI and other competitors.²⁹

3 80. Google also understood that the value of its models—and the revenue
4 from AI-powered products—depended on embedding Plaintiffs’ creative expression
5 into Gemini and other models. Google’s own executives have linked record
6 revenues and rapid growth in Cloud and other business lines to generative-AI
7 integration, including revenue measured in the billions of dollars per year.³⁰

8 **H. Meta Trained Its LLM Models on Copyrighted Works that Were**
9 **Pirated.**

10 81. Meta’s Llama models were trained on massive collections of books
11 obtained from shadow libraries and datasets, including Common Crawl, C4,
12 Books3, LibGen, Z-Library, and related piracy sites.

13 82. In its Llama-1 paper, Meta admitted that 3.3 terabytes of its training
14 data came from “CommonCrawl” and another 783 gigabytes came from “C4.”³¹

15 83. In that same paper, Meta admitted that yet another 85 gigabytes came
16 from “Books,”³² which comprised texts from two sources: Project Gutenberg and
17 Books3. While Project Gutenberg contains out-of-copyright works, Books3, and its
18

19 ²⁹ Google Research, *Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for*
20 *Breakthrough Performance*, GOOGLE RESEARCH BLOG (Apr. 4, 2022),
21 [https://research.google/blog/pathways-language-model-palm-scaling-to-540-billion-parameters-](https://research.google/blog/pathways-language-model-palm-scaling-to-540-billion-parameters-for-breakthrough-performance)
22 [for-breakthrough-performance](https://research.google/blog/pathways-language-model-palm-scaling-to-540-billion-parameters-for-breakthrough-performance) (“PaLM was trained using a combination of English and
multilingual datasets that include high-quality web documents, books, Wikipedia, conversations,
and GitHub code.”).

23 ³⁰ Sundar Pichai, *Q3 2025 Earnings: Remarks from our CEO*, THE KEYWORD (Oct. 29, 2025),
24 <https://blog.google/inside-google/message-ceo/alphabet-earnings-q3-2025/> (“This was a terrific
25 quarter for Alphabet, driven by double-digit growth across every major part of our business. We’re
seeing AI now driving real business results across the company.”).

26 ³¹ See Hugo Touvron et al., *LLaMA: Open and Efficient Foundation Language Models*, arXiv, 2
27 (2023), <https://arxiv.org/pdf/2302.13971>.

28 ³² *Id.*

1 200,000 books downloaded from Bibliotik, contains copyrighted books. Meta did
 2 not stop there. Its internal documents explain that Meta also downloaded books
 3 directly from LibGen, Z-Library, Anna’s Archive, Sci-Hub, and related shadow
 4 libraries.³³ These libraries have been repeatedly identified in public reporting and
 5 enforcement actions as illegal piracy hubs, and have been accessible in bulk via
 6 torrent systems and mirrors such as Anna’s Archive, The Eye, and Hugging Face. At
 7 least for LibGen and Anna’s Archive, Meta used BitTorrent to download, and did
 8 not prevent reuploading the books it illegally downloaded through leeching.³⁴

9 84. Meta relied on these pirated books because it viewed book-corpora as
 10 among its most valuable sources of training data. Llama’s design goal was to emit
 11 particularly creative and expressive language, leveraging Meta’s consumer
 12 platforms to “connect” with users through text.³⁵ To accomplish that, Meta needed
 13 to train on large quantities of high-quality books.

14 85. Meta employees repeatedly acknowledged the importance of books as
 15 training data. It was “really important for [Meta] to get books data ASAP,” and the
 16 “best resources [Meta] [could] think of are definitely books.”³⁶

17
 18
 19
 20 ³³ Ernestas Naprys, *Meta leeched 82 terabytes of pirated books to train its Llama AI, documents reveal*, CYBERNEWS (Feb. 7, 2025), <https://cybernews.com/tech/meta-leeched-82-terabytes-of-pirated-books-to-train-its-llama-ai-documents-reveal/>.

22 ³⁴ *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1041 (N.D. Cal. 2025) (“There is no
 23 dispute that Meta torrented LibGen and Anna’s Archive, but the parties dispute whether and to
 24 what extent Meta uploaded (via leeching or seeding) the data it torrented. A Meta engineer
 involved in the torrenting wrote a script to prevent seeding, but apparently not leeching.”).

