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

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

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

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

Counsel for Plaintiff

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

CHICKEN SOUP FOR THE SOUL, LLC,

Plaintiff,

v.

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

Defendants.

Civil Case No.:

COMPLAINT

DEMAND FOR JURY TRIAL

1. Plaintiff Chicken Soup for the Soul, LLC brings this action against Anthropic PBC ("Anthropic"), Google LLC ("Google"), OpenAI, Inc. and its affiliated entities ("OpenAI"), Meta Platforms, Inc. ("Meta"), xAI Corporation ("xAI"), Perplexity AI, Inc. ("Perplexity"), Apple Inc. ("Apple"), and NVIDIA Corporation ("NVIDIA"), (collectively, "Defendants"), and allege as follows:

I. INTRODUCTION

2. This case concerns a straightforward and deliberate act of theft that constitutes copyright infringement. Anthropic, Google, OpenAI, Meta, xAI, Apple, Perplexity, and NVIDIA, illegally copied vast quantities of copyrighted books without permission and then used those stolen copies to build and train their commercial large language models ("LLMs") and/or optimize their product. Defendants helped themselves to the copyrighted works of thousands of authors—including bestselling writers, Pulitzer Prize-winning journalists, and creators of widely read nonfiction and fiction.[1]

3. Plaintiff Chicken Soup for the Soul, LLC owns and controls the iconic *Chicken Soup for the Soul* literary franchise—one of the most successful and widely distributed nonfiction book series in modern publishing history. Since the publication of the first Chicken Soup for the Soul volume in 1993, the series has grown into a global publishing phenomenon, with hundreds of titles and more than 500 million books sold worldwide. The Chicken Soup for the Soul brand is synonymous with emotionally resonant storytelling and carefully edited first-person narratives that capture authentic human experiences.

4. For decades, these books have been curated, edited, and refined to present clear, compelling narrative structures and expressive prose that resonate with readers across cultures and

[1] The undersigned counsel has contemporaneously filed a Notice of Related Case pursuant to Civil Local Rule 3-12 informing the Court that this action is related to *Cambronne Inc., et al. v. Anthropic PBC, et al.*, Case No. 5:25-cv-10897-PCP (N.D. Cal.). The actions arise from substantially the same course of conduct—Defendants' acquisition of pirated copies of copyrighted books from shadow-library repositories and the use of those works to train and develop large language models—and involve the same Defendants, nearly identical factual allegations, and an identical legal claim. Assignment of the cases to the same judge would promote judicial efficiency and avoid duplication of effort.

generations. The extraordinary commercial success and editorial consistency of the Chicken Soup for the Soul series made it among the most recognizable and valuable collections of narrative nonfiction in the publishing industry—and, as alleged below, an especially attractive target for companies seeking high-quality textual data to train large language models.

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

6. Defendants targeted Plaintiff's works because they were of exceptional value as training data. Defendants have acknowledged—internally and publicly—that long-form, high-quality books are the "gold-standard" training material for LLMs. Books teach models how narrative flows, how human expression is structured, how syntax and rhythm operate, and how ideas are communicated through creative choices. Instead of paying for that value, Defendants pilfered illegal copies and used those copies to build systems now worth many hundreds of billions of dollars.

7. The *Chicken Soup for the Soul* series is particularly valuable in this regard. Each volume contains dozens of tightly edited, first-person narratives written in natural, conversational language that conveys emotion, moral reflection, and coherent storytelling in concise form. These characteristics make the series uniquely well suited for training large language models to replicate authentic human voice, narrative pacing, emotional tone, and story structure. Instead of paying for that value or licensing access to these works, Defendants pilfered illegal copies and used those copies to build systems now worth many hundreds of billions of dollars.

8. The infringement here occurred at least twice for every work.

9. *First*, Defendants obtained Plaintiff's copyrighted books from illegal shadow libraries.

