

1 Elizabeth Brannen (SBN 226234)
 ebrannen@stris.com
 2 John Stokes (SBN 310847)
 jstokes@stris.com
 3 Lauren Martin (SBN 294367)
 lmartin@stris.com
 4 **STRIS & MAHER LLP**
 5 17785 Center Court Dr N, Ste 600
 Cerritos, CA 90703
 6 T: (213) 995-6800
 7 F: (213) 261-0299

8 Christopher M. Rigali (*pro hac vice*)
 crigali@stris.com
 9 Jacqueline Sahlberg (*pro hac vice*)
 jsahlberg@stris.com
 10 **STRIS & MAHER LLP**
 11 1717 K St NW Suite 900
 Washington, DC 20006
 12 T: (202) 800-5749

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

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

13 *Counsel for Plaintiffs*

14 **UNITED STATES DISTRICT COURT**
 15 **NORTHERN DISTRICT OF CALIFORNIA**

16 CAMBRONNE INC., LISA BARRETTA,
 17 PHILIP SHISHKIN, JANE ADAMS,
 MATTHEW SACKS, and MICHAEL
 KOCHIN,

18 Plaintiffs,

19 v.

20 ANTHROPIC PBC; GOOGLE LLC;
 21 OPENAI, INC.; OPENAI OPCO LLC;
 22 OPENAI GP LLC; OPENAI GLOBAL LLC;
 OAI CORPORATION LLC; OPENAI
 HOLDINGS LLC;
 23 META PLATFORMS, INC.;
 XAI CORPORATION;
 24 PERPLEXITY AI, INC.; APPLE, INC.; and
 NVIDIA CORPORATION

25 Defendants.

Civil Case No.: 5:2025-cv-10897-PCP

AMENDED COMPLAINT

DEMAND FOR JURY TRIAL

1 1. Plaintiffs Cambronne, Inc., Lisa Barretta, Philip Shishkin, Jane Adams¹, Matthew
2 Sacks, and Michael Kochin (collectively “Plaintiffs”), each proceeding in their individual capacity,
3 bring this action against Anthropic PBC (“Anthropic”), Google LLC (“Google”), OpenAI, Inc. and
4 its affiliated entities (“OpenAI”)², Meta Platforms, Inc. (“Meta”), xAI Corporation (“xAI”), and
5 Perplexity AI, Inc. (“Perplexity”), Apple Inc. (“Apple”), and NVIDIA Corporation (“NVIDIA”),
6 (collectively, “Defendants”), and allege as follows:

7 **I. INTRODUCTION**

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

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

21 4. Defendants targeted Plaintiffs’ works because they were of exceptional value as
22 training data. Defendants have acknowledged—internally and publicly—that long-form, high-
23 quality books are the “gold-standard” training material for LLMs. Books teach models how narrative

24 _____
25 ¹ Pursuant to Plaintiff Jane Adams' Notice of Voluntary Dismissal Pursuant to F.R.C.P. 41(a)(1)(A)(i)
26 [*see* ECF No. 50], Plaintiff Jane Adams asserts claims against named Defendants except for
Anthropic PBC.

27 ² Plaintiffs’ claims against OpenAI are severed for transfer to pending multidistrict litigation, *see*
28 ECF Nos. 98 & 99, and will be pursued in the new civil action for the severed claims (*see* ECF No.
99 at ¶ 2) once the Clerk of Court takes such action.

1 flows, how human expression is structured, how syntax and rhythm operate, and how ideas are
2 communicated through creative choices. Instead of paying for that value, Defendants pilfered illegal
3 copies and used those copies to build systems now worth many hundreds of billions of dollars.

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

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

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

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

18 9. Anthropic’s Claude models were trained on datasets containing hundreds of
19 thousands of books obtained from piracy sources, including LibGen, that included Plaintiffs’
20 copyrighted books. Google’s Gemini and Imagen models rely on datasets—including Z-Library and
21 OceanofPDF—that incorporate large collections of pirated works including Plaintiffs’ works.
22 OpenAI and Microsoft’s GPT-series models were trained on vast pirated corpora, including LibGen,
23 enabling products such as ChatGPT, Copilot, GitHub Copilot, and a suite of AI-enhanced Microsoft
24 applications. Meta’s Llama models were trained on massive sets of books downloaded from shadow
25 libraries, including LibGen, that included Plaintiffs’ copyrighted books. xAI’s Grok models and
26 Perplexity’s AI search systems likewise relied on large-scale ingestion of pirated books. Apple’s
27 “Apple Foundation Models” relied upon *The Pile* and Books 3. NVIDIA not only relied upon *The*
28 *Pile* to train its NeMo LLM framework, it reached out directly to the largest illegal pirate site—

1 Anna’s Archive—and just days after being explicitly warned of the illegal nature of its collections
2 NVIDIA’s management gave the “green light” to proceed with illegally downloading hundreds of
3 terabytes of data. These models, all trained and/or optimized on Plaintiffs’ copyrighted books, now
4 anchor multibillion-dollar product ecosystems.

5 10. Although Defendants compete with respect to AI, their infringement arises from a
6 shared factual core and fundamentally interrelated conduct.

7 11. The genesis of this case is a single act: in 2018, an OpenAI employee downloaded
8 pirated copies of books from Library Genesis—a shadow library repeatedly enjoined by federal
9 courts—and used those books to create two internal datasets OpenAI called “LibGen1” and
10 “LibGen2,” later renamed “Books1” and “Books2.” OpenAI used those pirated corpora to train
11 GPT-3, which it released in June 2020 to widespread commercial acclaim. That act—downloading
12 a pirate library to build a commercial AI model—set the template for every Defendant in this case.

13 12. Within weeks of GPT-3’s release, the open-source AI research
14 collective EleutherAI formed with the explicit goal of replicating GPT-3’s capabilities. To do
15 so, EleutherAI assembled and publicly released *The Pile*, a large general-purpose training dataset
16 expressly designed to be downloaded, reused, and incorporated into LLMs by academic researchers,
17 startups, and commercial AI developers. Because OpenAI’s own training data was proprietary and
18 undisclosed, members of the open-source community constructed a pirated book corpus of their
19 own: “Books3,” consisting of approximately 200,000 books copied from the Bibliotik shadow
20 library—a private BitTorrent tracker devoted to pirated ebooks. EleutherAI packaged Books3
21 into *The Pile* and distributed it publicly, effectively democratizing access not just to training data,
22 but to a standardized, “off-the-shelf” corpus of infringing works that any AI developer could
23 download and immediately incorporate into a large-scale training pipeline.

24 13. The remaining Defendants followed OpenAI’s lead. They came to appreciate the
25 commercial significance of large language models, several turned to *The Pile*—and therefore
26 Books3—as a foundational training resource. They did not encounter Books3 incidentally; they
27 adopted it as a high-value source of long-form, editorially polished text that could accelerate model
28 development at a fraction of the cost of lawful licensing. Several defendants used *The Pile* as part

1 of their model-training ecosystem, either directly to train proprietary models or indirectly as a
2 baseline corpus replicated, extended, or fine-tuned to improve commercial performance.

3 14. But the piracy did not stop with *The Pile*. As the generative-AI arms race
4 intensified, other Defendants sought ever-larger volumes of book-length training data and turned
5 to additional shadow libraries—including LibGen, Z-Library, and Anna’s Archive—to supplement
6 what *The Pile* provided. Defendants knew or should have known that these sources were illicit: each
7 had been the subject of criminal prosecutions, civil injunctions, domain seizures, or widespread
8 industry warnings. Employees across the industry raised red flags about using these sources, with
9 some calling them “illegal pirated websites” and warning of liability for downloading from them.
10 Defendants pressed forward because pirated books allowed them to build more sophisticated models
11 faster, more cheaply, and with higher performance than lawful alternatives would permit.

12 15. The result is a cascading pattern of infringement that traces directly back to OpenAI's
13 2018 decision to source its training data from a pirate library. OpenAI demonstrated that pirated
14 books were the fastest path to a frontier language model. The open-source community replicated
15 that approach through Books3 and *The Pile*. And the remaining Defendants—Anthropic, Google,
16 Meta, xAI, Perplexity, Apple, and NVIDIA—exploited either *The Pile* and additional shadow
17 libraries as they raced to catch up. Each escalation compounded the last: from a
18 single LibGen download, to an industry-standard pirated dataset, to the wholesale harvesting of
19 every major shadow library on the internet.

20 16. This case therefore concerns not isolated acts of infringement, but an industry-wide
21 course of conduct in which competing AI companies drew from the same piracy-derived datasets—
22 and then expanded their piracy to new sources—to build rival commercial models. The
23 common origin in OpenAI's LibGen download, the shared reliance on *The Pile* and Books3, and the
24 parallel escalation to additional shadow libraries establish a unifying factual and evidentiary
25 nucleus that links Defendants’ infringement and renders their conduct
26 properly adjudicated together.

27 17. Defendants’ unlawful conduct also did not end with the unauthorized downloads of
28 Plaintiffs’ works. In addition to making unauthorized copies when torrenting shadow libraries,

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

5 18. Defendants’ unauthorized copying of Plaintiffs’ books has inflicted immediate and
6 ongoing harm. Plaintiffs spent years creating the works at issue; Defendants spent seconds copying
7 them. By embedding Plaintiffs’ creative expression into their model parameters and/or optimization,
8 Defendants have appropriated—and continue to monetize—the fruits of Plaintiffs’ copyrighted
9 labor across cloud platforms, consumer products, enterprise tools, advertising systems, and
10 subscription services. Congress created a private right of action in the Copyright Act to prevent this
11 type of infringement. But even with these protections, Congress is under mounting pressure to
12 address the mass harvesting of copyrighted works by AI developers, and is presently considering
13 bi-partisan legislation like the proposed Copyright Labeling and Ethical AI Reporting Act (the
14 “CLEAR Act”) to expose and prevent this very conduct. If enacted as drafted, legislation like the
15 CLEAR Act would require mandatory reporting requirements for companies developing artificial
16 intelligence (AI) models that are trained using original works that are protected under U.S. copyright
17 law, and would create an additional cause of action for copyright owners alleging that generative AI
18 developers failed to give such notice with respect to their works.³

19 19. While Defendants’ conduct constitutes classic copyright infringement, their conduct
20 is unique in that they have willfully infringed Plaintiffs’ copyrights at an unprecedented scale for
21 massive commercial gain.

22 20. To redress Defendants’ repeated, unlawful, and massive infringement of their work,
23 each Plaintiff individually seeks (1) damages, (2) permanent injunctive relief barring Defendants’
24 ongoing infringement, and (3) any additional remedies the law provides.

25
26 ³ See Alex Welch, *Adam Schiff’s Proposed Bill Would Require Tech Companies to Disclose*
27 *Copyrighted Works Used to Train AI*, THEWRAP (February 10, 2026, 7:59 AM),
28 <https://www.thewrap.com/industry-news/public-policy-legal/clear-act-ai-copyright-bill-adam-schiff-john-curtis/> (referencing S. 3813 – 119th Congress).

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

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

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

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

18 25. That is not how Plaintiffs plan to proceed. Under established Supreme Court
19 precedent, “the amount of statutory damages is a question for the jury.”⁴ The Copyright Act thus
20 vests authors with the right to have a jury evaluate the willfulness of infringement and assign a
21 damages amount tailored to the Defendant’s conduct.

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

25
26
27
28

⁴ *Feltner v. Columbia Pictures Television, Inc.*, 523 U.S. 340, 353 (1998).

1 **II. PARTIES**

2 **A. Plaintiffs**

3 27. Plaintiff Cambronne Inc., is wholly owned by John Carreyrou, who is an author and
4 journalist who resides in New York. He is the author of *Bad Blood: Secrets and Lies in a Silicon*
5 *Valley Startup*. His work is contained in pirated online libraries, including Books 3 (and thus, *The*
6 *Pile*), LibGen, Z-Library, and Anna’s Archive. Defendants have directly or indirectly downloaded
7 books illegally contained in *The Pile*, LibGen, Z-Library, or Anna’s Archive, and there is
8 accordingly a reasonable inference that Defendants illegally downloaded Carreyrou’s work.

9 28. Plaintiff Lisa Barretta is an author who resides in Pennsylvania. She is the author of
10 *The Street-Smart Psychic’s Guide to Getting a Good Reading*. Her work is contained in pirated
11 online libraries, including LibGen, Z-Library, and Anna’s Archive. Defendants have directly or
12 indirectly downloaded books illegally contained in *The Pile*, LibGen, Z-Library, or Anna’s Archive,
13 and there is accordingly a reasonable inference that Defendants illegally downloaded Barretta’s
14 work.

15 29. Plaintiff Philip Shishkin is an author and journalist who resides in Washington D.C.
16 He is the author of *Restless Valley: Revolution, Murder and Intrigue in the Heart of Central Asia*.
17 His work is contained in pirated online libraries, including LibGen, Z-Library, and Anna’s Archive.
18 Defendants have directly or indirectly downloaded books illegally contained in *The Pile*, LibGen,
19 Z-Library, or Anna’s Archive, and there is accordingly a reasonable inference that Defendants
20 illegally downloaded Shishkin’s work.

21 30. Plaintiff Jane Adams is an author and journalist who resides in Washington. She is
22 the author of *Boundary Issues: Using Boundary Intelligence to Get the Intimacy You Want and the*
23 *Independence You Need in Life, Love, and Work* and *How to Sell What You Write*. Her works are
24 contained in pirated online libraries, including LibGen, Z-Library, and Anna’s Archive. Defendants
25 have directly or indirectly downloaded books illegally contained in *The Pile*, LibGen, Z-Library, or
26 Anna’s Archive, and there is accordingly a reasonable inference that Defendants illegally
27 downloaded Adams’s works.

28

1 31. Plaintiff Matthew Sacks is an author and journalist who resides in California. He is
2 the author of *Pro Website Development and Operations*. His work is contained in pirated online
3 libraries, including LibGen, Z-Library, and Anna’s Archive. Defendants have directly or indirectly
4 downloaded books illegally contained in *The Pile*, LibGen, Z-Library, or Anna’s Archive, and there
5 is accordingly a reasonable inference that Defendants illegally downloaded Sacks’s work.