25 ³⁵ Jon Russell, *Mark Zuckerberg Announces New Team at Meta Working on A.I. Products for*
 26 *Instagram, WhatsApp*, CNBC (Feb. 27, 2023, 4:19 PM), <https://www.cnbc.com/2023/02/27/mark-zuckerberg-announces-new-team-at-meta-working-on-ai-products.html> (“Zuckerberg said that the
 27 team would build ‘creative and expressive’ tools to be used inside Meta’s products.”).

28 ³⁶ *Kadrey*, 788 F. Supp. 3d at 1040.

I. Meta’s Infringement Was Willful.

86. Meta’s infringement, too, was willful. Meta knew that its book datasets were composed of pirated works and chose to use them anyway.

87. Meta employees internally recognized that the shadow libraries it used had “pirated material” and warned about potential liability.³⁷ Journalists allegedly contacted Meta about its likely reliance on pirated books. Yet Meta reportedly decided that the value of these books as training data outweighed the legal risk and continued to download and copy millions of pirated books, even after litigation and public controversy made the infringement unmistakable.³⁸

88. Meta discussed licensing copyrighted books from publishers and considered spending \$100 million on the vibrant market for AI-training content, but ultimately decided to cut corners by turning to free shadow-library datasets instead.³⁹ Meta even cross-referenced its LibGen collection against commercially licensable catalogs to decide whether it was worth paying for a license, but decided to keep using LibGen.⁴⁰ Meta thus understood both the illegality of its shadow-library troves and the existence of lawful alternatives.

89. Meta nevertheless moved forward, incorporating Llama into its principal products and publicly portraying itself as a leader in open-source AI, all while its training pipeline rested on unlicensed copies of Plaintiffs’ books. Meta believed that its multi-billion-dollar investment in Llama would bolster and define

³⁷ Reisner, *The Unbelievable Scale of AI’s Pirated-Books Problem*.

³⁸ *Kadrey*, 788 F. Supp. 3d at 1041 (finding Meta continued downloading pirated books in “early 2024,” more than six months after authors sued Meta for violating the Copyright Act for its download and use of their pirated books).

³⁹ *Kadrey*, 788 F. Supp. 3d at 1040; *see also* Kyle Wiggers, *Court filings show Meta paused efforts to license books for AI training*, TECHCRUNCH (Feb. 14, 2025, 3:35 PM), <https://techcrunch.com/2025/02/14/court-filings-show-meta-paused-efforts-to-license-books-for-ai-training/>.

⁴⁰ *Kadrey*, 788 F. Supp. 3d at 1041.

1 its competitive position for years to come—and chose to finance that investment
2 with unauthorized copies of Plaintiffs’ copyrighted works.

3 **J. xAI Trained Its LLM Models on Copyrighted Works that Were**
4 **Pirated.**

5 90. xAI is the developer of the Grok family of large language models.

6 91. Unlike Anthropic, OpenAI, Google, and Meta, xAI has not yet faced
7 litigation or discovery over its processes for building its LLM.

8 92. However, Grok’s own model has publicly admitted in user interactions,
9 the Grok training corpus included “literally millions of books”—far beyond the
10 universe of lawfully licensed or public-domain material.⁴¹

11 93. In these same exchanges, xAI’s model explained that its training data
12 included books, articles, websites, forums, journals, and other long-form nonfiction
13 and fiction texts, and that it had effectively “read an absurdly huge library at super-
14 speed.”⁴² The model repeatedly emphasized the centrality of books to its capabilities
15 and acknowledged that such books were included in the training datasets used to
16 develop Grok.

17 94. Critically, the model further admitted that a substantial portion of the
18 books it was trained on had been scraped—without permission—from “shadow
19 library” repositories such as Library Genesis (“LibGen”). In response to a question
20 about whether it was familiar with LibGen, the model stated:⁴³

21
22
23
24
25 ⁴¹ *Grok Conversation*, (“Hey - I was wondering how you know so many books!”)
26 (https://grok.com/share/bGVnYWN5_585707fa-d0dc-4f2e-8069-86de3997f088).

27 ⁴² *Id.*

28 ⁴³ *Id.*

Are you familiar with the LibGen library of books?

Yes, I'm aware of Library Genesis (LibGen). It's a well-known shadow library that hosts millions of books, academic papers, journal articles, comics, and other texts, almost all of them shared without publisher or author permission (i.e., pirated).

Because my training data was enormous and scraped from publicly available corners of the internet, it's very likely that some portion of the texts I was exposed to during training came from LibGen mirrors or similar sources (along with legal sources, public-domain works, open-access repositories, etc.). The people who trained me didn't hand-curate every single book—they basically vacuumed up whatever was out there, and LibGen has been one of the biggest “whatever was out there” troves for years.