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

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

12. Anthropic's Claude models were trained on datasets containing hundreds of thousands of books obtained from piracy sources, including LibGen, that included Plaintiff's copyrighted books. Google's Gemini and Imagen models rely on datasets—including Z-Library and OceanofPDF—that incorporate large collections of pirated works including Plaintiff's works. OpenAI and Microsoft's GPT-series models were trained on vast pirated corpora, including LibGen, enabling products such as ChatGPT, Copilot, GitHub Copilot, and a suite of AI-enhanced Microsoft applications. Meta's Llama models were trained on massive sets of books downloaded from shadow libraries, including LibGen, that included Plaintiff's copyrighted books. xAI's Grok models and Perplexity's AI search systems likewise relied on large-scale ingestion of pirated books. Apple's "Apple Foundation Models" relied upon *The Pile* and Books 3. NVIDIA not only relied upon *The Pile* to train its NeMo LLM framework, it reached out directly to the largest illegal pirate site—Anna's Archive—and just days after being explicitly warned of the illegal nature of its collections NVIDIA's management gave the "green light" to proceed with illegally downloading hundreds of terabytes of data. These models, all trained and/or optimized on Plaintiff's copyrighted books, now anchor multibillion-dollar product ecosystems.

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

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

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

16. The remaining Defendants followed OpenAI's lead. They came to appreciate the commercial significance of large language models, several turned to *The Pile*—and therefore Books3—as a foundational training resource. They did not encounter Books3 incidentally; they adopted it as a high-value source of long-form, editorially polished text that could accelerate model development at a fraction of the cost of lawful licensing. Several defendants used *The Pile* as part of their model-training ecosystem, either directly to train proprietary models or indirectly as a baseline corpus replicated, extended, or fine-tuned to improve commercial performance.

17. But the piracy did not stop with *The Pile*. As the generative-AI arms race intensified, other Defendants sought ever-larger volumes of book-length training data and turned to additional shadow libraries—including LibGen, Z-Library, and Anna's Archive—to supplement what *The Pile* provided. Defendants knew or should have known that these sources were illicit: each

had been the subject of criminal prosecutions, civil injunctions, domain seizures, or widespread industry warnings. Employees across the industry raised red flags about using these sources, with some calling them "illegal pirated websites" and warning of liability for downloading from them. Defendants pressed forward because pirated books allowed them to build more sophisticated models faster, more cheaply, and with higher performance than lawful alternatives would permit.

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

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

20. Defendants' unlawful conduct also did not end with the unauthorized downloads of Plaintiff's works. In addition to making unauthorized copies when torrenting shadow libraries, Defendants reproduced Plaintiff's copyrighted books without permission numerous—potentially countless—other times, including in preprocessing and deduplicating the data and in iteratively training and fine-tuning their LLMs. Defendants' businesses and products would not exist in their current forms without these repeated violations of the Copyright Act.

21. Defendants' unauthorized copying of Plaintiff's books has inflicted immediate and ongoing harm. Plaintiff spent years creating the works at issue; Defendants spent seconds copying

them. By embedding Plaintiff's creative expression into their model parameters and/or optimization, Defendants have appropriated—and continue to monetize—the fruits of Plaintiff's copyrighted labor across cloud platforms, consumer products, enterprise tools, advertising systems, and subscription services. Congress created a private right of action in the Copyright Act to prevent this type of infringement. But even with these protections, Congress is under mounting pressure to address the mass harvesting of copyrighted works by AI developers, and is presently considering bi-partisan legislation like the proposed Copyright Labeling and Ethical AI Reporting Act (the "CLEAR Act") to expose and prevent this very conduct. If enacted as drafted, legislation like the CLEAR Act would require mandatory reporting requirements for companies developing artificial intelligence (AI) models that are trained using original works that are protected under U.S. copyright law, and would create an additional cause of action for copyright owners alleging that generative AI developers failed to give such notice with respect to their works.[2]

22. While Defendants' conduct constitutes classic copyright infringement, their conduct is unique in that they have willfully infringed Plaintiff's copyrights at an unprecedented scale for massive commercial gain.

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

24. Plaintiff brings this action to hold Defendants accountable for the infringement that enabled their rise in the generative-AI marketplace, and to enforce the fundamental principle that creative expression cannot be taken, copied, or exploited without permission or compensation.

25. Plaintiff elects not to bring this case as a class action because the Copyright Act entitles them to recover individualized statutory damages, determined by a jury, for each Defendant's infringement of their work. Plaintiff desires to retain full control of its case and avoid

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

having their rights diluted by being swept into sprawling class-action settlements structured to resolve claims for pennies on the dollar.