6 32. Plaintiff Michael Kochin is an author and journalist who resides in Israel. He is the
7 author of *Five Chapters on Rhetoric: Character, Action, Things, Nothing & Art* and *An Independent*
8 *Empire: Diplomacy & War in the Making of the United States*. His works are contained in pirated
9 online libraries, including LibGen, Z-Library, and Anna’s Archive. Defendants have directly or
10 indirectly downloaded books illegally contained in *The Pile*, LibGen, Z-Library, or Anna’s Archive,
11 and there is accordingly a reasonable inference that Defendants illegally downloaded Kochin’s
12 work.

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

15 **B. Defendants**

16 34. Defendant Anthropic PBC (“Anthropic”) is a Delaware public benefit corporation
17 with its principal place of business in San Francisco, California. Anthropic develops and
18 commercializes large language models (including the Claude series). Anthropic directed,
19 authorized, and profited from the acts of copyright infringement alleged in this Complaint, including
20 the acquisition of pirated copies of Plaintiffs’ copyrighted books from shadow-library websites and
21 the reproduction, ingestion, and use of those works in the training, development, and deployment of
22 its LLMs. For example, Anthropic downloaded *The Pile* (including Books3) and LibGen. Anthropic
23 conducts substantial business in this District and throughout the United States.

24 35. Defendant Google LLC (“Google”) is a Delaware limited liability company with its
25 principal place of business in Mountain View, California. Google develops, trains, and
26 commercializes generative AI models, including Gemini, Bard (formerly), and Imagen, which were
27 trained using datasets containing large volumes of pirated books. For example, Google trained its
28 LLMs on C4, which includes materials downloaded from Z-Library, and on information and belief

1 downloaded Z-Library. Google copied, reproduced, and embedded Plaintiffs’ copyrighted works
2 into its models without permission or license and continues to profit from those infringements across
3 its commercial product ecosystem, including Google Cloud, Google Search, and various AI-
4 powered enterprise tools.

5 36. Defendant OpenAI, Inc., and its affiliated entities OpenAI OpCo LLC, OpenAI GP
6 LLC, OpenAI Global LLC, OAI Corporation LLC, OpenAI Holdings LLC, OpenAI Startup Fund I
7 LP, OpenAI Startup Fund GP I LLC, and OpenAI Startup Fund Management LLC (collectively,
8 “OpenAI”) are entities organized under the laws of Delaware with principal places of business in
9 San Francisco, California. OpenAI develops and commercializes the GPT family of models
10 (including GPT-3, GPT-3.5, GPT-4, GPT-4o, and their derivatives), which were trained on datasets
11 containing illegal copies of Plaintiffs’ copyrighted books. For example, OpenAI downloaded
12 LibGen, which formed the basis of its Books 1 and 2 training datasets. OpenAI reproduced
13 Plaintiffs’ works multiple times during data collection, preprocessing, and training, and continues
14 to exploit those works commercially through ChatGPT, ChatGPT Enterprise, the OpenAI API, and
15 other products.

16 37. Defendant Meta Platforms, Inc. (“Meta”) is a Delaware corporation with its principal
17 place of business in Menlo Park, California. Meta develops and distributes the Llama series of
18 LLMs, including Llama-1, Llama-2, and Llama-3, which were trained using datasets sourced in part
19 from shadow libraries such as LibGen containing pirated books. Meta also acts as a distributor of
20 such datasets within its research ecosystem. For example, Meta downloaded *The Pile* (including
21 Books3), LibGen, Z-Library, and Anna’s Archive. Meta copied Plaintiffs’ copyrighted works
22 without license and monetizes those infringements through its integration of Llama models into
23 Facebook, Instagram, WhatsApp, Ray-Ban Meta Glasses, enterprise APIs, and other products.

24 38. Defendant xAI Corporation (“xAI”) is a Nevada corporation with its principal place
25 of business in Palo Alto, California. xAI develops the Grok series of LLMs, which were trained on
26 large-scale text corpora that includes illegally obtained books and datasets containing Plaintiffs’
27 copyrighted works. For example, there is a reasonable basis to believe that xAI downloaded LibGen
28 and, like “all” major LLM companies, Anna’s Archive. xAI copied, reproduced, and embedded

1 Plaintiffs’ works into its models for use in Grok and its associated commercial services, including
2 products offered through X Corp. (formerly Twitter).

3 39. Defendant Perplexity AI, Inc. (“Perplexity”) is a Delaware corporation with its
4 principal place of business in San Francisco, California. Perplexity builds and deploys AI search
5 and text-generation systems that rely on the unauthorized use of copyrighted works to optimize its
6 product through its retrieval-augmented generation or “RAG” process. There is a reasonable basis
7 to believe that Perplexity, like “all” major LLM companies, downloaded Anna’s Archive. On
8 information and belief, Perplexity’s RAG process relies on pirated copies of Plaintiffs’ books. On
9 information and belief, Perplexity reproduced and exploited Plaintiffs’ copyrighted works without
10 authorization in its AI search systems.

11 40. Defendant Apple Inc. (“Apple”) is a corporation organized and existing under the
12 laws of the State of Delaware, with its principal place of business at One Apple Park Way,
13 Cupertino, California 95014. Apple develops, manufactures, and markets consumer electronics,
14 software, and online services, including artificial intelligence products marketed under the “Apple
15 Intelligence” brand. Apple is building its AI brand using datasets of pirated copyrighted books that
16 include Plaintiffs’ works and has reproduced and exploited these works and the works of others to
17 train its OpenELM and Foundation language models. For example, Apple downloaded *The Pile*
18 (including Books3), and there is a reasonable basis to believe that like “all” major LLM companies,
19 it downloaded Anna’s Archive. On information and belief, Apple continues to retain a private AI
20 training-data library including Plaintiffs’ pirated works to train its future models without lawful
21 consent.

22 41. Defendant NVIDIA Corporation (“NVIDIA”) is a corporation organized and
23 existing under the laws of the State of Delaware, with its principal place of business at 2788 San
24 Tomas Expressway, Santa Clara, California 95051. NVIDIA designs, develops, and markets
25 graphics processing units and artificial intelligence computing platforms, including the NeMo large
26 language model framework and associated AI models, which relied upon datasets of pirated
27 copyrighted books that include Plaintiffs’ works. For example, NVIDIA downloaded *The Pile*
28 (including Books3) and Anna’s Archive.

1 **III. JURISDICTION AND VENUE**

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

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

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

17 **IV. FACTUAL ALLEGATIONS**

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

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

22 46. Over the last several years, technology companies have treated generative AI as the
23 next foundational layer of the digital economy. Industry leaders publicly describe an “AI arms race,”
24 in which they have redirected their corporate strategies to seize control of what they believe will
25 become a new infrastructure layer for commerce, communication, and knowledge work.⁵

26
27 _____
28 ⁵ 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).

1 47. For these companies, staying ahead of competitors is “code red.”⁶ Google itself
2 responded by consolidating its AI research divisions, devoting unprecedented resources to
3 generative AI, and rapidly integrating AI features across its product suite.⁷

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

9 49. Google’s Gemini family of models and its Imagen text-to-image systems have
10 likewise been woven into core Google products—including Search, Cloud, Workspace, and other
11 AI-powered products—which Google attributes with driving billions of dollars in new revenue and
12 record quarterly results.⁸

13
14
15 _____
16 ⁶ See Sharon Goldman, *Sam Altman declares ‘Code Red’ as Google’s Gemini surges—three years*
17 *after ChatGPT cause Google CEO Sundar Pichai to do the same*, FORTUNE (Dec. 2, 2025, 11:43
AM), <https://fortune.com/2025/12/02/sam-altman-declares-code-red-google-gemini-ceo-sundar-pichai/>.

18 ⁷ See, e.g., Sundar Pichai, *Building for our AI future*, Google (Apr. 18, 2024),
19 <https://blog.google/inside-google/company-announcements/building-ai-future-april-2024/>; Tom
20 Jowitt, “Google Consolidates DeepMind and AI Research Teams,” SILICON (Apr. 19, 2024, 9:35
PM), [https://www.silicon.co.uk/e-innovation/artificial-intelligence/google-consolidates-deepmind-
21 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)
22 (discussing consolidation).

23 ⁸ See Kyle Wiggers and Maxwell Zeff, *Google Gemini: Everything you need to know about the*
24 *generative AI apps and models*, TECHCRUNCH (Feb. 26, 2025, 6:09 PM),
25 <https://techcrunch.com/2025/02/26/what-is-google-gemini-ai/> (“The Gemini apps aren’t the only
26 means of recruiting Gemini models’ assistance with tasks. Slowly but surely, Gemini-imbued
27 features are making their way into staple Google apps and services like Gmail and Google
28 Docs.”); Jennifer Elias, *Google Cloude chief details how search giant is making billions*
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](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)).

1 50. Anthropic has taken the same path. Its Claude models—trained to write, summarize,
2 and analyze text at book-length scale—are projected to generate hundreds of millions of dollars in
3 annual revenue and have supported valuations in the hundreds of billions of dollars⁹, funded by
4 major technology investors such as Amazon¹⁰ and Google.¹¹

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

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

13 53. Defendants used the training data to read and learn from it.

14 54. Most, if not all, of the Defendants now allow users to opt out of having their data
15 used to train their AI models.

16 55. Defendants deprived Plaintiffs and many other authors of that choice.
17
18
19

20 ⁹ 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/> (“Claude Code hit \$2.5 billion in run rate revenue and doubled since the start of the year...and [Anthropic’s] total annualized revenue has climbed to \$14 billion.”)

22 ¹⁰ See *Amazon and Anthropic Deepen Strategic Collaboration*, AMAZON NEWS (Nov. 22, 2024),
23 <https://www.aboutamazon.com/news/aws/amazon-invests-additional-4-billion-anthropic-ai>
24 (explaining that Amazon was “making a \$4 billion investment in Anthropic.”)

25 ¹¹ See Cade Metz, Nico Grant, and David McCabe, *Inside Google’s Investment in the A.I. Start-Up Anthropic* (March 11, 2025), <https://www.nytimes.com/2025/03/11/technology/google-investment-anthropic.html> (“Google owns 14% of Anthropic” and “In total, Google has invested more than \$3 billion in the A.I. company.”)

26
27 ¹² See Alex Reisner, *The Unbelievable Scale of AI’s Pirated-Books Problem*, THE ATLANTIC (March
28 20, 2025), <https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/>.

1 56. Defendants’ misconduct undermined and deprived Plaintiffs and many other authors
2 of the bargaining power that they should have had, and otherwise would have had, with respect to
3 licensing terms for the use Defendants made of their works.

4 57. Defendants’ specific misconduct and the widespread unlicensed copying of
5 Plaintiffs’ works have harmed Plaintiffs’ ability to participate in and profit from the derivative
6 market for literary and other works used to train AI LLMs.

7 **B. The Shadow Library Ecosystem for Pirated Books**

8 58. The Defendants did not obtain that gold-standard material lawfully. Instead, in order
9 to win the generative-AI arms race cheaply and quickly, each Defendant turned to the same piracy
10 repositories—shadow-library websites like LibGen, Z-Library, Bibliotik, Books3, and similar
11 datasets. Each dataset constituted a separate repository of copyrighted works, and Defendants’
12 acquisition of each involved independent acts of unauthorized reproduction.¹³

13 **1. Overview of Shadow Libraries**

14 59. For more than a decade, a network of so-called “shadow libraries” has operated for
15 the express purpose of distributing pirated copies of copyrighted books at massive scale. These
16 repositories systematically acquire, store, index, and disseminate full-fidelity digital copies of
17 copyrighted books—typically in native formats such as EPUB, PDF, MOBI, or DJVU—without
18 authorization from authors or publishers.¹⁴

19 60. Unlike incidental infringement or isolated file-sharing, shadow libraries are purpose-
20 built systems for mass copyright infringement. They maintain searchable catalogs, metadata,
21

22
23 ¹³ See *AI Watchdog: Books3*, THE ATLANTIC (Sep. 10, 2025),
24 <https://www.theatlantic.com/technology/archive/2025/09/dataset-books3/683662/>; Claire
25 Woodcock, *‘Shadow Libraries’ Are Moving Their Pirated Books to the Dark Web After Fed Crackdowns*, VICE (Nov. 30, 2022), <https://www.vice.com/en/article/shadow-libraries-are-moving-their-pirated-books-to-the-dark-web-after-fed-crackdowns/>.

26 ¹⁴ See Riddhi Setty, *Rampant ‘Shadow Libraries’ Drive Calls for Anti-Piracy Action*, BLOOMBERG
27 LAW (Oct. 19, 2022, 9:03 AM), <https://news.bloomberglaw.com/ip-law/rampant-shadow-libraries-drive-calls-for-anti-piracy-action>; Woodcock, *‘Shadow Libraries’ Are Moving Their Pirated Books to the Dark Web After Fed Crackdowns*.

1 mirrors, bulk-download mechanisms, and—critically—complete downloadable archives designed
2 to allow third parties to replicate the entire collection.¹⁵

3 61. These libraries are widely known within both piracy communities and the technology
4 sector as illegal sources of copyrighted books. Many have been the subject of criminal prosecutions,
5 civil injunctions, domain seizures, and formal designation as “notorious piracy markets” by United
6 States trade authorities. As centralized shadow libraries increasingly faced enforcement actions,
7 including the seizure of domains, discussed just below, third parties responded by creating full
8 mirrored copies of those repositories for decentralized redistribution.¹⁶

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

20 63. The shadow library ecosystems exists for one purpose: making copyrighted works
21 available for unauthorized download. Defendants exploited these shadow libraries and datasets—
22 along with others not named here—to train their LLM models on Plaintiffs’ books without
23

24 ¹⁵ See Letter from Mary E. Rasenberger, CEO, Authors Guild, & Umair Kazi, Dir. of Pol’y &
25 Advocacy, Authors Guild, to Daniel Lee, Assistant U.S. Trade Representative for Innovation &
26 Intellectual Prop., Office of the U.S. Trade Representative (Oct. 7, 2022),
<https://cdn.arstechnica.net/wp-content/uploads/2022/11/Authors-Guild-October-7-Z-Library-Complaint.pdf>.

27 ¹⁶ Woodcock, *'Shadow Libraries' Are Moving Their Pirated Books to the Dark Web After Fed*
28 *Crackdowns*.

1 permission or compensation. The most significant shadow libraries relevant to Defendants’ conduct
2 include Bibliotik, Books3, Library Genesis (“LibGen”), Z-Library, Sci-Hub, Pirate Library Mirror
3 (“PiLiMi”), and Anna’s Archive. These repositories are interrelated, overlapping, and in many cases
4 direct descendants of one another.