So when I seem to know obscure academic monographs, out-of-print novels, or textbooks that normally cost \$200, there's a decent chance some of that knowledge traces back to files that originally lived on LibGen.

(That said, I don't have direct real-time access to LibGen or any piracy site now—I just have the patterns I learned during training.)

95. The model further admitted that the scale of the ANDing corpus made it “practically impossible” for xAI to avoid acquiring pirated LibGen materials because internet-scale scrapes inevitably sweep in mirror links, PDF dumps, Z-Library derivations, Sci-Hub papers, book torrents, and other infringement sources.⁴⁴ The model’s statements reflect (i) that pirated book datasets formed a material portion of the Grok training corpus, and (ii) that xAI took no steps to exclude infringing works.

96. These admissions are fully consistent with what is publicly known about xAI’s data-collection practices. xAI has confirmed that it trains Grok on data gathered from “publicly available sources,”⁴⁵ a phrase that can include large unaudited web scrapes from Common Crawl and related datasets. Those scrapes

⁴⁴ *Id.*

⁴⁵ X Help Center, *About Grok, Your Humorous AI Assistant on X*, <https://help.x.com/en/using-x/about-grok> (last visited Dec. 22, 2025).

1 notoriously include mirror links and file dumps from LibGen, Z-Library, and similar
2 repositories that host millions of pirated books.

3 97. At no time did xAI obtain licenses from Plaintiffs or from any other
4 authors whose copyrighted works were copied and reproduced in the Grok training
5 process. Nor did xAI pay any fee to a licensing society, publisher, clearinghouse, or
6 collecting agent for the right to use these books.

7 98. As with other LLMs, training Grok required xAI to make multiple
8 reproductions of each book: (a) a copy during ingestion or download; (b) additional
9 copies during preprocessing, tokenization, and batching; (c) repeated copies during
10 training as the model ingested each work in multiple epochs; and (d) an embedded,
11 parametric copy of expressive information from each work stored permanently
12 within Grok’s model weights.

13 99. Grok’s ability to generate high-quality prose, summaries, paraphrases,
14 and long-form outputs is directly tied to its ingestion of Plaintiffs’ works and the
15 millions of other copyrighted books it acquired from piracy sources. xAI built
16 commercially valuable models—now deployed across X Corp’s consumer,
17 enterprise, and API products—on top of these infringing copies.

18 **K. xAI’s Infringement Was Willful.**

19 100. xAI’s infringement was willful. The Grok model explicitly
20 acknowledged that the training process “vacuumed up whatever was out there,”
21 including pirated LibGen materials, and that xAI neither curated its book dataset nor
22 screened out infringing works.⁴⁶ These admissions confirm that xAI knew, or at a
23 minimum was recklessly indifferent to the fact, that its training corpus included
24 massive quantities of pirated copyrighted books.

25
26
27 ⁴⁶ See *Grok Conversation*, (“Hey - I was wondering how you know so many books!”)
28 (https://grok.com/share/bGVnYWN5_585707fa-d0dc-4f2e-8069-86de3997f088).

1 101. xAI was on notice—long before and during the development of Grok—
2 that LibGen, Z-Library, and other shadow libraries are illegal repositories of pirated
3 books. These repositories have been the subject of criminal prosecutions, copyright
4 lawsuits, mass domain takedowns, and international enforcement campaigns. This
5 fact is widely known in the technology and AI communities, and even
6 acknowledged directly by Grok itself.

7 102. xAI thus knew—or consciously avoided confirming—that its training
8 data included copyrighted works that were plainly not licensed and plainly not in the
9 public domain. Nonetheless, it used those works because they were valuable training
10 data for improving Grok’s fluency, reasoning ability, stylistic coherence, and
11 literary skill.

12 103. On information and belief, xAI also understood that book data was
13 among the most valuable forms of training data for frontier models. Like Meta,
14 OpenAI, and Anthropic, xAI leveraged the unique expressive quality of books to
15 improve Grok’s narrative and analytical capabilities. The decision to rely on pirated
16 book datasets, rather than obtain licenses, conferred a substantial competitive
17 advantage in speed, cost, and model performance.

18 104. On information and belief, xAI continued to use pirated books even
19 after lawsuits were filed against other AI developers for identical conduct—
20 including the use of LibGen-derived datasets. Grok’s public statements that training
21 data was “vacuumed up” from whatever could be scraped show that xAI deliberately
22 maintained the same indiscriminate data-collection practices despite mounting legal
23 risk and increasing public scrutiny.