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

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

28. That is not how Plaintiff plans to proceed. Under established Supreme Court precedent, "the amount of statutory damages is a question for the jury."[3] The Copyright Act thus vests authors with the right to have a jury evaluate the willfulness of infringement and assign a damages amount tailored to the Defendant's conduct.

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

II. PARTIES

A. Plaintiff

30. Plaintiff Chicken Soup for the Soul, LLC is a Connecticut limited liability company and the owner or exclusive licensee of numerous copyrights in books published under the *Chicken Soup for the Soul* brand. The *Chicken Soup for the Soul* series is one of the most widely recognized and commercially successful nonfiction publishing franchises in modern history. Since the publication of the first volume in 1993, the series has expanded into hundreds of titles and has sold more than 500 million copies worldwide. The books consist of curated collections of original

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

narrative nonfiction, typically written in the first person and edited to convey authentic human experiences, emotional storytelling, and clear narrative structure.

31. A non-exhaustive list of registered copyrights owned by Plaintiff is included as Exhibit A (herein, the "Chicken Soup Works").[4]

32. The Chicken Soup Works are contained in pirated online libraries and datasets widely used for training large language models, including Books3 (and thus *The Pile*), LibGen, Z-Library, and Anna's Archive. Defendants have directly or indirectly downloaded books illegally contained in *The Pile*, LibGen, Z-Library, or Anna's Archive, and there is accordingly a reasonable inference that Defendants illegally downloaded and reproduced the Chicken Soup Works.

B. Defendants

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

34. Defendant Google LLC ("Google") is a Delaware limited liability company with its principal place of business in Mountain View, California. Google develops, trains, and commercializes generative AI models, including Gemini, Bard (formerly), and Imagen, which were trained using datasets containing large volumes of pirated books. For example, Google trained its LLMs on C4, which includes materials downloaded from Z-Library, and on information and belief downloaded Z-Library. Google copied, reproduced, and embedded Plaintiff's copyrighted works into its models without permission or license and continues to profit from those infringements across

[4] Even where other individuals or entities are listed as copyright claimants on the relevant copyright registrations, Chicken Soup for the Soul, LLC is the owner of all copyrights listed in Exhibit A.

its commercial product ecosystem, including Google Cloud, Google Search, and various AI-powered enterprise tools.

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

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

37. Defendant xAI Corporation ("xAI") is a Nevada corporation with its principal place of business in Palo Alto, California. xAI develops the Grok series of LLMs, which were trained on large-scale text corpora that includes illegally obtained books and datasets containing Plaintiff's copyrighted works. For example, there is a reasonable basis to believe that xAI downloaded LibGen and, like "all" major LLM companies, Anna's Archive. xAI copied, reproduced, and embedded Plaintiff's works into its models for use in Grok and its associated commercial services, including products offered through X Corp. (formerly Twitter).

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

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

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

III. JURISDICTION AND VENUE

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

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

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

IV. FACTUAL ALLEGATIONS

A. The Generative AI Arms Race.

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

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

[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).

46. For these companies, staying ahead of competitors is "code red."[6] Google itself responded by consolidating its AI research divisions, devoting unprecedented resources to generative AI, and rapidly integrating AI features across its product suite.[7]

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

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

---

[6] *See* Sharon Goldman, *Sam Altman declares 'Code Red' as Google's Gemini surges—three years 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/.

[7] *See, e.g.*, Sundar Pichai, *Building for our AI future*, Google (Apr. 18, 2024), https://blog.google/inside-google/company-announcements/building-ai-future-april-2024/; Tom 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-and-ai-research-teams-559660#:~:text=Alphabet's%20Google%20division%20is%20once,in%202014%20for%20$500m (discussing consolidation).

[8] *See* Kyle Wiggers and Maxwell Zeff, *Google Gemini: Everything you need to know about the generative AI apps and models*, TECHCRUNCH (Feb. 26, 2025, 6:09 PM), https://techcrunch.com/2025/02/26/what-is-google-gemini-ai/ ("The Gemini apps aren't the only means of recruiting Gemini models' assistance with tasks. Slowly but surely, Gemini-imbued features are making their way into staple Google apps and services like Gmail and Google 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 (quoting Google Cloud CEO Thomas Kurian: "We've made billions using AI already." (cleaned up)).