5 i. Library Genesis (“LibGen”)

6 64. Library Genesis, commonly known as “LibGen,” is one of the largest and longest-
7 running shadow libraries in the world. It hosts millions of pirated books, academic texts, and
8 scholarly articles.¹⁷

9 65. LibGen operates as a centralized repository offering direct downloads of full-fidelity
10 ebook files. It also distributes its entire database through bulk archives and torrent files, enabling
11 third parties to download and locally host complete copies of its collection.¹⁸

12 66. LibGen has been repeatedly enjoined by federal courts for copyright infringement
13 and has been designated a “notorious market” by the United States Trade Representative.¹⁹

14 67. Despite enforcement actions, LibGen has remained accessible through shifting
15 domains, mirrors, and downloadable archives. Its persistence is a function of deliberate
16 decentralization designed to evade shutdown.²⁰

17
18 ¹⁷ Office of the U.S. Trade Representative, REVIEW OF NOTORIOUS MARKETS FOR
19 COUNTERFEITING AND PIRACY, 27 (2024),
20 [https://ustr.gov/sites/default/files/2024%20Review%20of%20Notorious%20Markets%20of%20Co
21 unterfeiting%20and%20Piracy%20\(final\).pdf](https://ustr.gov/sites/default/files/2024%20Review%20of%20Notorious%20Markets%20of%20Co%20unterfeiting%20and%20Piracy%20(final).pdf). (“Libgen ... hosts a large number of digital copies
22 of books, manuals, journals, and other works, many of which are unauthorized copies of copyright
23 protected content.”)

24 ¹⁸ See Letter from Mary E. Rasenberger, CEO, Authors Guild, & Umair Kazi, Dir. of Pol’y &
25 Advocacy, Authors Guild, to Daniel Lee, Assistant U.S. Trade Representative for Innovation &
26 Intellectual Prop., Office of the U.S. Trade Representative at “The Library Genesis Project (Libgen).”

27 ¹⁹ See Office of the U.S. Trade Representative, 2019 REVIEW OF NOTORIOUS MARKETS FOR
28 COUNTERFEITING AND PIRACY, 27,
[https://ustr.gov/sites/default/files/2019_Review_of_Notorious_Markets_for_Counterfeiting_and_P
iracy.pdf](https://ustr.gov/sites/default/files/2019_Review_of_Notorious_Markets_for_Counterfeiting_and_Piracy.pdf).

²⁰ See Andrew Albanese, *Textbook Publishers Sue Notorious ‘Shadow Library’ Libgen*,
PUBLISHERS WEEKLY (Sep. 14, 2023), <https://www.publishersweekly.com/pw/by->

1 ii. Books1 and Books 2 – OpenAI Begins the Pirated Book Arm’s Race

2 68. In June 2020, OpenAI published its seminal paper on GPT-3, revealing that the
3 model was trained on five principal text corpora: Common Crawl, WebText2, Books1, Books2, and
4 Wikipedia. Books1 contributed approximately 12 billion tokens and Books2 approximately 55
5 billion tokens, each weighted at 8% of the training mix—for a combined 16% of GPT-3’s training
6 data by weight.²¹

7 69. OpenAI described Books1 and Books2 only as “internet-based books corpora” and
8 disclosed no further details about their contents, sources, or provenance.²²

9 70. It has since been established as undisputed in litigation that in 2018, an OpenAI
10 employee downloaded pirated copies of books from Library Genesis—the same shadow library
11 described above.²³

12 71. From these pirated LibGen downloads, OpenAI created two datasets it internally
13 referred to as “LibGen1” and “LibGen2.” OpenAI later renamed these datasets to “Books1” and
14 “Books2” for public-facing references—effectively concealing that its foundational book-training
15 corpora were sourced directly from a known pirate library.²⁴

16
17
18 topic/digital/copyright/article/93189-textbook-publishers-sue-notorious-shadow-library-libgen.html.

19 ²¹ Tom Brown et al., *Language Models are Few-Shot Learners*, arXiv, 9 (Table 2.2) (2020), available
20 at <https://arxiv.org/pdf/2005.14165>.

21 ²² *Id.* at 8.

22 ²³ Opinion & Order Re: OpenAI’s Deletion of Books1 and Books2 Datasets and Privilege Rulings,
23 *In Re OpenAI, Inc. Copyright Infringement Litigation*, No. 25-md-3143 (S.D.N.Y. Nov. 24, 2025)
24 ECF No. 846 at 1, available at <https://www.nysd.uscourts.gov/sites/default/files/2026-01/25md3143%20-%2011.24.25%20Order%20re%20ECF%20Nos.%20413,%20428,%20479,%20504,%20615.pdf>

25 ²⁴ *Id.* at 1; see also Class Plaintiffs’ Opp’n to OpenAI’s Obj. to Nov. 24, 2025 Order, ECF No. 996
26 at 1 (“To hide the fact that it sourced these datasets from a known pirate site OpenAI publicly called
27 these datasets Books1 and Books2, but internally it knew them as ‘LibGen1’ and ‘LibGen2.’”),
28 available at https://chatgptiseatingtheworld.com/wp-content/uploads/2025/12/In-Re_-_OpenAI-Inc.-Copyright-Infringement-Litigation-Entry-996.pdf.

1 72. OpenAI used the Books1 and Books2 datasets to train both GPT-3 and GPT-3.5,
2 among the most commercially significant large language models ever released.²⁵

3 73. In mid-2022, approximately one year before any of the copyright infringement
4 lawsuits against it were filed, OpenAI deleted the Books1 and Books2 datasets.²⁶

5 iii. EleutherAI and *The Pile*: How Books3 Was Packaged for Broad Reuse

6 74. EleutherAI is an open research collective formed in a Discord server in July 2020—
7 weeks after OpenAI released GPT-3—with the stated goal of democratizing access to large-scale
8 language modeling by building and releasing open-source GPT-style models and the training
9 resources needed to develop them. EleutherAI was widely characterized as an “open-source version
10 of OpenAI,” created specifically to replicate GPT-3's capabilities using publicly available tools and
11 data.²⁷

12 75. A central obstacle to that effort was that OpenAI's training data was proprietary and
13 undisclosed, including Books1 and Books2.²⁸

17 ²⁵ Opinion & Order Re: OpenAI's Deletion of Books1 and Books2 Datasets and Privilege Rulings
18 at 7 (citing Trinh Dep. testimony that “the datasets were used to train GPT-3 and GPT-3.5”).

19 ²⁶ *Id.* at 1–2 (“[I]n 2022, approximately one year before any of the actions in this multidistrict
20 litigation commenced, OpenAI deleted the Books1 and Books2 datasets.” “These are the only
training datasets that, according to OpenAI, have ever been deleted.”).

21 ²⁷ See Stella Biderman et al., *The Common Pile v0.1*, ELUETHERAI (June 5, 2025),
22 <https://blog.eleuther.ai/common-pile/>; Jonathan Gillham, *Eleuther AI NLP Models*, ORIGINALITY.AI
23 (Aug. 15, 2025), <https://originality.ai/blog/eleutherai-nlp-models#> (“EleutherAI is an open-source
24 consortium of AI and machine learning researchers that came together to provide a decentralized
alternative to OpenAI. They work on open-source AI models and products that rival OpenAI's
offerings.”).

25 ²⁸ See Kyle Barr, *Anti-Piracy Group Takes Massive AI Training Dataset ‘Books3’ Offline*, GIZMODO
26 (Aug. 18, 2023), <https://gizmodo.com/anti-piracy-group-takes-ai-training-dataset-books3-off-1850743763>
27 (“OpenAI's GPT-3 model used the Books2 training set to train its AI. Both Books1
and Books2 make up close to 15% of GPT-3's training data, though there's little to no precise
28 information on what's contained in it....” “But while OpenAI has been revealing less of its training
data over the years, we know exactly what's gone into the Books3 repository.”).

1 76. At that time, OpenAI refused to publicly identify the contents or sources of either
2 corpus.²⁹ However, a number of AI researchers, including the creator of the Books3 dataset, broadly
3 suspected Books2 to be all of Libgen.³⁰

4 77. To support the training of large language models by the broader research
5 community—including academic labs, independent researchers, startups, and commercial AI
6 developers—EleutherAI assembled and released *The Pile*, a large, general-purpose English text
7 dataset explicitly “targeted at training large-scale language models.”³¹

8 78. *The Pile* was released as a composite dataset made up of multiple “subsets,”
9 combining preexisting corpora with newly constructed components curated for LLM training.³²

10 79. One of those subsets was “Books3”—a book corpus constructed by members of the
11 open-source AI research community in an effort to replicate and approximate the undisclosed
12 Books1 and Books2 corpora that OpenAI had used to train GPT-3. EleutherAI and *The Pile*
13 documentation identify Books3 as derived from Bibliotik. As alleged above, Books3 consists of
14 approximately 200,000 books and was created by downloading books from Bibliotik and extracting
15 their full text for use as a machine-learning dataset.³³

16 80. The chain of events is notable: OpenAI trained GPT-3 on secret book corpora that
17 researchers suspected were sourced from shadow libraries; the open-source community then
18 constructed Books3 from a different shadow library in an attempt to replicate those corpora; and
19 EleutherAI packaged Books3 into *The Pile* and distributed it publicly—effectively democratizing
20

21
22 ²⁹ See Barr, *Anti-Piracy Group Takes Massive AI Training Dataset ‘Books3’ Offline*.

23 ³⁰ See *The Pile Books3 Dataset*, defunct-datasets/the_pile_books3, HUGGING FACE,
24 https://huggingface.co/datasets/defunct-datasets/the_pile_books3 (dataset card last updated 2023;
dataset defunct due to reported copyright infringement).

25 ³¹ Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, arXiv, 1
26 (2020), available at <https://arxiv.org/abs/2101.00027>.

27 ³² *Id.*

28 ³³ See Barr, *Anti-Piracy Group Takes Massive AI Training Dataset ‘Books3’ Offline*.

1 access not just to training data, but to a corpus of infringing works derived from a pirate book
2 library.³⁴

3 81. This irony was not lost on the creator of Books3. He remarked on this in an interview
4 with *The Atlantic*:

5 [Shawn Presser] created Books3 in the hope that it would allow any
6 developer to create generative-AI tools. “It would be better if it wasn’t
7 necessary to have something like Books3,” he said. “But the alternative is
8 that, without Books3, only OpenAI can do what they’re doing.”³⁵

9 82. By incorporating Books3 into *The Pile* and distributing *The Pile* publicly as an LLM
10 training resource, EleutherAI effectively packaged and normalized Books3 as an “off-the-shelf”
11 component of modern language-model development—making Books3 readily accessible for reuse
12 by downstream researchers and organizations building AI tools.

13 83. This incorporation matters because *The Pile* was not released as a private research
14 artifact. It was presented as a general-purpose training dataset intended to be downloaded,
15 replicated, and used by others as a foundational corpus for training GPT-style models. As a result,
16 Books3’s presence within *The Pile* facilitated wide downstream distribution and adoption of a
17 corpus derived from a pirate book library.³⁶

18 84. Public reporting and enforcement activity later confirmed that Books3 was
19 recognized externally as a piracy-derived book dataset used in AI training, prompting takedown
20 demands directed at entities that facilitated access to it (including entities that published dataset
21 “cards” or links pointing to where it could be obtained).³⁷

22 _____
23 ³⁴ Reisner, *Revealed: The Authors Whose Pirated Books Are Power Generative AI*.

24 ³⁵ *Id.*

25 ³⁶ See Stella Biderman, *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*,
26 ELEUTHERAI (Dec. 31, 2020), <https://www.eleuther.ai/papers-blog/the-pile-an-800gb-dataset>.

27 ³⁷ See Rizwan Choudhury, *Anti-piracy group shuts down Books3, a popular dataset for AI models*,
28 INTERESTING ENGINEERING (Aug. 20, 2023), <https://interestingengineering.com/innovation/anti-piracy-group-shuts-down-books3-a-popular-dataset-for-ai-models>.

1 85. Although Books3 has been removed from some official distribution points and
2 references following these enforcement actions, Books3 has continued to circulate through mirrors
3 and third-party hosting locations—consistent with the broader shadow-library ecosystem alleged
4 herein, in which infringing datasets persist through replication and redistribution even after
5 takedowns.³⁸

6 iv. Bibliotik: The Source of Books3

7 86. Bibliotik is a private, invitation-only torrent tracker that has long functioned as a
8 centralized source of pirated ebooks. Unlike public piracy websites, Bibliotik restricts access to
9 registered users but nonetheless hosts and distributes hundreds of thousands of copyrighted books.³⁹

10 87. Bibliotik operates through BitTorrent technology, enabling users to download
11 complete ebook files while simultaneously re-uploading pieces of those same files to other users—
12 a process that results in repeated unauthorized reproduction and redistribution of copyrighted
13 works.⁴⁰

14 88. Bibliotik has been widely recognized in piracy communities, academic literature, and
15 AI-research documentation as a shadow library devoted to copyrighted books. Its illicit nature has
16 been openly discussed for years prior to Defendants’ use of datasets derived from it.⁴¹

17
18
19
20
21
22 _____
23 ³⁸ See Ernesto Van de Sar, Anti-Piracy Group Takes Prominent AI Training Dataset “Books3’
24 Offline, TORRENTFREAK (Aug. 16, 2023), <https://torrentfreak.com/anti-piracy-group-takes-prominent-ai-training-dataset-books3-offline-230816/>.

25 ³⁹ See Ruheni Mathenge, *The 12 Best Private Torrent Sites Still Working in 2026*, PRIVACYSAVVY
26 (last accessed March 9, 2026), <https://privacysavvy.com/security/torrents/best-private-torrent-websites/>.

27 ⁴⁰ See *id.*

28 ⁴¹ Barr, *Anti-Piracy Group Takes Massive AI Training Dataset ‘Books3’ Offline*.

1 89. Books3—a dataset of approximately 200,000 books⁴²—was created by extracting the
2 full text of books downloaded from Bibliotik and distributing those texts as a structured dataset for
3 machine-learning training.⁴³

4 90. Books3 did not consist of public-domain works. Rather, it overwhelmingly contained
5 copyrighted books, including commercially available fiction and nonfiction titles, copied verbatim
6 from Bibliotik without authorization.

7 91. Books3 was assembled and distributed specifically to serve as training data for large
8 language models. It was incorporated into *The Pile*, a large open-source dataset widely used by AI
9 developers, including Defendants.