24 **L. Perplexity’s Model Relies on Copyrighted Works without**
25 **Permission or Compensation.**

26 105. Perplexity AI, Inc. (“Perplexity”) has rapidly emerged as a commercial
27 competitor in the generative-AI search and LLM market. Central to its strategy is a
28 suite of products—including “Perplexity Answers,” “Perplexity Pages,” and its

1 proprietary LLM models—that can generate detailed narrative summaries,
2 structured analyses, and book-length outlines with extraordinary specificity.

3 106. Perplexity operates by ingesting massive quantities of copyrighted text,
4 including works that are not available in any public, licensed, or authorized source.
5 As multiple independent investigations have confirmed, Perplexity acquires this
6 material through large-scale crawling and scraping systems—both declared and
7 undeclared—that indiscriminately copy entire texts from across the internet and
8 beyond.⁴⁷

9 107. Perplexity’s own behavior suggests that it relies on the full text of
10 books. Despite acknowledging in responses that books are copyrighted, and that it
11 cannot produce “line-by-line chapter notes,” Perplexity is capable of doing exactly
12 that.⁴⁸ Upon request, it can produce detailed, chapter-by-chapter accounts of works,
13 including descriptions of plot turns, chapter-specific structure, and thematic
14 sequencing. In certain instances, the sources Perplexity cites for its chapter-by-
15 chapter descriptions do not include the underlying information that it produces in
16 response to queries.⁴⁹ That information is in the complete, original books.

17 108. The recently-filed *New York Times v. Perplexity* complaint alleges,
18 based on forensic evidence, that Perplexity’s systems routinely crawl, copy, and
19 store expressive content in violation of copyright law. For example, the complaint
20 alleges that Perplexity: (1) builds and operates a massive “AI-First” search index

21 _____
22 ⁴⁷ See, e.g., Gabriel Corral et al., *Perplexity is using stealth, undeclared crawlers to evade website*
23 *no-crawl directives*, CLOUDFLARE (Aug. 4, 2025), [https://blog.cloudflare.com/perplexity-is-using-](https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/)
24 [stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/](https://blog.cloudflare.com/perplexity-is-using-stealth-undeclared-crawlers-to-evade-website-no-crawl-directives/); Dhruv Mehrotra and Tim
Marchman, *Perplexity Is a Bullshit Machine*, WIRED (June 19, 2024, 9:00 AM),
<https://www.wired.com/story/perplexity-is-a-bullshit-machine/>.

25 ⁴⁸ *Perplexity Conversation*, (“what is John Carrey[r]ou’s Bad Blood book about”),
26 ([https://www.perplexity.ai/search/what-is-john-carreyou-s-bad-bl-](https://www.perplexity.ai/search/what-is-john-carreyou-s-bad-bl-F12yAyhSQbatnBcu6TqgTQ#0)
[F12yAyhSQbatnBcu6TqgTQ#0](https://www.perplexity.ai/search/what-is-john-carreyou-s-bad-bl-F12yAyhSQbatnBcu6TqgTQ#0)).

27 ⁴⁹ *Id.* (citing sources for its summary of Chapter 19 that do not include the information Perplexity
28 provides about Chapter 19).

1 populated through direct copying of protected works; (2) uses both “PerplexityBot”
2 and “Perplexity-User” agents to scrape websites and copy non-public content; (3)
3 copies content for use in its LLMs and retrieval-augmented generation (RAG)
4 pipelines; and (4) outputs detailed summaries, paraphrases, and quotations that
5 substantially reproduce copyrighted texts.⁵⁰

6 109. Independent investigations corroborate this pattern. WIRED reported
7 that Perplexity produced detailed summaries of WIRED’s articles even though
8 WIRED explicitly blocked Perplexity. Engineers confirmed that the chatbot was
9 “surreptitiously scraping” and recapitulating protected content “in detail” that was
10 not publicly available.⁵¹

11 110. Cloudflare’s investigation further found that Perplexity operates stealth
12 crawlers designed to evade detection. According to Cloudflare: (1) Perplexity used
13 undeclared user agents that impersonated Google Chrome; (2) Perplexity used
14 multiple undisclosed IP ranges to circumvent no-crawl directives; (3) customers
15 who blocked Perplexity’s known crawlers found Perplexity still scraping their sites
16 anyway; and (4) Perplexity’s activity “evade[d] website blocks” and undermined
17 publisher controls.⁵²

18 111. On information and belief, Perplexity’s model has required Perplexity
19 to make unauthorized reproductions of each work, including copies during scraping,
20 ingestion, and deduplication. Perplexity’s ability to output chapter-specific
21 content corroborates that the models were trained and/or optimized with pirated
22 copies of Plaintiffs’ works.