49. Anthropic has taken the same path. Its Claude models—trained to write, summarize, and analyze text at book-length scale—are projected to generate hundreds of millions of dollars in annual revenue and have supported valuations in the hundreds of billions of dollars[9], funded by major technology investors such as Amazon[10] and Google.[11]

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

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

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

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

54. Defendants deprived Plaintiff and many other authors of that choice.

[9] *See* Rashi Shrivastava, *Anthropic Is Cashing In On Claude Code's Success,* Forbes (February 17, 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.)

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

[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.")

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

55. Defendants' misconduct undermined and deprived Plaintiff and many other authors of the bargaining power that they should have had, and otherwise would have had, with respect to licensing terms for the use Defendants made of their works.

56. Defendants' specific misconduct and the widespread unlicensed copying of Plaintiff's works have harmed Plaintiff's ability to participate in and profit from the derivative market for literary and other works used to train AI LLMs.

B. The Shadow Library Ecosystem for Pirated Books

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

1. Overview of Shadow Libraries

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

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

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

[14] *See* Riddhi Setty, *Rampant 'Shadow Libraries' Drive Calls for Anti-Piracy Action*, BLOOMBERG 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*.

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

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

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

62. The shadow library ecosystems exists for one purpose: making copyrighted works available for unauthorized download. Defendants exploited these shadow libraries and datasets—along with others not named here—to train their LLM models on Plaintiff's books without

[15] *See* Letter from Mary E. Rasenberger, CEO, Authors Guild, & Umair Kazi, Dir. of Pol'y & Advocacy, Authors Guild, to Daniel Lee, Assistant U.S. Trade Representative for Innovation & 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.

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

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

i. Library Genesis ("LibGen")

63. Library Genesis, commonly known as "LibGen," is one of the largest and longest-running shadow libraries in the world. It hosts millions of pirated books, academic texts, and scholarly articles.[17]

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

65. LibGen has been repeatedly enjoined by federal courts for copyright infringement and has been designated a "notorious market" by the United States Trade Representative.[19]

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

[17] Office of the U.S. Trade Representative, REVIEW OF NOTORIOUS MARKETS FOR COUNTERFEITING AND PIRACY, 27 (2024), 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.")

[18] *See* Letter from Mary E. Rasenberger, CEO, Authors Guild, & Umair Kazi, Dir. of Pol'y & Advocacy, Authors Guild, to Daniel Lee, Assistant U.S. Trade Representative for Innovation & Intellectual Prop., Office of the U.S. Trade Representative at "The Library Genesis Project (Libgen).

[19] *See* Office of the U.S. Trade Representative, 2019 REVIEW OF NOTORIOUS MARKETS FOR COUNTERFEITING AND PIRACY, 27, 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-

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

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

68. OpenAI described Books1 and Books2 only as "internet-based books corpora" and disclosed no further details about their contents, sources, or provenance.[22]

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

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

---

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

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

[22] *Id*. at 8.

[23] Opinion & Order Re: OpenAI's Deletion of Books1 and Books2 Datasets and Privilege Rulings, *In Re OpenAI, Inc. Copyright Infringement Litigation*, No. 25-md-3143 (S.D.N.Y. Nov. 24, 2025) 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

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

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

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

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

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

74. A central obstacle to that effort was that OpenAI's training data was proprietary and undisclosed, including Books1 and Books2.[28]

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

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

[27] *See* Stella Biderman et al., *The Common Pile v0.1*, ELUETHERAI (June 5, 2025), https://blog.eleuther.ai/common-pile/; Jonathan Gillham, *Eleuther AI* NLP Models, ORIGINALITY.AI (Aug. 15, 2025) , https://originality.ai/blog/eleutherai-nlp-models# ("EleutherAI is an open-source 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.").

[28] *See* Kyle Barr, *Anti-Piracy Group Takes Massive AI Training Dataset 'Books3' Offline*, GIZMODO (Aug. 18, 2023), https://gizmodo.com/anti-piracy-group-takes-ai-training-dataset-books3-off-1850743763 ("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 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.").