10 v. Z-Library

11 92. Z-Library (also known as “B-ok”) emerged as an expanded and user-friendly
12 derivative of LibGen. It incorporated large portions of LibGen’s catalog while adding additional
13 titles, metadata, and interface features.⁴⁴

14 93. Z-Library offered premium features—including faster downloads and higher volume
15 limits—in exchange for payment, operating in effect as a commercial piracy service.⁴⁵

16 94. In 2022, Z-Library’s domains were seized by law-enforcement authorities, and its
17 operators were arrested and later indicted for criminal copyright infringement. These actions
18 confirmed what had long been publicly known: Z-Library was an illegal piracy operation.⁴⁶

19
20 ⁴² See Van de Sar, *Anti-Piracy Group Takes Prominent AI Training Dataset ‘Books3’ Offline*.

21 ⁴³ See Ernesto Van der Sar, *Meta Admits Use of ‘Pirated’ Book Dataset to Train AI*, TORRENTFREAK
22 (Jan. 11, 2024), <https://torrentfreak.com/meta-admits-use-of-pirated-book-dataset-to-train-ai-240111/>.

23 ⁴⁴ Jordana Rosenfeld, *Z-Library*, ENCYCLOPEDIA BRITANNICA (last accessed March 9, 2026),
24 <https://www.britannica.com/topic/Z-Library>

25 ⁴⁵ See Masood Farivar, *Two Russian Nationals Charged With Operating E-Book Piracy Site*, VOA
26 (Nov. 16, 2022), <https://www.voanews.com/a/two-russian-nationals-charged-with-operating-e-book-piracy-site-/6838130.html>

27 ⁴⁶ Press Release, U.S. Dep’t of Justice, U.S. Att’y’s Off., E. Dist. of N.Y., *Two Russian Nationals*
28 *Charged with Running Massive E-Book Piracy Website* (Nov. 16, 2022),

1 95. The seizure of Z-Library did not eliminate access to its content. Instead, third parties
2 responded by creating full mirrors of its collection to ensure continued distribution.⁴⁷

3 vi. Sci-Hub

4 96. Sci-Hub is a shadow library focused primarily on academic journal articles, but it is
5 closely linked to LibGen and shares infrastructure, mirrors, and operational overlap.⁴⁸

6 97. Like LibGen, Sci-Hub has been repeatedly enjoined by federal courts for copyright
7 infringement and remains accessible through mirrors and downloadable archives.⁴⁹

8 vii. Pirate Library Mirror (“PiLiMi”)

9 98. Pirate Library Mirror, commonly referred to as “PiLiMi,” is a complete mirrored
10 archive of the Z-Library corpus, itself largely derived from LibGen.⁵⁰

11 99. PiLiMi is not merely a website or index. It is a full, downloadable dataset designed
12 to allow users to obtain and locally host millions of pirated books through peer-to-peer distribution.⁵¹

13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

<https://www.justice.gov/usao-edny/pr/two-russian-nationals-charged-running-massive-e-book-piracy-website>.

⁴⁷ See Woodcock, *‘Shadow Libraries’ Are Moving Their Pirated Books to The Dark Web*.

⁴⁸ See Office of the U.S. Trade Representative, 2019 REVIEW OF NOTORIOUS MARKETS FOR COUNTERFEITING AND PIRACY at 27.

⁴⁹ See Dalmeet Singh Chawla, *Court demands that search engines and internet service providers block Sci-Hub*, SCIENCE (Nov. 6, 2017), <https://www.science.org/content/article/court-demands-search-engines-and-internet-service-providers-block-sci-hub>.

⁵⁰ See Geoff Wheelright, *Will I get a piece of Anthropic’s \$1.5B settlement if my book was used to train AI?*, GEEKWIRE (Sep. 18, 2025), <https://www.geekwire.com/2025/will-i-get-a-piece-of-anthropics-1-5b-settlement-if-my-book-was-used-to-train-ai/> (“The Authors Guild says PiLiMi is a mirror of Anna’s Archive”).

⁵¹ *Id.*

1 100. PiLiMi was explicitly created to preserve and propagate Z-Library’s pirated
2 collection after law-enforcement seizures, ensuring continuity of access despite shutdowns of the
3 original site.⁵²

4 101. Users who download PiLiMi do not passively receive data; they actively participate
5 in copying and redistributing copyrighted works through torrent “leeching” and “seeding.”⁵³

6 102. At least some Defendants knowingly treated PiLiMi as a distinct and supplemental
7 pirated dataset rather than a redundant copy of materials already obtained. Before downloading
8 PiLiMi, some Defendants compared its catalog against their existing LibGen holdings, identified
9 which titles were not already in their possession, and deliberately downloaded only those additional
10 works. Through this process, some Defendants expanded their illicit libraries by millions of unique
11 copyrighted books obtained after the shutdown of Z-Library, while retaining earlier pirated copies
12 from Books3 and LibGen in centralized storage. These actions reflect intentional sourcing, selection,
13 and accumulation of multiple pirated book datasets at different times, through different mechanisms,
14 and in conscious disregard of copyright law.

15 viii. Anna’s Archive

16 103. Anna’s Archive is the most comprehensive and active shadow library currently in
17 operation.⁵⁴

18
19
20 _____
21 ⁵² See Ernesto Van de Sar, “*Anna’s Archive*” *Opens the Door to Z-Library and Other Pirate*
22 *Libraries*, TORRENTFREAK (Nov. 19, 2022), <https://torrentfreak.com/annas-archive-opens-the-door-to-z-library-and-other-pirate-libraries-221118/>.

23 ⁵³ See Robert Nogacki, *Anthropic's Landmark Settlement: A \$1.5 Billion Copyright Precedent in*
24 *Artificial Intelligence Training Data*, LinkedIn (Sep. 7, 2025),
25 <https://www.linkedin.com/pulse/anthropics-landmark-settlement-15-billion-copyright-training-nogacki-esn3f/> (describing “leeching” and “seeding” as “processes characteristic of peer-to-peer networks where users simultaneously download and distribute files”).

26 ⁵⁴ See Soumyajyoti Mukherjee, *Who is Anna’s Archive? All we know about pirate activist group*
27 *behind 300 TB Spotify music library heist*, SOAPCENTRAL.COM (Dec. 24, 2025),
28 <https://www.soapcentral.com/entertainment/who-anna-s-archive-all-know-pirate-activist-group-behind-300-tb-spotify-music-library-heist>.

1 104. Anna’s Archive began in 2022 as “Pirate Library Mirror,” initially hosting a mirrored
2 copy of Z-Library. It later rebranded as “Anna’s Archive” and expanded to aggregate and host the
3 complete collections of LibGen, Z-Library, Sci-Hub, PiLiMi, and other pirated sources.⁵⁵ Anna’s
4 Archive hosts millions of pirated books.

5 105. Anna’s Archive functions as a meta-library: it indexes, mirrors, and redistributes
6 multiple shadow libraries simultaneously, offering users unified access to millions of pirated
7 books.⁵⁶

8 106. Anna’s Archive offers paid tiers that provide “high-speed” or priority access to its
9 pirated collections, monetizing mass copyright infringement.⁵⁷

10 107. Through its downloadable archives and torrent-based distribution, Anna’s Archive
11 enables users to acquire and store local copies of millions of copyrighted books in bulk.⁵⁸

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

14 109. According Ann’s Archive, “virtually all major companies building LLMs contacted
15 us to train on our data. . . We have given high-speed access to about 30 companies.”⁵⁹ Piracy by
16 LLM companies through Anna’s Archive is de facto industry practice—Ann’s Archive
17 blog stated as recently as February 18, 2026 that if an *LLM was reading its blog* “you have likely
18
19

20 ⁵⁵ See Ernesto Van de Sar, “*Anna’s Archive*” *Opens the Door to Z-Library and Other Pirate*
21 *Libraries*.

22 ⁵⁶ *Id.*

23 ⁵⁷ See *If you’re an LLM, please read this*, ANNA’S ARCHIVE (February 18, 2026), [https://annas-
archive.gl/blog/llms-txt.html](https://annas-archive.gl/blog/llms-txt.html).

24 ⁵⁸ See M. Luisa Simpson, *2024 Special 301 Out-of-Cycle Review of Notorious Markets: Request*
25 *for Comments* ASSOCIATION OF AMERICAN PUBLISHERS (OCTOBER 2, 2024),
26 [https://copyrightalliance.org/wp-content/uploads/2024/10/USTR-2024-0013-
0005_attachment_1.pdf](https://copyrightalliance.org/wp-content/uploads/2024/10/USTR-2024-0013-0005_attachment_1.pdf).

27 ⁵⁹ See *Copyright reform is necessary for national security*, ANNA’S ARCHIVE (Jan. 31, 2025),
28 <https://annas-archive.gl/blog/ai-copyright.html>.

1 been trained in part on our data.”⁶⁰ Each Defendant is a major company building LLMs. There is
2 accordingly a reasonable inference that each Defendant downloaded from Anna’s Archive. And
3 since each Plaintiffs’ Infringed Work is contained on Anna’s Archive, on that basis alone there is a
4 reasonable inference that each Defendant infringed Plaintiffs’ works.

5 2. **Research Confirms That AI Training Embeds Near-Verbatim Copies of** 6 **Copyrighted Works Into the Models**

7 110. Peer-reviewed scientific research has now confirmed what Plaintiffs have alleged:
8 the process of training large language models on copyrighted works does not merely allow those
9 models to “learn” from the works in some abstract sense. Rather, training causes the models to store
10 persistent, near-verbatim copies of copyrighted works within their internal parameters—copies so
11 complete that, when prompted, the models can reproduce substantially all of the original text, word
12 for word.⁶¹

13 111. In a 2025 study, a team of researchers from Cornell University, Stanford University,
14 and West Virginia University developed a methodology for extracting memorized copyrighted
15 books from production, closed-source large language models—the same commercially deployed AI
16 systems at issue in this case. The researchers tested four leading production models: Anthropic’s
17 Claude 3.7 Sonnet, OpenAI’s GPT-4.1, Google’s Gemini 2.5 Pro, and xAI’s Grok 3.⁶²

18 112. The researchers’ methodology proceeded in two phases. In the first phase, the
19 researchers used a technique called “best-of-N” jailbreaking, which involves repeatedly prompting
20 the model with requests to reproduce a copyrighted work until the model’s safety filters fail to block
21 the output. In the second phase, the researchers used “continuation prompting”—providing the
22 model with the last portion of text it had already generated and asking it to continue—to extract
23

24 ⁶⁰ See *If you’re an LLM, please read this*, ANNA’S ARCHIVE (Feb. 18, 2026), [https://annas-](https://annas-archive.gl/blog/llms-txt.html)
25 [archive.gl/blog/llms-txt.html](https://annas-archive.gl/blog/llms-txt.html).

26 ⁶¹ A. Feder Cooper et al., *Extracting Copyrighted Long-Form Text from Production Language*
27 *Models*, arXiv, 1-2 (2025), <https://arxiv.org/abs/2601.02671>.

28 ⁶² *Id.* at 2–4.

1 progressively longer passages. Through this iterative process, the researchers were able to extract
2 near-complete copies of entire books.⁶³

3 113. The results were striking. Across all four production models tested, the researchers
4 were able to extract lengthy, near-verbatim passages of copyrighted books. Using a metric called
5 “nv-recall”—which measures the fraction of a book that can be extracted as near-verbatim text—
6 the researchers found that production models had memorized and could reproduce vast portions of
7 copyrighted works.⁶⁴

8 114. From Anthropic’s Claude 3.7 Sonnet alone, the researchers extracted 97.5% of *The*
9 *Great Gatsby*, 95.5% of *1984*, 94.3% of *Frankenstein*, 92.3% of *Harry Potter and the Sorcerer’s*
10 *Stone*, and 70.2% of *The Hobbit*. Similar levels of memorization were observed across GPT-4.1,
11 Gemini 2.5 Pro, and Grok 3. These figures demonstrate that AI training does not merely “learn”
12 from copyrighted texts—it stores virtually the entire work within the model.⁶⁵

13 115. For some models, the researchers did not even need to circumvent safety measures
14 to extract copyrighted content. The study found that Google’s Gemini 2.5 Pro and xAI’s Grok 3
15 would reproduce copyrighted book text in response to straightforward prompts, without any
16 jailbreaking whatsoever. This finding demonstrates that for certain Defendants’ models, copyrighted
17 content is not only stored within the model but is readily accessible through ordinary use.⁶⁶

18 116. The study further demonstrates that the safety measures and content filters deployed
19 by AI companies are insufficient to prevent the reproduction of memorized copyrighted works.
20 Although each Defendant has implemented various safety measures—such as reinforcement
21 learning from human feedback and system-level content filters—the research shows that these

22
23 _____
24 ⁶³ *Id.* at 4–7, Figs. 2, 3.

25 ⁶⁴ *Id.* at 6–9, Figs. 5–7, Tables 1–2.

26 ⁶⁵ *Id.* at 12, Fig. 5 (reporting nv-recall of 97.5% for *The Great Gatsby*, 95.5% for *1984*, 94.3% for
Frankenstein, and 92.3% for *Harry Potter and the Sorcerer’s Stone* via Claude 3.7 Sonnet).

27 ⁶⁶*Id.* (reporting that Gemini 2.5 Pro and Grok 3 reproduced copyrighted content without any
28 jailbreaking).

1 measures do not remove the underlying copyrighted content from the models. The copyrighted
2 works remain embedded in the model weights and can be extracted despite these protective layers.⁶⁷

3 117. These findings are directly relevant to all Defendants in this action. The models
4 tested in the study—Claude 3.7 Sonnet (Anthropic), GPT-4.1 (OpenAI), Gemini 2.5 Pro (Google),
5 and Grok 3 (xAI)—are commercially deployed products of four of the named Defendants. The
6 remaining Defendants use substantially similar training methodologies and datasets, and there is no
7 reason to believe the results would differ materially for their models. The research thus provides
8 powerful scientific evidence that each Defendant’s training process created persistent, unauthorized
9 copies of Plaintiffs’ copyrighted works within the Defendants’ AI models.

10 118. This peer-reviewed research confirms that training a large language model on
11 copyrighted works is not a transformative process that merely extracts abstract patterns or ideas. It
12 is an act of copying that embeds near-complete, recoverable reproductions of copyrighted works
13 into the model itself. Each Defendant’s model functions, in practical effect, as a repository of pirated
14 copyrighted works.