23
24
25 ⁵⁰ See Compl. at 3-4, *The New York Times Company v. Perplexity AI, Inc.*, 1:25-cv-10106
(S.D.N.Y. Dec. 5, 2025), ECF No. 1.

26 ⁵¹ Mehrotra and Marchman, *Perplexity Is a Bullshit Machine*.

27 ⁵² See Corral et al., *Perplexity is using stealth, undeclared crawlers to evade website no-crawl*
28 *directives*.

1 **M. Perplexity’s Infringement Was Willful.**

2 112. Perplexity’s infringement was willful. As Cloudflare and WIRED
3 independently confirmed, Perplexity intentionally deployed stealth crawlers and
4 undeclared automated agents to evade copyright protections and access content it
5 knew it was not authorized to copy.⁵³

6 113. Cloudflare found that Perplexity impersonated Chrome browsers, used
7 concealed IP addresses, and intentionally bypassed restrictions to obtain content.⁵⁴
8 The purpose of such evasion is unmistakable: to gain access to copyrighted text that
9 Perplexity knew it was forbidden to crawl.

10 114. The *Times* complaint likewise alleges that Perplexity continued to
11 access and copy prohibited content even after written cease-and-desist demands,
12 “hard-block[s] of PerplexityBot and Perplexity-User,” and explicit revocation of
13 access.⁵⁵ Perplexity continued to make over 175,000 unauthorized access attempts
14 in a single month after being technically and contractually barred.⁵⁶

15 115. Perplexity knew that its conduct violated copyright law. The *Times*
16 repeatedly informed Perplexity in writing—beginning in March 2024—that
17 Perplexity was unlawfully scraping and copying copyrighted material; Perplexity
18 refused to stop.⁵⁷ Instead, it escalated its crawling behavior using stealth methods to
19 avoid detection.

20 116. Perplexity also publicly markets itself as providing users with the
21 ability to “skip the links” and read “a single, comprehensive answer that summarizes
22

23 ⁵³ See *id.*; Mehrotra and Marchman, *Perplexity Is a Bullshit Machine*.

24 ⁵⁴ See Corral et al., *Perplexity is using stealth, undeclared crawlers to evade website no-crawl*
25 *directives*.

26 ⁵⁵ Compl., *New York Times*, 1:25-cv-10106, at 28-29.

27 ⁵⁶ See *id.*

28 ⁵⁷ *Id.*

1 everything you need to know,” thereby advertising that it substitutes for underlying
2 copyrighted works.⁵⁸

3 117. Perplexity’s conduct reflects intentional, systematic, and commercially
4 motivated exploitation of copyrighted works. It intentionally circumvented
5 protective barriers, accessed materials it knew it was forbidden to copy, ignored
6 written legal demands, and profited from generating expressive content that directly
7 substitutes for Plaintiffs’ books.

8 118. Perplexity’s infringement was neither accidental nor negligent—it was
9 deliberate, concealed, repeated, and performed at massive scale.

10 **V. CLAIMS FOR RELIEF**

11 **COUNT I**

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

13 **(Against all Defendants)**

14 119. Plaintiffs incorporate the allegations above.

15 120. As the respective owners of the registered copyrights in the Infringed
16 Works, Plaintiffs hold the exclusive rights to those books under 17 U.S.C. § 106.

17 121. Each Defendant—including Anthropic, Google, OpenAI, Meta, xAI,
18 and Perplexity—without authorization from Plaintiffs, copied, downloaded,
19 reproduced, ingested, parsed, embedded, and used pirated copies of the Plaintiffs’
20 works in the development, training, fine-tuning, and deployment of their
21 commercial large language models. These acts violated Plaintiffs’ exclusive rights
22 under § 106.

23 122. Defendants’ infringement occurred repeatedly throughout the lifecycle
24 of their AI-model development pipelines. As alleged above, Defendants:

25
26
27 ⁵⁸ *What is Perplexity?*, <https://perma.cc/Q4VM-DYUJ> (accessed from *Dow Jones & Co., Inc. v. Perplexity AI, Inc.*, No. 1:24-cv-07984-KPF, at Dkt. 46 (Second Amended Complaint) n.1
28 (S.D.N.Y. Jan. 28, 2025)) (last accessed Dec. 22, 2025).