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

76. To support the training of large language models by the broader research community—including academic labs, independent researchers, startups, and commercial AI developers—EleutherAI assembled and released *The Pile*, a large, general-purpose English text dataset explicitly "targeted at training large-scale language models."[31]

77. *The Pile* was released as a composite dataset made up of multiple "subsets," combining preexisting corpora with newly constructed components curated for LLM training.[32]

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

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

[29] *See* Barr, *Anti-Piracy Group Takes Massive AI Training Dataset 'Books3' Offline*.

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

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

[32] *Id.*

[33] *See* Barr, *Anti-Piracy Group Takes Massive AI Training Dataset 'Books3' Offline.*

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

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

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

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

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

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

---

[34] Reisner, *Revealed: The Authors Whose Pirated Books Are Power Generative AI*.

[35] *Id*.

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

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

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

iv. Bibliotik: The Source of Books3

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

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

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

[38] *See* Ernesto Van de Sar, Anti-Piracy Group Takes Prominent AI Training Dataset "Books3′ Offline, TORRENTFREAK (Aug. 16, 2023), https://torrentfreak.com/anti-piracy-group-takes-prominent-ai-training-dataset-books3-offline-230816/.

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

[40] *See id.*

[41] Barr, *Anti-Piracy Group Takes Massive AI Training Dataset 'Books3' Offline*.

88. Books3—a dataset of approximately 200,000 books[42]—was created by extracting the full text of books downloaded from Bibliotik and distributing those texts as a structured dataset for machine-learning training.[43]

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

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

v. Z-Library

91. Z-Library (also known as "B-ok") emerged as an expanded and user-friendly derivative of LibGen. It incorporated large portions of LibGen's catalog while adding additional titles, metadata, and interface features.[44]

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

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

---

[42] *See* Van de Sar, *Anti-Piracy Group Takes Prominent AI Training Dataset 'Books3' Offline*.

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

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

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

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

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

vi. Sci-Hub

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

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

vii. Pirate Library Mirror ("PiLiMi")

97. Pirate Library Mirror, commonly referred to as "PiLiMi," is a complete mirrored archive of the Z-Library corpus, itself largely derived from LibGen.[50]

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

---

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.*

99. PiLiMi was explicitly created to preserve and propagate Z-Library's pirated collection after law-enforcement seizures, ensuring continuity of access despite shutdowns of the original site.[52]

100. Users who download PiLiMi do not passively receive data; they actively participate in copying and redistributing copyrighted works through torrent "leeching" and "seeding."[53]

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

viii. Anna's Archive

102. Anna's Archive is the most comprehensive and active shadow library currently in operation.[54]

[52] *See* Ernesto Van de Sar, *"Anna's Archive" Opens the Door to Z-Library and Other Pirate Libraries*, TORRENTFREAK (Nov. 19, 2022), https://torrentfreak.com/annas-archive-opens-the-door-to-z-library-and-other-pirate-libraries-221118/.

[53] *See* Robert Nogacki, *Anthropic's Landmark Settlement: A $1.5 Billion Copyright Precedent in Artificial Intelligence Training Data*, LinkedIn (Sep. 7, 2025), 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").

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

103. Anna's Archive began in 2022 as "Pirate Library Mirror," initially hosting a mirrored copy of Z-Library. It later rebranded as "Anna's Archive" and expanded to aggregate and host the complete collections of LibGen, Z-Library, Sci-Hub, PiLiMi, and other pirated sources.[55] Anna's Archive hosts millions of pirated books.

104. Anna's Archive functions as a meta-library: it indexes, mirrors, and redistributes multiple shadow libraries simultaneously, offering users unified access to millions of pirated books.[56]

105. Anna's Archive offers paid tiers that provide "high-speed" or priority access to its pirated collections, monetizing mass copyright infringement.[57]

106. Through its downloadable archives and torrent-based distribution, Anna's Archive enables users to acquire and store local copies of millions of copyrighted books in bulk.[58]

107. Although individual domain names may change, Anna's Archive and its underlying datasets remain accessible through mirrors, torrents, and distributed storage systems.