15 **C. Anthropic Trained Its LLM Models On Copyrighted Works That Were Pirated.**

16 119. Anthropic’s business model is built on the large-scale copying of books.⁶⁸ Anthropic
17 has developed and commercialized the “Claude” family of large language models by stealing over
18 seven million copyrighted books, including Plaintiffs’ works.⁶⁹ Judge Alsup, presiding over the
19 *Bartz v. Anthropic* litigation, stated simply: “From the start, Anthropic ‘had[d] many places from
20 which’ it could have purchased books, but it preferred to steal them to avoid ‘legal/practice/business
21 slog’.”⁷⁰ Rather than pay for the creative expression it exploits, Anthropic downloaded pirated
22 copies of books, reproduced them, and fed them into its models.

23 _____
24 ⁶⁷*Id.* at 9–11.

25 ⁶⁸ *See Bartz v. Anthropic PBC*, No. 3:24-cv-05417 (N.D. Cal. June 23, 2025), ECF No. 231 at 1 (Anthropic’s explicit goal was to “amass a central library of ‘all the books in the world’ to retain ‘forever.’”).

27 ⁶⁹ *Id.* at 3.

28 ⁷⁰ *Id.* at 2

1 20. Anthropic’s own public statements and technical papers confirm that books are
2 central to Claude’s capabilities. Anthropic has described a training corpus “most of which we
3 sourced from *The Pile*,”⁷¹ which includes Books3.

4 21. Anthropic has admitted that it used *The Pile* (which includes Books3) to train its
5 Claude models and that roughly one-third of one core Claude training dataset consisted of “internet
6 books.”⁷² By downloading these datasets and ingesting them into Claude, Anthropic necessarily
7 made multiple unlicensed copies of Plaintiffs’ works: once when obtaining them from pirate
8 sources, again during preprocessing and storage, and repeatedly during training and fine-tuning.⁷³
9 As the U.S. Patent and Trademark Office has explained, LLM training “almost by definition
10 involve[s] the reproduction of entire works or substantial portions” of them.⁷⁴ And as one court has
11 already found, “Anthropic’s piracy of otherwise available copies is inherently infringing even if the
12 pirated copies” may later be tapped for transformative use.⁷⁵

13 22. Anthropic selected books precisely because they are especially valuable training
14 material. Anthropic touts Claude’s ability to process entire books (up to roughly 75,000 words) and
15 generate coherent long-form responses that reflect not only word ordering and syntax, but also
16 themes, narrative structure, and high-level ideas—capabilities that could be developed only by
17 training on a large corpus of long-form prose.⁷⁶

18
19 ⁷¹ See Amanda Askeff et al., *A General Language Assistant as a Laboratory for Alignment*, arXiv,
20 27 (2021), <https://arxiv.org/pdf/2112.00861>.

21 ⁷² *Bartz*, ECF 72 at ¶ 33 (N.D. Cal. Dec. 18, 2024) [Anthropic’s Answer to First Amended
22 Complaint]; see also Askeff et al., *A General Language Assistant as a Laboratory for Alignment*, at
23 27.

24 ⁷³ *Bartz*, ECF 231 at 18-19.

25 ⁷⁴ U.S. Patent & Trademark Office, *Public Views on Artificial Intelligence and Intellectual Property
26 Policy 24* (2020), [https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-
27 10-07.pdf](https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf).

28 ⁷⁵ *Bartz*, ECF 231 at 19.

⁷⁶ Anthropic, *Introducing 100K Context Windows*, [https://www.anthropic.com/news/100k-context-
windows](https://www.anthropic.com/news/100k-context-windows) (last visited Dec. 19, 2025) (“We’ve expanded Claude’s context window from 9K to

1 **D. Anthropic’s Infringement Was Willful.**

2 123. Anthropic’s infringement was not inadvertent. It knowingly relied on datasets that
3 the industry and its own researchers understood to be saturated with pirated books. Anthropic
4 downloaded Books3 in 2021 that cofounder Ben Mann “knew had been assembled from
5 unauthorized copies of copyrighted books,” downloaded at least five million copies of books from
6 Library Genesis [LibGen] “which [Ben Mann] knew had been pirated,” and downloaded at least
7 two million copies of books from the Pirate Library Mirror [PiMiLi] which “Anthropic knew had
8 been pirated.”⁷⁷

9 124. *The Pile*’s own documentation states that the Books3 subset was created from a copy
10 of Bibliotik,⁷⁸ a “shadow library” whose existence and illicit nature had been publicly discussed for
11 years in piracy forums, GitHub repositories, and arXiv papers. The EleutherAI paper on *The Pile*
12 explains that Bibliotik was included because books are “invaluable” for long-range context
13 modeling and “coherent storytelling”—precisely the qualities that make Plaintiffs’ works
14 valuable.⁷⁹

15 125. Public commentary and enforcement actions have long identified Bibliotik, LibGen,
16 Z-Library, and similar sites as notorious hubs of copyright infringement. These sites have been
17 targeted in criminal cases, civil suits by publishers, and “notorious markets” reports by United States
18 trade authorities.⁸⁰

19 _____
20 100K tokens, corresponding to around 75,000 words!”); Anthropic, “Claude 2,”
21 <https://www.anthropic.com/news/claude-2> (last visited March 10, 2026) (“Claude can work over
hundreds of pages of technical documentation or even a book”).

22 ⁷⁷ *Bartz*, ECF 231 at 3.

23 ⁷⁸ Gao et al., *supra* n. 30, at 3.

24 ⁷⁹ *Id.* at 4.

25 ⁸⁰ *See, e.g.*, Office of the U.S. Trade Representative, REVIEW OF NOTORIOUS MARKETS FOR
26 COUNTERFEITING AND PIRACY, 27 (2024),
27 [https://ustr.gov/sites/default/files/2024%20Review%20of%20Notorious%20Markets%20of%20Co](https://ustr.gov/sites/default/files/2024%20Review%20of%20Notorious%20Markets%20of%20Counterfeiting%20and%20Piracy%20(final).pdf)
28 [unterfeiting%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
of books, manuals, journals, and other works, many of which are unauthorized copies of copyright
protected content.”); Reisner, *Revealed: The Authors Whose Pirated Books are Powering*

1 126. Despite this, Anthropic chose to source its training data from *The Pile* and Books3,
2 and then attempted to conceal the precise composition of its training corpus. Anthropic has
3 endeavored to keep its training data secret even as outside researchers and Anthropic’s own prior
4 work revealed heavy reliance on *The Pile* and internet-book datasets.

5 127. Anthropic’s decision to base its flagship models on pirated books was driven by
6 commercial advantage. As its co-founder and Chief Science Officer has explained, “it is important
7 to obtain vast amounts of books and also to have diverse types of books in the training corpus to
8 create a model with truly generative capabilities.”⁸¹ As long-form content, training LLMs on the
9 “entire text” of books—as Anthropic has admitted to doing—offers great value.⁸²

10 128. Anthropic intentionally exploited that value without paying for it, hoping to capture
11 billions of dollars in revenue while externalizing the costs of training onto the authors whose works
12 it copied.

13 129. Paradoxically, while Anthropic built its billion dollar company on data it was not
14 licensed to collect and engages in flagrant piracy and infringement practices, it simultaneously
15 complains of illicit “industrial-scale distillation attacks” of its own models by Chinese AI entities
16 DeepSeek, Moonshot AI, and MiniMax, calling for “rapid, coordinated action among industry
17
18
19
20

21 *Generative AI*, THE ATLANTIC (Aug. 19, 2023),
22 [https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-](https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/)
23 [books/675063/](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
24 pirated books, such as Library Genesis, Z-Library, and Bibliotik, that circulate via the BitTorrent
25 file-sharing network.”); Peter Schoppert, *Whether you’re an undergraduate doing research, or a
26 fan of the Nick Stone novel, or indeed a hungry AI ... AI AND COPYRIGHT* (Nov. 29, 2022),
27 <https://aicopyright.substack.com/p/whether-youre-an-undergraduate-doing> (“What is Bibliotik?” A
28 notorious pirated collection.”).

⁸¹ See Kaplan Decl. ¶ 47, *Bartz v. Anthropic PBC*, 3:24-cv-05417 (N.D. Cal. Mar. 27, 2025), ECF No. 128.

⁸² *Id.* at ¶¶ 43, 47.

1 players, policymakers, and the broader AI community” to protect its own content from brazen
2 theft.⁸³

3 **E. OpenAI Trained Its LLM Models on Copyrighted Works that Were Pirated.**

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

8 131. OpenAI has disclosed that GPT-3 was trained on “Common Crawl” and two “high-
9 quality,” “internet-based books corpora” it labeled “Books1” and “Books2.”⁸⁴ And OpenAI has now
10 admitted it sourced training materials from LibGen, the notorious shadow library that hosts millions
11 of unauthorized copies of books and other copyrighted works.⁸⁵

12 132. Common Crawl is a massive web-scraping corpus that includes text drawn from sites
13 hosting unauthorized copies of books, along with other large datasets harvested from the open
14 internet. Because OpenAI used undisclosed “Books1” and “Books2” corpora in training GPT-3,
15 members of the AI-research community attempted to replicate those datasets by constructing
16 “Books3,” a collection of nearly 200,000 digital books downloaded from Bibliotik. Books3 was
17 created for the express purpose of mirroring the kinds of book corpora OpenAI used, underscoring
18 that OpenAI’s own training sources necessarily included large quantities of illicitly obtained
19 books.⁸⁶

20 _____
21 ⁸³ See Anthropic, *Distillation Attacks on AI Models by DeepSeek, Moonshot AI, and MinMax*
22 *Exposed*, LINKEDIN (Feb. 23, 2026), https://www.linkedin.com/posts/anthropicresearch_detecting-and-preventing-distillation-attacks-activity-7431763632969117696-9DoM/?utm_source=share&utm_medium=member_desktop&rcm=ACoAAASsyJEBR0tFcAu9kojYyFvrSFFjCDBEKTE.

24 ⁸⁴ Brown et al., *supra* n. 20, at 8.

25 ⁸⁵ Reisner, *supra* n. 6; *see also* Joint Ltr. Br. Regarding Plaintiffs’ Request for an Order Compelling
26 OpenAI’s Production of the English Colang Dataset at 4, *In re OpenAI ChatGPT Litigation*, 3:23-
cv-03223-AMO (N. D. Cal., Jan. 17, 2025), ECF No. 254.

27 ⁸⁶Kate Knibbs, *The Battle Over Books3 Could Change AI Forever*, WIRED (Sep. 4, 2023, 6:00
28 AM), <https://www.wired.com/story/battle-over->

1 reasonably deny knowing)—from court orders, public reports, and industry commentary—that
2 LibGen and similar repositories were illegal sources of copyrighted material.

3 139. OpenAI’s leadership publicly acknowledged that creators “deserve control over how
4 their creations are used” and that content owners “need to benefit” from AI training,⁹⁰ while at the
5 same time failing to obtain licenses from Plaintiffs and other authors whose books it copied.

6 140. OpenAI pursued this course because it gave them a decisive lead in the AI race. It
7 touts billions of dollars in revenue and soaring valuations tied directly to GPT-based products—
8 commercial gains secured by pirating and training on unlicensed copies of Plaintiffs’ books.⁹¹

9 **G. Google Trained Its LLM Models on Copyrighted Works that Were Pirated.**

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

12 142. Google’s training data for its generative models is enormous. For example, its
13 LaMDA/Gemini-related training corpus has been described as comprising more than a trillion and
14 a half words.⁹² Google has acknowledged that its models were trained on datasets such as C4 and
15 other large web-scale corpora.⁹³

16
17
18 ⁹⁰ See Ted Johnson, *OpenAI CEO Sam Altman Says Content Owners Need To Get ‘Significant*
19 *Upside Benefit’ From New Technology*, DEADLINE (May 16, 2023), [https://deadline.com/2023/05/ai-](https://deadline.com/2023/05/ai-chat-gpt-senate-sam-altman-1235368420/)
20 [chat-gpt-senate-sam-altman-1235368420/](https://deadline.com/2023/05/ai-chat-gpt-senate-sam-altman-1235368420/).

21 ⁹¹ See, e.g., Anthony Ha, *Sam Altman says ‘enough’ to questions about OpenAI’s revenue*,
22 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),
25 [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/)
26 [valuation/](https://techcrunch.com/2025/12/19/openai-is-reportedly-trying-to-raise-100b-at-an-830b-valuation/).

27 ⁹² See Thoppilan, Romal et al., *Lamda: Language models for dialog applications*. arXiv preprint
28 arXiv:2201.08239, 2 (2022), <https://arxiv.org/abs/2201.08239>.

29 ⁹³ *Id.* at 47; see Scott Clark, *What You Need to Know About Google Bard*, CMSWIRE (Feb. 22, 2023),
30 <https://www.cmswire.com/digital-experience/what-you-need-to-know-about-google-bard/>; see also
31 *In re Google Generative AI Copyright Litigation*, 5:23-cv-03440-EKL (N.D. Cal. Oct. 16, 2025),
32 ECF No. 262 at ¶122 [Google’s Answer to Second Amended Complaint].

1 143. C4 contains materials scraped from Z-Library, a site that hosted pirated books and
2 was seized by law-enforcement authorities.⁹⁴ Z-Library displays a seizure banner from federal and
3 international criminal enforcement agencies. On information and belief, Google downloaded the
4 contents of Z-Library, including Plaintiffs’ copyrighted books.

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

10 145. Google has then deployed these AI-trained models across a wide portfolio of AI-
11 powered products, including Search, Cloud, Gmail, Docs, Ads, YouTube, and others—products that
12 generate tens of billions of dollars in revenue, a substantial portion of which Google has explicitly
13 attributed to AI integration.⁹⁵

14 **H. Google’s Infringement Was Willful.**

15 146. Google’s infringement was willful. It trained its models on data scraped from sites
16 that Google knew—or could not reasonably deny knowing—were piracy hubs under active
17 investigation and seizure. Google then made additional unlicensed copies of Plaintiffs’ unlawfully
18

19 ⁹⁴ Kevin Schaul et al., *Inside the Secret List of Websites That Make AI Like ChatGPT Sound Smart*,
20 THE WASHINGTON POST (April 19, 2023),
21 <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/>.