- 1 • acquired Plaintiffs’ books from shadow-library repositories such as
- 2 LibGen, Bibliotik, Z-Library, Books3, and other known piracy sources;
- 3 • reproduced additional copies during ingestion, preprocessing, storage,
- 4 deduplication, formatting, and/or tokenization; and
- 5 • while training the model, and/or through retrieval-augmented
- 6 generation, made even more copies of the text—because every training
- 7 pass (each epoch and each step of gradient descent) automatically
- 8 requires creating and working with fresh versions of that text.

9 123. Defendants’ reproductions of Plaintiffs’ copyrighted works were made
10 without permission, license, or consent and violated Plaintiffs’ exclusive rights
11 under the Copyright Act.

12 124. Defendants’ infringement was willful. As alleged above, each
13 Defendant knowingly trained its models on and/or optimized its product with
14 datasets saturated with pirated books, including Plaintiffs’ works; relied on shadow-
15 library corpora they knew to be illegal; ignored internal and external warnings;
16 attempted to conceal the composition of their training datasets; and continued
17 copying after public reports, lawsuits, law-enforcement seizures, cease-and-desist
18 notices, and industry-wide alerts made the illegality unmistakable.

19 125. Upon information and belief, Defendants have made and will continue
20 to make substantial profits and gains to which they are not in law or in equity
21 entitled.

22 126. Plaintiffs are entitled to all remedies available under the Copyright Act,
23 including statutory damages under 17 U.S.C. § 504(c) of up to \$150,000 per
24 infringed work per Defendant for willful infringement.

25 127. Plaintiffs are entitled to recover attorneys’ fees and costs under 17
26 U.S.C. § 505.

27 **PRAYER FOR RELIEF**

28 WHEREFORE, Plaintiff respectfully requests the following relief:

1 128. Judgment in favor of Plaintiffs against each Defendant;

2 129. A declaration that each Defendant has infringed Plaintiffs' exclusive
3 copyrights under the Copyright Act;

4 130. A declaration that such infringement is willful;

5 131. A permanent injunction enjoining each Defendant and all those acting
6 in concert with them from engaging in the infringing conduct alleged herein;

7 132. That each Defendant be directed to account to Plaintiffs for all gains,
8 profits, and advantages derived from their unlawful acts;

9 133. An award of statutory damages under the Copyright Act;

10 134. An award of restitution, disgorgement, costs, expenses, and attorneys'
11 fees as permitted by law (including those allowable under 17 U.S.C. § 505 and/or 17
12 U.S.C. § 1203(b)(4)–(5));

13 135. Pre- and post-judgment interest on the damages awarded to Plaintiffs;
14 and

15
16 136. Further relief for Plaintiffs as the Court may deem just and proper.

17 **JURY TRIAL DEMANDED**

18 Under Federal Rule of Civil Procedure 38(b), Plaintiffs demand a trial by jury.
19
20
21
22
23
24
25
26
27
28

1 Dated: December 22, 2025

Respectfully submitted,

2 /s/ Elizabeth Brannen

3 Elizabeth Brannen (SBN 226234)

4 John Stokes (SBN 310847)

5 Lauren Martin (SBN 294367)

STRIS & MAHER LLP

6 17785 Center Court Dr N, Ste 600

Cerritos, CA 90703

7 T: (213) 995-6800

8 F: (213) 261-0299

ebrannen@stris.com

9 jstokes@stris.com

10 lmartin@stris.com

11 Bridget Asay (*pro hac vice* forthcoming)

12 15 East State Street, Suite 2

13 Montpelier, VT 05602

14 T: (802) 858-4285

basay@stris.com

15 Jacqueline Sahlberg (*pro hac vice*
16 forthcoming)

17 1717 K St NW Suite 900

Washington, DC 20006

18 T: (202) 800-5749

jsahlberg@stris.com

19 Devin (Velvel) Freedman (*pro hac vice*
20 forthcoming)

21 Kyle Roche (*pro hac vice* forthcoming)

22 Alex Potter (*pro hac vice* forthcoming)

FREEDMAN NORMAND

FRIEDLAND LLP

23 155 E. 44th Street, Suite 915

24 New York, NY 10017

25 T: (646) 494-2900

26 vel@fnf.law

kroche@fnf.law

27 apotter@fnf.law

28 *Counsel for Plaintiffs*