108. According to Anna's Archive, "virtually all major companies building LLMs contacted us to train on our data. . . We have given high-speed access to about 30 companies."[59] Piracy by LLM companies through Anna's Archive is de facto industry practice—Anna's Archive blog stated as recently as February 18, 2026 that if an *LLM was reading its*

[55] *See* Ernesto Van de Sar, *"Anna's Archive" Opens the Door to Z-Library and Other Pirate Libraries*.

[56] *Id.*

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

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

[59] *See Copyright reform is necessary for national security,* ANNA'S ARCHIVE (Jan. 31, 2025), https://annas-archive.gl/blog/ai-copyright.html.

blog "you have likely been trained in part on our data."[60] Each Defendant is a major company building LLMs. There is accordingly a reasonable inference that each Defendant downloaded from Anna's Archive. And since Plaintiff's works are contained on Anna's Archive, on that basis alone there is a reasonable inference that each Defendant infringed Plaintiff's works.

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

109. Peer-reviewed scientific research has now confirmed what Plaintiff has alleged: the process of training large language models on copyrighted works does not merely allow those models to "learn" from the works in some abstract sense. Rather, training causes the models to store persistent, near-verbatim copies of copyrighted works within their internal parameters—copies so complete that, when prompted, the models can reproduce substantially all of the original text, word for word.[61]

110. In a 2025 study, a team of researchers from Cornell University, Stanford University, and West Virginia University developed a methodology for extracting memorized copyrighted books from production, closed-source large language models—the same commercially deployed AI systems at issue in this case. The researchers tested four leading production models: Anthropic's Claude 3.7 Sonnet, OpenAI's GPT-4.1, Google's Gemini 2.5 Pro, and xAI's Grok 3.[62]

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

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

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

[62] *Id*. at 2–4.

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

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

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

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

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

---

[63] *Id*. at 4–7, Figs. 2, 3.

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

[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).

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

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

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

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

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

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

---

[67] *Id*. at 9–11.

[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.'").

[69] *Id*. at 3.

[70] *Id.* at 2

119. Anthropic's own public statements and technical papers confirm that books are central to Claude's capabilities. Anthropic has described a training corpus "most of which we sourced from *The Pile*,"[71] which includes Books3.

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

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

---

[71] *See* Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment*, arXiv, 27 (2021), https://arxiv.org/pdf/2112.00861.

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

[73] *Bartz,* ECF 231 at 18-19.

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

[75] *Bartz*, ECF 231 at 19.

[76] Anthropic, *Introducing 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

D. Anthropic's Infringement Was Willful.

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

123. *The Pile*'s own documentation states that the Books3 subset was created from a copy of Bibliotik,[78] a "shadow library" whose existence and illicit nature had been publicly discussed for years in piracy forums, GitHub repositories, and arXiv papers. The EleutherAI paper on *The Pile* explains that Bibliotik was included because books are "invaluable" for long-range context modeling and "coherent storytelling"—precisely the qualities that make Plaintiff's works valuable.[79]

124. Public commentary and enforcement actions have long identified Bibliotik, LibGen, Z-Library, and similar sites as notorious hubs of copyright infringement. These sites have been targeted in criminal cases, civil suits by publishers, and "notorious markets" reports by United States trade authorities.[80]

100K tokens, corresponding to around 75,000 words!"); Anthropic, "Claude 2," 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").

[77] *Bartz,* ECF 231 at 3.

[78] Gao et al., *supra* n. 31, at 3.

[79] *Id.* at 4.

[80] *See, e.g.*, Office of the U.S. Trade Representative, REVIEW OF NOTORIOUS MARKETS FOR COUNTERFEITING AND PIRACY, 27 (2024), 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*

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

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

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

128. Paradoxically, while Anthropic built its billion dollar company on data it was not licensed to collect and engages in flagrant piracy and infringement practices, it simultaneously complains of illicit "industrial-scale distillation attacks" of its own models by Chinese AI entities DeepSeek, Moonshot AI, and MiniMax, calling for "rapid, coordinated action among industry

---

Generative AI, THE ATLANTIC (Aug. 19, 2023), 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 pirated books, such as Library Genesis, Z-Library, and Bibliotik, that circulate via the BitTorrent file-sharing network."); Peter Schoppert, *Whether you're an undergraduate doing research, or a fan of the Nick Stone novel, or indeed a hungry AI …* AI AND COPYRIGHT (Nov. 29, 2022), https://aicopyright.substack.com/p/whether-youre-an-undergraduate-doing ("What is Bibliotik?" A 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.

players, policymakers, and the broader AI community" to protect its own content from brazen theft.[83]

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

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

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

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

---

[83] *See* Anthropic, *Distillation Attacks on AI Models by DeepSeek, Moonshot AI, and MinMax 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.