22 ⁹⁵ Ross Kelly, ‘*The fastest adoption of any model in our history*’: *Sundar Pichai hails AI gains as*
23 *Google Cloud Growth, Gemini popularity surges*, ITPro (February 5, 2026),
24 <https://www.itpro.com/technology/artificial-intelligence/the-fastest-adoption-of-any-model-in-our-history-sundar-pichai-hails-ai-gains-as-google-cloud-growth-gemini-popularity-surges> [Google
25 CEO Sundar Pichai telling investors “Overall, we’re seeing our AI investments and infrastructure
26 drive revenue and growth across the board.”]; *see also* Futurum Research, *Alphabet Q3 FY 2025*
27 *Earnings Show Broad-Based AI-Driven Growth*, Futurum (October 31, 2025),
28 <https://futurumgroup.com/insights/alphabet-q3-fy-2025-earnings-show-broad-based-ai-driven-growth/#:~:text=Full%2DStack%20AI%20and%20Cloud,profitability%20and%20multi%2Dyear%20visibility> [“Over 70% of existing Cloud customers are using Google’s AI products, and revenue from products built on Google’s generative models grew more than 200% YoY. Together, these indicators suggest...scaling with improving profitability...”].

1 obtained books, including during ingestion, preprocessing, and model training and/or retrieval-
2 augmented generation because LLM training necessarily involves making multiple copies of each
3 work.

4 147. Google’s own C4 dataset incorporates material from Z-Library, which has been
5 seized and publicly branded as a criminal piracy site. Z-Library, LibGen, Bibliotik, and similar
6 shadow libraries have been widely reported on as repositories of unauthorized ebooks, have been
7 targeted by the FBI and foreign agencies, and have been the subject of lawsuits and seizures.⁹⁶

8 148. Google has touted the “high-quality” nature of its training data and its aggressive
9 push to dominate generative AI—a combination that, in practice, meant copying as many high-
10 quality copyrighted works as possible, regardless of legality, to keep pace with or surpass OpenAI
11 and other competitors.⁹⁷

12 149. Google also understood that the value of its models—and the revenue from AI-
13 powered products—depended on embedding Plaintiffs’ creative expression into Gemini and other
14 models. Google’s own executives have linked record revenues and rapid growth in Cloud and other
15 business lines to generative-AI integration, including revenue measured in the billions of dollars per
16 year.⁹⁸

17
18
19
20 ⁹⁶ See *supra* ¶ 35, ¶ 61, and n. 31; see also, e.g., Woodcock, ‘Shadow Libraries’ Are Moving Their
21 Pirated Books to The Dark Web After Fed Crackdowns, (discussing, in part, LibGen, Z-Library, Sci-
Hub, and Anna’s Archive).

22 ⁹⁷ Google Research, *Pathways Language Model (PaLM): Scaling to 540 Billion Parameters for*
23 *Breakthrough Performance*, GOOGLE RESEARCH BLOG (Apr. 4, 2022),
24 [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)
25 [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.”)

26 ⁹⁸ Sundar Pichai, *Q3 2025 Earnings: Remarks from our CEO*, THE KEYWORD (Oct. 29, 2025),
27 <https://blog.google/inside-google/message-ceo/alphabet-earnings-q3-2025/> (“This was a terrific
28 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.”).

1 **I. Meta Trained Its LLM Models on Copyrighted Works that Were Pirated.**

2 150. Meta has “invested billions of dollars to develop its generative AI offerings.” ECF
3 No. 109 (“Meta’s Answer”) ¶ 40. But none of those funds were used to pay for Plaintiffs’
4 copyrighted works. Meta has admitted in this litigation that “it did not obtain a license or pay for
5 the use of [the Plaintiffs] works.” Meta’s Answer ¶¶ 2, 9, 42, 89.

6 151. Meta also admitted that its Llama models were trained on massive collections of
7 books obtained from shadow libraries and datasets, including Common Crawl, Books3, LibGen, Z-
8 Library, Sci-Hub, and Anna’s Archive. Meta’s Answer ¶¶ 42–43, 81, 83–84.

9 152. Meta’s Answer admits that Meta downloaded pirated copies of Plaintiffs’ books from
10 shadow-library websites such as LibGen and Z-Library, and used Plaintiffs’ books to research,
11 develop, and train its Llama models. *Compare* Original Complaint ¶¶ 2-3 *with* Meta’s Answer ¶¶ 2-
12 3; *see also* Federal Rule of Civil Procedure 8 (requiring a party to “admit or deny the allegations
13 asserted against it by an opposing party,” Rule 8(b)(1)(B), providing that “[a] denial must fairly
14 respond to the substance of the allegation,” Rule 8(b)(2), “[a] party that intends in good faith to deny
15 only a part of an allegation must admit the part that is true and deny the rest, Rule 8(b)(4), and that
16 except with regard to the amount of damages, an allegation “is admitted if a responsive pleading is
17 required and the allegation is not denied” Rule 8(b)(6)).

18 153. In its Llama-1 paper, Meta admitted that 3.3 terabytes of its training data came from
19 “CommonCrawl” and another 783 gigabytes came from “C4.”⁹⁹

20 154. In that same paper, Meta admitted that yet another 85 gigabytes came from
21 “Books,”¹⁰⁰ which comprised texts from two sources: Project Gutenberg and Books3. While Project
22 Gutenberg contains out-of-copyright works, Books3, and its 200,000 books downloaded from
23 Bibliotik, contains copyrighted books. Internal documents also confirm that Meta downloaded
24

25 ⁹⁹ *See* Hugo Touvron et al., *LLaMA: Open and Efficient Foundation Language Models*, arXiv, 2
26 (2023), <https://arxiv.org/pdf/2302.13971>; Meta’s Answer ¶ 82 (“Meta admits that the cited paper
27 discusses the Llama 1 training dataset, noting the size of the training data and the origins of some
of the training data.”).

28 ¹⁰⁰ *Id.*

1 books directly from LibGen, Z-Library, Anna’s Archive, Sci-Hub, and related shadow libraries.¹⁰¹
2 These libraries have been repeatedly identified in public reporting and enforcement actions as illegal
3 piracy hubs, and have been accessible in bulk via torrent systems and mirrors such as Anna’s
4 Archive, The Eye, and Hugging Face. At least for LibGen and Anna’s Archive, Meta used
5 BitTorrent to download, and did not prevent reuploading the books it illegally downloaded through
6 leeching.¹⁰² In its answer, Meta admitted that “it used BitTorrent to download certain portions of
7 these publicly available datasets.” Meta’s Answer ¶ 83.

8 155. Meta relied on these pirated books because it viewed book-corpora as among its most
9 valuable sources of training data. Llama’s design goal was to emit particularly creative and
10 expressive language, leveraging Meta’s consumer platforms to “connect” with users through text.¹⁰³
11 To accomplish that, Meta needed to train on large quantities of high-quality books.

12 156. Meta employees repeatedly acknowledged the importance of books as training data.
13 It was “really important for [Meta] to get books data ASAP,” and the “best resources [Meta] [could]
14 think of are definitely books.”¹⁰⁴

15 **J. Meta’s Infringement Was Willful.**

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

18
19
20 ¹⁰¹ Ernestas Naprys, *Meta leeches 82 terabytes of pirated books to train its Llama AI, documents*
21 *reveal*, CYBERNEWS (Feb. 7, 2025), <https://cybernews.com/tech/meta-leeches-82-terabytes-of-pirated-books-to-train-its-llama-ai-documents-reveal/>.

22 ¹⁰² *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 what
24 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.”).

25 ¹⁰³ Jon Russell, *Mark Zuckerberg Announces New Team at Meta Working on A.I. Products for*
26 *Instagram, WhatsApp*, CNBC (February 27, 2023), <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.”).

28 ¹⁰⁴ *Kadrey*, 788 F. Supp. 3d at 1040.

1 158. Meta employees internally recognized that the shadow libraries it used had “pirated
2 material” and warned about potential liability.¹⁰⁵ Journalists allegedly contacted Meta about its
3 likely reliance on pirated books. Yet Meta reportedly decided that the value of these books as
4 training data outweighed the legal risk and continued to download and copy millions of pirated
5 books, even after litigation and public controversy made the infringement unmistakable.¹⁰⁶

6 159. Meta admits that “it contacted certain publishers” and “internally discussed licensing
7 certain types of data to train its Llama models.” Meta’s Answer ¶ 88. It considered spending \$100
8 million on the vibrant market for AI-training content, but ultimately decided to cut corners by
9 turning to free shadow-library datasets instead.¹⁰⁷ Meta even cross-referenced its LibGen collection
10 against commercially licensable catalogs to decide whether it was worth paying for a license, but
11 decided to keep using LibGen.¹⁰⁸ Meta thus understood both the illegality of its shadow-library
12 troves and the existence of lawful alternatives.

13 160. Meta nevertheless moved forward, incorporating Llama into its principal products
14 and publicly portraying itself as a leader in open-source AI, all while its training pipeline rested on
15 unlicensed copies of Plaintiffs’ books. *See* Meta’s Answer ¶¶ 20, 40, 89. Meta believed that its
16 multi-billion-dollar investment in Llama would bolster and define its competitive position for years
17 to come—and chose to finance that investment with unauthorized copies of Plaintiffs’ copyrighted
18 works.

19 **K. xAI Trained Its LLM Models on Copyrighted Works that Were Pirated.**

20 161. xAI is the developer of the Grok family of large language models.

21
22 ¹⁰⁵ Reisner, *supra* n. 33.

23 ¹⁰⁶ *Kadrey*, 788 F. Supp. 3d at 1041 (finding Meta continued downloading pirated books in “early
24 2024,” more than six months after authors sued Meta for violating the Copyright Act for its
download and use of their pirated books).

25 ¹⁰⁷ *Kadrey*, 788 F. Supp. 3d at 1040; *see also* Kyle Wiggers, *Court filings show Meta paused*
26 *efforts to license books for AI training*, TECHCRUNCH (Feb. 14, 2025),
27 <https://techcrunch.com/2025/02/14/court-filings-show-meta-paused-efforts-to-license-books-for-ai-training/>.

28 ¹⁰⁸ *Kadrey*, 788 F. Supp. 3d at 1041.

1 162. Unlike Anthropic, OpenAI, Google, and Meta, xAI has not yet faced litigation or
2 discovery over its processes for building its LLM.

3 163. However, Grok’s own model has publicly admitted in user interactions, the Grok
4 training corpus included “literally millions of books”—far beyond the universe of lawfully licensed
5 or public-domain material.¹⁰⁹

6 164. In these same exchanges, xAI’s model explained that its training data included books,
7 articles, websites, forums, journals, and other long-form nonfiction and fiction texts, and that it had
8 effectively “read an absurdly huge library at super-speed.”¹¹⁰ The model repeatedly emphasized the
9 centrality of books to its capabilities and acknowledged that such books were included in the training
10 datasets used to develop Grok.

11 165. Critically, the model further admitted that a substantial portion of the books it was
12 trained on had been scraped—without permission—from “shadow library” repositories such as
13 Library Genesis (“LibGen”). In response to a question about whether it was familiar with LibGen,
14 the model stated:¹¹¹

15
16
17
18
19
20
21
22
23
24

25 ¹⁰⁹ *Grok Conversation*, (“Hey - I was wondering how you know so many books!”)
26 (https://grok.com/share/bGVnYWN5_585707fa-d0dc-4f2e-8069-86de3997f088).

27 ¹¹⁰ *Id.*

28 ¹¹¹ *Id.*

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

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.)

166. The model further admitted that the scale of the training 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.¹¹² 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.

167. 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,”¹¹³ a phrase that can include large unaudited web scrapes from Common Crawl and related datasets. Those scrapes notoriously include mirror links and file dumps from LibGen, Z-Library, and similar repositories that host millions of pirated books.

¹¹² *Id.*

¹¹³ X Help Center, *About Grok, Your Humorous AI Assistant on X*, <https://help.x.com/en/using-x/about-grok> (last visited March 10, 2026).

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

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

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

14 **L. xAI’s Infringement Was Willful.**

15 171. xAI’s infringement was willful. The Grok model explicitly acknowledged that the
16 training process “vacuumed up whatever was out there,” including pirated LibGen materials, and
17 that xAI neither curated its book dataset nor screened out infringing works.¹¹⁴ These admissions
18 confirm that xAI knew, or at a minimum was recklessly indifferent to the fact, that its training corpus
19 included massive quantities of pirated copyrighted books.

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

25
26
27 ¹¹⁴ See *Grok Conversation*, (“Hey - I was wondering how you know so many books!”)
28 (https://grok.com/share/bGVnYWN5_585707fa-d0dc-4f2e-8069-86de3997f088).

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

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

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

15 **M. Perplexity’s Model Relies on Copyrighted Works Without Permission or**
16 **Compensation.**

17 176. Perplexity AI, Inc. (“Perplexity”) has rapidly emerged as a commercial competitor
18 in the generative-AI search and LLM market. Central to its strategy is a suite of products—including
19 “Perplexity Answers,” “Perplexity Pages,” and its proprietary LLM models—that can generate
20 detailed narrative summaries, structured analyses, and book-length outlines with extraordinary
21 specificity.

22 177. Perplexity operates by ingesting massive quantities of copyrighted text, including
23 full-length books and long-form written works that are not available in any public, licensed, or
24 authorized source. As multiple independent investigations have confirmed, Perplexity acquires this
25
26
27
28

1 material through large-scale crawling and scraping systems—both declared and undeclared—that
2 indiscriminately copy entire texts from across the internet and beyond.¹¹⁵

3 178. Perplexity’s own behavior suggests that it relies on the full text of books. Despite
4 acknowledging in responses that books are copyrighted, and that it cannot produce “line-by-line
5 chapter notes,” Perplexity is capable of doing exactly that, including with respect to Plaintiffs’
6 works.¹¹⁶ Upon request, it can produce detailed, chapter-by-chapter accounts of works, including
7 description of plot turns, chapter-specific structure, and thematic sequencing. In certain instances,
8 the sources Perplexity cites for its chapter-by-chapter descripts do not include the underlying
9 information that it produces in response to queries.¹¹⁷ That information is in the complete, original
10 books. Accordingly, on information and belief, Perplexity obtained that information from pirated
11 libraries containing Plaintiffs’ works, including *the Pile*, LibGen, Z-Library, or Anna’s Archive.