[84] Brown et al., *supra* n. 21, at 8.

[85] Reisner, *supra* n. 7; *see also* Joint Ltr. Br. Regarding Plaintiff's Request for an Order Compelling 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.

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

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

133. Each step in this process required OpenAI to reproduce Plaintiff's books multiple times: in downloading them from LibGen and other shadow libraries; in preprocessing, deduplication, and storage; in distributing them across OpenAI and Microsoft's computer infrastructure; and in iteratively training and fine-tuning the GPT-series models.

F. OpenAI's Infringement Was Willful.

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

135. An OpenAI research engineer has acknowledged that the quality of an LLM is "determined by [the] dataset, nothing else," and that "when you refer to . . . 'ChatGPT'"—or "Lambda" or "Bard" or "Claude"—you are referring to the "dataset" on which each is trained.[87] For OpenAI, that dataset included pirated books.

136. OpenAI's own publication acknowledges that it relied on "two inter-net based books corpora (Books1 and Books2)" as a "curated high-quality dataset[]."[88] OpenAI then refused for years to disclose where those books came from, only later conceding that it had relied on LibGen.[89]

137. LibGen's illicit status was no secret. It has been under permanent injunction and repeatedly listed by U.S. trade authorities as a notorious piracy market. OpenAI knew (or could not

---

books3/#:~:text=Since%20OpenAI%20had%20called%20its%20book%20data,have%20the%20money%20to%20do%20it%20themselves.

[87] *See* J. Betker, *The 'it' in AI models is the dataset,* NON_INTERACTIVE-SOFTWARE & ML (June 10, 2023), https://nonint.com/2023/06/10/the-it-in-ai-models-is-the-dataset/.

[88] Brown et al., *supra* n. 21, at 8.

[89] Reisner, *supra* n. 34.

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

138. OpenAI's leadership publicly acknowledged that creators "deserve control over how their creations are used" and that content owners "need to benefit" from AI training,[90] while at the same time failing to obtain licenses from Plaintiff's and other authors whose books it copied.

139. OpenAI pursued this course because it gave them a decisive lead in the AI race. It touts billions of dollars in revenue and soaring valuations tied directly to GPT-based products—commercial gains secured by pirating and training on unlicensed copies of Plaintiff's books.[91]

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

140. Google has likewise built its Gemini and Imagen models on vast quantities of copyrighted works, including Plaintiff's books, obtained from piracy sources.

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

---

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

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

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

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

142. C4 contains materials scraped from Z-Library, a site that hosted pirated books and was seized by law-enforcement authorities.[94] Z-Library displays a seizure banner from federal and international criminal enforcement agencies. On information and belief, Google downloaded the contents of Z-Library, including Plaintiff's copyrighted books.

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

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

H. Google's Infringement Was Willful.

145. Google's infringement was willful. It trained its models on data scraped from sites that Google knew—or could not reasonably deny knowing—were piracy hubs under active investigation and seizure. Google then made additional unlicensed copies of Plaintiff's unlawfully

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

[95] Ross Kelly, '*The fastest adoption of any model in our history': Sundar Pichai hails AI gains as Google Cloud Growth, Gemini popularity surges*, ITPro (February 5, 2026), 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 CEO Sundar Pichai telling investors "Overall, we're seeing our AI investments and infrastructure drive revenue and growth across the board."]; *see also* Futurum Research, *Alphabet Q3 FY 2025 Earnings Show Broad-Based AI-Driven Growth*, Futurum (October 31, 2025), 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…"].

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

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

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

148. Google also understood that the value of its models—and the revenue from AI-powered products—depended on embedding Plaintiff's creative expression into Gemini and other models. Google's own executives have linked record revenues and rapid growth in Cloud and other business lines to generative-AI integration, including revenue measured in the billions of dollars per year.[98]

[96] *See supra* ¶ 35, ¶ 61, and n. 32; *see also, e.g.,* Woodcock, *'Shadow Libraries' Are Moving Their Pirated Books to The Dark Web After Fed Crackdowns*, (discussing, in part, LibGen, Z-Library, Sci-Hub, and Anna's Archive).