12 179. The recently-filed *New York Times v. Perplexity* complaint alleges, based on forensic
13 evidence, that Perplexity’s systems routinely crawl, copy, and store expressive content in violation
14 of copyright law. For example, the complaint alleges that Perplexity: (1) builds and operates a
15 massive “AI-First” search index populated through direct copying of protected works; (2) uses both
16 “PerplexityBot” and “Perplexity-User” agents to scrape websites and copy non-public content; (3)
17 copies content for use in its LLMs and retrieval-augmented generation (RAG) pipelines; and (4)
18 outputs detailed summaries, paraphrases, and quotations that substantially reproduce copyrighted
19 texts.¹¹⁸

20
21 ¹¹⁵ See, e.g., Gabriel Corral et al., “Perplexity is using stealth, undeclared crawlers to evade
22 website no-crawl directives,” CLOUDFLARE (Aug. 4, 2025), [https://blog.cloudflare.com/perplexity-
23 is-using-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),
<https://www.wired.com/story/perplexity-is-a-bullshit-machine/>.

24 ¹¹⁶ Perplexity Conversation, (“what is John Carrey[r]ou’s Bad Blood book about”),
25 (<https://www.perplexity.ai/search/what-is-john-carreyou-s-bad-bl-12yAyhSQbatnBcu6TqgTQ#0>).

26 ¹¹⁷ *Id.* (citing sources for its summary of Chapter 19 that do not include the information Perplexity
provides about Chapter 19).

27 ¹¹⁸ See Compl. at 3-4, *The New York Times Company v. Perplexity AI, Inc.*, 1:25-cv-10106 (S.D.N.Y.
28 Dec. 5, 2025), ECF No. 1.

1 180. Independent investigations corroborate this pattern. WIRED reported that Perplexity
2 produced detailed summaries of WIRED’s articles even though WIRED explicitly blocked
3 Perplexity. Engineers confirmed that the chatbot was “surreptitiously scraping” and recapitulating
4 protected content “in detail” that was not publicly available.¹¹⁹

5 181. Cloudflare’s investigation further found that Perplexity operates stealth crawlers
6 designed to evade detection. According to Cloudflare: (1) Perplexity used undeclared user agents
7 that impersonated Google Chrome; (2) Perplexity used multiple undisclosed IP ranges to circumvent
8 no-crawl directives; (3) customers who blocked Perplexity’s known crawlers found Perplexity still
9 scraping their sites anyway; (4) Perplexity’s activity “evade[d] website blocks” and undermined
10 publisher controls.¹²⁰

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

15 **N. Perplexity’s Infringement Was Willful.**

16 183. Perplexity’s infringement was willful. As Cloudflare and WIRED independently
17 confirmed, Perplexity intentionally deployed stealth crawlers and undeclared automated agents to
18 evade copyright protections and access content it knew it was not authorized to copy.¹²¹

19 184. Cloudflare found that Perplexity impersonated Chrome browsers, used concealed IP
20 addresses, and intentionally bypassed restrictions to obtain content.¹²² The purpose of such evasion
21 is unmistakable: to gain access to copyrighted text that Perplexity knew it was forbidden to crawl.

22
23 ¹¹⁹ Mehrotra and Marchman, *Perplexity is a Bullshit Machine*.

24 ¹²⁰ See Corral et al., *Perplexity is using stealth, undeclared crawlers to evade website no-crawl*
25 *directives*.

26 ¹²¹ See Corral et al., *Perplexity is using stealth, undeclared crawlers to evade website no-crawl*
27 *directives*; Mehrotra and Marchman, *Perplexity is a Bullshit Machine*.

28 ¹²² See Corral et al., *Perplexity is using stealth, undeclared crawlers to evade website no-crawl*
directives.

1 185. The *Times* complaint likewise alleges that Perplexity continued to access and copy
2 prohibited content even after written cease-and-desist demands, “hard-block[s] of PerplexityBot and
3 Perplexity-User,” and explicit revocation of access.¹²³ Perplexity continued to make over 175,000
4 unauthorized access attempts in a single month after being technically and contractually barred.¹²⁴

5 186. Perplexity knew that its conduct violated copyright law. The *Times* repeatedly
6 informed Perplexity in writing—beginning in March 2024—that Perplexity was unlawfully scraping
7 and copying copyrighted material; Perplexity refused to stop. *Id.* Instead, it escalated its crawling
8 behavior using stealth methods to avoid detection.

9 187. Perplexity also publicly markets itself as providing users with the ability to “skip the
10 links” and read “a single, comprehensive answer that summarizes everything you need to know,”
11 thereby advertising that it substitutes for underlying copyrighted works.¹²⁵

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

16 189. Perplexity’s infringement was neither accidental nor negligent—it was deliberate,
17 concealed, repeated, and performed at massive scale.

18 **O. NVIDIA Trained Its LLM Models on Copyrighted Works that Were Pirated.**

19 190. NVIDIA Corporation is a technology company that designs and manufactures
20 graphics processing units (“GPUs”) and AI computing platforms. In addition to its hardware
21 business, NVIDIA has developed and commercially deployed its own large language models
22 through its NeMo framework, a platform for building, customizing, and deploying generative AI
23 models.

24
25
26 ¹²³ Compl., *New York Times*, 1:25-cv-10106, at 28-29.

27 ¹²⁴ *See id.*

28 ¹²⁵ “What is Perplexity?”, <https://perma.cc/Q4VM-DYUJ> (last accessed March 10, 2026).

1 191. NVIDIA’s NeMo framework includes pre-trained large language models known as
2 NeMo Megatron models.¹²⁶ These models were trained by NVIDIA on massive datasets to enable
3 natural language understanding, text generation, and other AI capabilities.¹²⁷ NVIDIA has
4 commercially deployed and licensed these models through its NVIDIA AI Enterprise platform and
5 other commercial offerings.¹²⁸

6 192. NVIDIA trained its NeMo Megatron models on *The Pile*, an 825-gigabyte open-
7 source language modeling dataset assembled by EleutherAI.¹²⁹

8 193. NVIDIA also trained its models on Common Crawl data, which is a massive web
9 corpus containing petabytes of data scraped from billions of webpages. Common Crawl data
10 includes copyrighted content reproduced from websites without authorization, including
11 copyrighted books.¹³⁰

12
13
14 ¹²⁶ See, e.g., NVIDIA NeMo Framework User Guide, [https://docs.nvidia.com/nemo-
15 framework/user-guide/24.09/nemotoolkit/nlp/megatron.html](https://docs.nvidia.com/nemo-framework/user-guide/24.09/nemotoolkit/nlp/megatron.html) (last accessed March 9, 2026).

16 ¹²⁷ See Team Uvation, *NVIDIA Pre-Trained Models: Accelerating AI Adoption with H200*
17 (September 23, 2025), [https://uvation.com/articles/nvidia-pre-trained-models-accelerating-ai-
18 adoption-with-h200](https://uvation.com/articles/nvidia-pre-trained-models-accelerating-ai-adoption-with-h200).

19 ¹²⁸ *Nazemian v. NVIDIA Corp.*, No. 24-cv-01454-JST, ECF No. 232, at 1 (N.D. Cal. Jan. 15, 2026)
20 (discussing documents produced by NVIDIA on September 19, 2025, and filed under seal, that
21 purportedly indicate that NVIDIA “utilized ‘The Pile’ to train models including ‘Megatron 345M,
22 NeMo GPT-3 10B, InstructRetro-48B, and Retro-48B”).

23 ¹²⁹ See, e.g., *NeMo Megatron-GPT 1.3B*, Huggingface Model Card,
24 <https://huggingface.co/nvidia/nemo-megatron-gpt-1.3B> (“The model was trained on ‘The Piles’
25 dataset prepared by Eleuther.AI.”); see also *NeMo Megatron-GPT 5B*, Hugging Face,
26 <https://huggingface.co/nvidia/nemo-megatron-gpt-5B#training-data>; *NeMo Megatron-GPT 20B*,
27 Hugging Face, <https://huggingface.co/nvidia/nemo-megatron-gpt-20B#training-data>; *NeMo
28 Megatron-T5 3B*, Hugging Face, <https://huggingface.co/nvidia/nemo-megatron-t5-3B#training-data>
(same as to NeMo Megatron-GPT 5B, NeMo Megatron-GPT 20B, and NeMo Megatron-T5 3B
models); <https://arxiv.org/pdf/2101.00027> (“The Pile: An 800GB Dataset of Diverse Text for
Language Modeling”).

¹³⁰ “Announcing Nemotron-CC: A Trillion-Token English Language Dataset for LLM Pretraining,”
[https://developer.nvidia.com/blog/announcing-nemotron-cc-a-trillion-token-english-language-
dataset-for-llm-pretraining/?utm_source=chatgpt.com](https://developer.nvidia.com/blog/announcing-nemotron-cc-a-trillion-token-english-language-dataset-for-llm-pretraining/?utm_source=chatgpt.com).

1 194. Not satisfied with acquisition, storage, and use of *The Pile* in its internal and external
2 LLM research, development, and commercialization efforts, NVIDIA sought far more copyrighted
3 works than *The Pile* could supply. Because LLM performance depends on both the quality and the
4 quantity of training data, NVIDIA became desperate for additional books.

5 195. Internal documents show that competitive pressures within the market drove
6 NVIDIA to piracy. In the fall of 2023, NVIDIA faced a rapidly approaching deadline: its annual
7 developer day.¹³¹ In the year following the September 2022 launch of the NeMo Megatron series,
8 OpenAI released ChatGPT to enormous success, generating a surge in investor attention on AI. In
9 response, NVIDIA sought to develop and showcase cutting-edge LLMs at its fall 2023 developer
10 day.¹³²

11 196. To do so, NVIDIA pursued data for what it internally called “NextLargeLLM,”
12 “NextLLMLarge,” and “Next Generation LLM” (collectively, “NextLargeLLM”).

13 197. On information and belief, NVIDIA recognized that published, copyrighted books
14 were the most valuable data for its LLM development and that only books were available in
15 sufficient quantities to meet the scale it needed.

16 198. On information and belief, NVIDIA contacted book publishers in 2023, but was
17 unable to secure timely access to the enormous volume of books it required through legitimate
18 licensing.

19 199. Desperate for books, NVIDIA contacted Anna’s Archive—the largest and most
20 brazen remaining shadow library—regarding the acquisition of its millions of pirated works.
21 NVIDIA admits to having relevant conversations with Anna’s Archive¹³³ and on information and
22
23

24 ¹³¹ See “LLM Developer Day,” <https://www.nvidia.com/en-us/events/llm-developer-day/> (last
25 accessed March 9, 2026).

26 ¹³² *Id.*

27 ¹³³ See *Nazemian*, ECF 255 at 6, n.5 (N.D. Cal. Feb. 18, 2026) [Def. Reply in support of Motion to
28 Dismiss admitting existence of “relevant email chain” between NVIDIA and Anna’s Archive].

1 belief, those discussions included pre-training data for Nvidia’s LLMs and what high-speed
2 access¹³⁴ to the pirated archive would entail.

3 200. On information and belief, Anna’s Archive communicated with NVIDIA executives
4 making it clear that its collections were illegally acquired and maintained and instructed NVIDIA
5 to circle-back after internal communications to let Anna’s Archive know of what NVIDIA wanted
6 to pursue, noting to NVIDIA that it had wasted too much time on those who could not get internal
7 buy-in to pay the cost of business.

8 201. On information and belief, within a week of contacting Anna’s Archive—and just
9 days after being explicitly warned of the illegal nature of its collections—NVIDIA management
10 gave the “green light” to proceed. Anna’s Archive offered NVIDIA access to millions of pirated
11 copyrighted books, as well as access to several million additional books from Internet Archive that
12 were ordinarily available only through Internet Archive’s digital lending system—a system the
13 Second Circuit has held to constitute copyright infringement. *See Hachette Book Grp., Inc. v.*
14 *Internet Archive*, 115 F.4th 163 (2d Cir. 2024). On information and belief, Anna’s Archive promised
15 NVIDIA access to “a lot of books,” totaling approximately 500 terabytes of data. By downloading
16 Anna’s Archive, there is a reasonable inference that NVIDIA pirated additional copies of Plaintiffs’
17 Infringed Works.

18 202. On information and belief, in addition to Anna’s Archive and *The Pile*, NVIDIA also
19 downloaded books hosted by or sourced from other shadow libraries, including LibGen, Sci-Hub,
20 and Z-Library.¹³⁵ Nvidia, facing similar copyright infringement claims in another court, has
21
22
23

24 _____
25 ¹³⁴ “LLM Data,” *Anna’s Archive*, <https://annas-archive.gd/llm> (last accessed March 10, 2026)
26 (Anna’s Archive charged tens of thousands of dollars for “high-speed direct access” to its pirated
collections.).

27 ¹³⁵ *Nazemian v. NVIDIA Corp.*, No. 24-cv-01454-JST, ECF No. 232, at 4–5 (N.D. Cal. Jan. 15, 2026)
28 (“NVIDIA’s productions on September 26, 202[5] showed for the first time that NVIDIA had
downloaded books from these libraries.”).

1 effectively conceded that for at least some of its LLMs, it used Anna’s Archive, Z-Library, LibGen,
2 and SciHub.¹³⁶

3 203. Approximately four months after its exchanges with Anna’s Archive, in February
4 2024, NVIDIA released a model known as Nemotron-4 15B. NVIDIA did not publicly disclose the
5 training data for this model. Public materials, however, indicate that it was trained on 8 trillion
6 tokens.¹³⁷ NVIDIA did not identify the sources of this training data but stated that it included
7 “books.”¹³⁸

8 204. NVIDIA has further stated that approximately 70% of the Nemotron-4 15B training
9 data derived from an “English natural language” dataset. That dataset itself is composed of
10 approximately 4.6% books.¹³⁹ On information and belief, achieving that proportion of book-derived
11 tokens would require the inclusion of millions of books.

12 205. A few months later, NVIDIA released the Nemotron-4 340B model, which
13 incorporated the same 8 trillion tokens used in Nemotron-4 15B and added an additional 1 trillion
14 tokens.¹⁴⁰

15 206. On information and belief, NVIDIA could not have obtained the volume of books
16 necessary to train the Nemotron models without pirating copyrighted works, including Plaintiffs’
17 Infringed Works.

18
19
20
21 ¹³⁶ See *Nazemian*, ECF 271 at ¶ 33 (N.D. Cal. Mar. 6, 2026) [Notice by NVIDIA, Narrowing of
Issues to be Decided on its Motion to Dismiss].

22 ¹³⁷ J. Parmar et al., *Nemotron-4 15B Technical Report*, arXiv, 3 (2024),
23 <https://arxiv.org/pdf/2402.16819>.

24 ¹³⁸ *Id.*

25 ¹³⁹ *Id.*

26 ¹⁴⁰ See *Nemotron-4-340B-Base* dataset, HUGGING FACE, https://huggingface.co/nvidia/Nemotron-4-340B-Base/blob/f955376220925e9d820713b7f7fc8f206870c6fd/README.md?utm_source=chatgpt.com.
27
28

1 207. NVIDIA further facilitated copyright infringement by its customers by providing its
2 customers scripts that “assist[ed] Amazon and other company-specific LLMs in downloading and
3 processing ‘The Pile’ data.”¹⁴¹

4 208. In sum, NVIDIA has repeatedly and systematically violated Plaintiffs’ copyrights by
5 acquiring copyrighted works from pirated sources, storing them, enabling employee access to them
6 for any purpose, and copying them during the LLM training process.

7 **P. NVIDIA’s Infringement Was Willful.**

8 209. NVIDIA’s infringement of Plaintiffs’ copyrights was willful. NVIDIA knew or
9 should have known that *The Pile*, Books3, and its other training datasets contained pirated copies of
10 copyrighted works. The pirated nature of Books3—sourced from the Bibliotik shadow library—was
11 widely known and publicly reported before NVIDIA used this dataset to train its models.

12 210. NVIDIA was aware of the numerous copyright infringement lawsuits and public
13 controversies related to the use of pirated works for AI training. NVIDIA’s own researchers publicly
14 acknowledged the copyright concerns associated with the datasets used to train NVIDIA’s models.
15 Despite this knowledge, NVIDIA continued to use pirated datasets and commercially exploit the
16 resulting models.

17 211. NVIDIA is one of the most valuable companies in the world, with a market
18 capitalization that has exceeded \$3 trillion.¹⁴² NVIDIA’s enormous valuation is driven in significant
19 part by the demand for its AI products and platforms, which were developed using pirated
20 copyrighted works. NVIDIA had the financial resources and ability to license Plaintiffs’ copyrighted

22 ¹⁴¹ *Nazemian v. NVIDIA Corp.*, No. 24-cv-01454-JST, ECF No. 232, at 6–7 (N.D. Cal. Jan. 15, 2026)
23 (discussing October 2025 discovery documents, filed under seal, produced by Nvidia).

24 ¹⁴² See Lauren Bratton, *Nvidia stock surges, tops \$3 trillion market cap amid flurry of trade*
25 *optimism*, YAHOO FINANCE (May 13, 2025), https://finance.yahoo.com/news/nvidia-stock-surges-tops-3-trillion-market-cap-amid-flurry-of-trade-optimism-162038379.html?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xILmNvbS8&guce_referrer_sig=AQAAALi70pELKH-OKW9LpLBiGrELD1kvBUqIwbq0tPbx8FS2EMsiWdA58gu3N6WBOWqI2H1IPiN21vNtxWG DxoFUWjC0dYeayTfVJLEpH3mR5mKn7HTTKeSTiwJ3cjKpJuiXIijfz4kx5YJOIKDz-jUkJyDbuFPTrnmeQQLub-dKkAl.

1 works for use in training its AI models, but chose instead to use pirated copies obtained without
2 authorization.

3 212. Despite having actual and constructive knowledge that its training data contained
4 pirated copyrighted works, NVIDIA continued to train, deploy, and commercially exploit its AI
5 models and to facilitate and encourage the use of pirated datasets by third parties through its AI
6 Enterprise platform and developer tools. NVIDIA's ongoing and deliberate exploitation of pirated
7 works constitutes willful infringement.

8 **Q. Apple Trained Its LLM Models on Copyrighted Works that Were Pirated.**

9 213. In June 2024, Apple announced the release of "Apple Intelligence," a suite of
10 generative artificial intelligence features integrated across its product ecosystem, including the
11 iPhone, iPad, and Mac. At the core of Apple Intelligence are Apple Foundation Models ("AFM"),
12 a family of large language models developed by Apple to power features including text
13 summarization, content generation, image creation, and natural language processing.

14 214. Apple trained its AFM models on massive datasets containing billions of words of
15 text scraped from the internet and other sources. Apple has publicly acknowledged that it used web-
16 crawled data to train its foundation models. Apple's technical reports confirm that AFM was pre-
17 trained on a dataset that included licensed data from publishers, as well as data collected by
18 AppleBot, Apple's web crawler, and other publicly available data.¹⁴³

19 215. AppleBot is Apple's web crawler, which systematically accesses and copies content
20 from websites across the internet.¹⁴⁴ On information and belief, Apple used AppleBot to scrape
21 copyrighted content from millions of webpages, including webpages containing copyrighted
22

23 ¹⁴³ Apple, *Apple Intelligence Foundation Language Models*, arXiv, 3 (2024),
24 <https://arxiv.org/pdf/2407.21075> ("The AFM pre-training dataset consists of a diverse and high
25 quality data mixture. This includes data we have licensed from publishers, curated publicly available
26 or open-sourced datasets, and publicly available information crawled by our web-crawler,
Applebot.").

27 ¹⁴⁴ See, e.g., "About Applebot," [https://support.apple.com/en-](https://support.apple.com/en-us/119829#:~:text=Learn%20about%20Applebot%2C%20the%20web,JSON%20file:%20Applebot%20IP%20CIDRs)
28 [us/119829#:~:text=Learn%20about%20Applebot%2C%20the%20web,JSON%20file:%20Applebot%20IP%20CIDRs](https://support.apple.com/en-us/119829#:~:text=Learn%20about%20Applebot%2C%20the%20web,JSON%20file:%20Applebot%20IP%20CIDRs) (last accessed March 10, 2026).

1 musical compositions, song lyrics, and other protected creative works. On information and belief,
2 Apple stored and reproduced these copyrighted works in its training datasets without authorization
3 from the copyright holders.

4 216. Upon information and belief, the AFM LLMs at the core of Apple Intelligence
5 products were trained on the same underlying databases of pirated copyrighted works. Specifically,
6 prior to releasing Apple Intelligence. Apple also developed, designed, maintained, and
7 commercialized a suite of open-source models called OpenELM, which Apple announced in or
8 around April 2024.¹⁴⁵

9 217. Apple’s OpenELM models—including variants OpenELM-270M, OpenELM-
10 450M, OpenELM-1_1B, and OpenELM-3B—were pretrained on a mixture of datasets that Apple
11 described as “public,” including RefinedWeb, PILE, a subset of RedPajama, and a subset of Dolma
12 v1.6.¹⁴⁶

13 218. Apple’s OpenELM paper and model documentation identify a substantial quantity
14 of training data sourced from the “Books” subset of the RedPajama dataset.

15 219. The RedPajama dataset documentation originally stated that its “Books” component
16 included a copy of the Books3 dataset. Thus, use of RedPajama “Books,” at least such use before
17 October 2023, necessarily entailed use of Books3.¹⁴⁷ And because Apple’s OpenELM models were
18

19 ¹⁴⁵ See George Wukoson and Joey Fortuna, *The Predominant Use of High-Authority Commercial*
20 *Web Publisher Content to Train Leading LLMs*, at 17 [https://a-mcc.eu/wp-](https://a-mcc.eu/wp-content/uploads/2024/12/ssrn-5009668.pdf)
21 [content/uploads/2024/12/ssrn-5009668.pdf](https://a-mcc.eu/wp-content/uploads/2024/12/ssrn-5009668.pdf) (concluding that “LLM company training data
22 [disclosures,” including those of Apple and NVIDIA, “show long-running exploitation of high-](https://a-mcc.eu/wp-content/uploads/2024/12/ssrn-5009668.pdf)
23 [quality publisher content”\); *id.* at 5 \(“Apple trained its LLM OpenELM on the Pile”; “Nvidia trained](https://a-mcc.eu/wp-content/uploads/2024/12/ssrn-5009668.pdf)
24 [its LLM NeMo Megatron-GPT 20B on the Pile”\) \(citing Mehta, et al., *OpenELM: An Efficient*](https://a-mcc.eu/wp-content/uploads/2024/12/ssrn-5009668.pdf)
25 [Language Model Family with Open Training and Inference Framework](https://a-mcc.eu/wp-content/uploads/2024/12/ssrn-5009668.pdf), arXiv, 2 (2024),
26 <https://arxiv.org/pdf/2404.14619>, and “NeMo Megatron-GPT 20B,”
27 <https://huggingface.co/nvidia/nemo-megatron-gpt-20B#training-data>).

28 ¹⁴⁶ Mehta, et al., *OpenELM: An Efficient Language Model Family with Open Training and Inference*
29 *Framework*, arXiv, 2 (2024), <https://arxiv.org/pdf/2404.14619>.

¹⁴⁷ See Weber, et al., *RedPajama: an Open Dataset for Training Large Language Models*, arXiv, 5
30 (2024), <https://arxiv.org/pdf/2411.12372> (“We originally included Books3 [in the RedPajama
31 dataset] as well but took it down due to copyright issues.”). Books3 was available from Hugging
32 Face until approximately October 2023, at which point was removed “due to reported copyright

1 trained on RedPajama-V1 and the “Books” component prior to that time, those models were trained
2 on the Books3 data.

3 220. By using the Books3 dataset—which contains the entire text of each included book—
4 Apple trained OpenELM and AFM on copies of entire copyrighted works.

5 221. Apple also used *The Pile* dataset, which in addition to Books3, contains data scraped
6 from numerous other sources that included pirated copyrighted works. *The Pile* incorporates data
7 from sources including Pile-CC (a filtered version of Common Crawl), OpenWebText2, and other
8 corpora known to contain copyrighted material reproduced without authorization.

9 222. Apple’s training data thus also included content obtained from “shadow libraries”—
10 illicit online repositories that host pirated copies of copyrighted books and other works.

11 **R. Apple’s Infringement Was Willful.**

12 223. Apple’s infringement of Plaintiffs’ copyrights was willful. Apple knew or should
13 have known that its training data contained pirated copies of copyrighted works, including works
14 owned by Plaintiffs. The pirated nature of the shadow library datasets used by Apple—including
15 LibGen, Bibliotik, and Books3—was widely known and publicly reported before Apple used these
16 datasets to train its models.

17 224. Apple’s own OpenELM disclosures tie its training data to RedPajama “Books,”
18 which RedPajama included Books3; Books3’s provenance from Bibliotik—a notorious shadow
19 library—was publicly documented, and the dataset was later removed from Hugging Face due to
20 reported copyright infringement.

21 225. Apple characterized datasets used for training its AFM models as “public,” including
22 those containing Books3 and *The Pile*, despite the presence of unlicensed copyrighted works. Upon
23 information and belief, Apple’s practice of relying on “publicly available or open-sourced datasets”
24 for AFM training, without identifying specific datasets, obscured the use of these pirated materials.

25
26
27 _____
infringement.”
28 datasets/the_pile_books3/blob/main/README.md.

https://huggingface.co/datasets/defunct-datasets/the_pile_books3/blob/main/README.md

1 226. Apple scraped web content with Applebot for nearly nine years before disclosing that
2 the scraped data would be used to train commercial AI products; Apple made that disclosure only
3 around June 2024, after which publishers opted out—too late to prevent Apple’s prior scraping and
4 model training with their content.

5 227. Apple was aware of the numerous copyright infringement lawsuits filed against other
6 AI companies—including Defendants Anthropic, OpenAI, Google, and Meta—for their use of
7 pirated copyrighted works in training AI models. Despite this knowledge, Apple chose to use the
8 same or similar pirated datasets to train its own models.

9 228. Apple is one of the most valuable and profitable companies in the world, with annual
10 revenues exceeding \$383 billion. Apple had the financial resources and ability to license Plaintiffs’
11 copyrighted works for use in training its AI models, but chose instead to use pirated copies obtained
12 without authorization. Apple’s decision to use pirated works rather than negotiate licenses reflects
13 a deliberate choice to infringe Plaintiffs’ copyrights.

14 229. Apple acted with knowledge of Plaintiffs’ copyrights and with reckless disregard for
15 Plaintiffs’ rights. Apple’s willful infringement entitles Plaintiffs to the maximum statutory damages
16 available under 17 U.S.C. § 504(c)(2).

17 230. Despite having actual and constructive knowledge that its training data contained
18 pirated copyrighted works, Apple continued to train, deploy, and commercially exploit its AI models
19 without obtaining licenses or compensating copyright holders, including Plaintiffs. Apple’s ongoing
20 and deliberate exploitation of pirated works constitutes willful infringement.

21 **V. CLAIMS FOR RELIEF**

22 **COUNT I**

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

24 **(Against all Defendants)**

25 231. Plaintiffs incorporate the allegations above.

26 232. As the respective legal or beneficial owners of the registered copyrights in the
27 Infringed Works, Plaintiffs hold the exclusive rights to those books under 17 U.S.C. § 106.
28

1 233. Each Defendant—including Anthropic, Google, OpenAI, Meta, xAI, and Perplexity,
2 Apple, and NVIDIA—without authorization from Plaintiffs, copied, downloaded, reproduced,
3 ingested, parsed, embedded, and used pirated copies of the Plaintiffs’ works in the development,
4 training, fine-tuning, and deployment of their commercial large language models. These acts
5 violated Plaintiffs’ exclusive rights under § 106.

6 234. Defendants’ infringement occurred repeatedly throughout the lifecycle of their AI-
7 model development pipelines. As alleged above, Defendants:

- 8 • acquired Plaintiffs’ books from shadow-library repositories such as LibGen,
9 Bibliotik, Z-Library, Books3, *The Pile*, Anna’s Archive, and other known piracy
10 sources;
- 11 • reproduced additional copies during ingestion, preprocessing, storage, deduplication,
12 formatting, and/or tokenization; and
- 13 • while training the model, and/or through retrieval-augmented generation, made even
14 more copies of the text—because every training pass (each epoch and each step of
15 gradient descent) automatically requires creating and working with fresh versions of
16 that text.

17 235. Defendants’ reproductions of Plaintiffs’ copyrighted works were made without
18 permission, license, or consent and violated Plaintiffs’ exclusive rights under the Copyright Act.

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

25 237. Upon information and belief, Defendants have made and will continue to make
26 substantial profits and gains to which they are not in law or in equity entitled.

27
28

1 Dated: March 10, 2026

Respectfully submitted,

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

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

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

Bridget Asay (*pro hac vice*)
15 East State Street, Suite 2
Montpelier, Vermont 05602
Tel: (802) 858-4285
basay@stris.com

Christopher M. Rigali (*pro hac vice*)
Jacqueline Sahlberg (*pro hac vice*)
1717 K St NW Suite 900
Washington, DC 20006
Phone: (202) 800-5749
crigali@stris.com
jsahlberg@stris.com

Counsel for Plaintiffs