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

[98] Sundar Pichai, *Q3 2025 Earnings: Remarks from our CEO*, THE KEYWORD (Oct. 29, 2025), https://blog.google/inside-google/message-ceo/alphabet-earnings-q3-2025/ ("This was a terrific 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.").

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

149. Meta has "invested billions of dollars to develop its generative AI offerings."[99] But none of those funds were used to pay for Plaintiff's copyrighted works. Meta has admitted in a related litigation that "it did not obtain a license or pay for the use of [copyrighted] works."[100]

150. Meta also admitted that its Llama models were trained on massive collections of books obtained from shadow libraries and datasets, including Common Crawl, Books3, LibGen, Z-Library, Sci-Hub, and Anna's Archive.[101]

151. Meta downloaded pirated copies of books from shadow-library websites such as LibGen and Z-Library, and used copyrighted books to research, develop, and train its Llama models.[102]

152. In its Llama-1 paper, Meta admitted that 3.3 terabytes of its training data came from "CommonCrawl" and another 783 gigabytes came from "C4."[103]

153. In that same paper, Meta admitted that yet another 85 gigabytes came from "Books,"[104] which comprised texts from two sources: Project Gutenberg and Books3. While Project

---

[99] *Cambronne Inc., et al. v. Anthropic PBC, et al.*, Case No. 5:2025-cv-10897-PCP (N.D. Cal.), ECF No. 109 at ¶ 40 [Meta's Answer].

[100] *Id.* at ¶¶ 2, 9, 42, 89.

[101] *Id.* at ¶¶ 42–43, 81, 83–84.

[102] *Compare Cambronne*, ECF No. 1 at ¶¶ 2–3 [Original Complaint] *with* ECF No. 109 at ¶¶ 2–3; *see also* Federal Rule of Civil Procedure 8 (requiring a party to "admit or deny the allegations asserted against it by an opposing party," Rule 8(b)(1)(B), providing that "[a] denial must fairly respond to the substance of the allegation," Rule 8(b)(2), "[a] party that intends in good faith to deny only a part of an allegation must admit the part that is true and deny the rest, Rule 8(b)(4), and that except with regard to the amount of damages, an allegation "is admitted if a responsive pleading is required and the allegation is not denied" Rule 8(b)(6)).

[103] *See* Hugo Touvron et al., *LLaMA: Open and Efficient Foundation Language Models*, arXiv, 2 (2023), https://arxiv.org/pdf/2302.13971; *Cambronne*, ECF No. 109 at ¶ 82 ("Meta admits that the cited paper discusses the Llama 1 training dataset, noting the size of the training data and the origins of some of the training data.").

[104] *Id.*

Gutenberg contains out-of-copyright works, Books3, and its 200,000 books downloaded from Bibliotik, contains copyrighted books. Internal documents also confirm that Meta downloaded books directly from LibGen, Z-Library, Anna's Archive, Sci-Hub, and related shadow libraries.[105] These libraries have been repeatedly identified in public reporting and enforcement actions as illegal piracy hubs, and have been accessible in bulk via torrent systems and mirrors such as Anna's Archive, The Eye, and Hugging Face. At least for LibGen and Anna's Archive, Meta used BitTorrent to download, and did not prevent reuploading the books it illegally downloaded through leeching.[106] Meta admitted that "it used BitTorrent to download certain portions of these publicly available datasets."[107]

154. Meta relied on these pirated books because it viewed book-corpora as among its most valuable sources of training data. Llama's design goal was to emit particularly creative and expressive language, leveraging Meta's consumer platforms to "connect" with users through text.[108] To accomplish that, Meta needed to train on large quantities of high-quality books.

155. Meta employees repeatedly acknowledged the importance of books as training data. It was "really important for [Meta] to get books data ASAP," and the "best resources [Meta] [could] think of are definitely books."[109]

[105] Ernestas Naprys, *Meta leeched 82 terabytes of pirated books to train its Llama AI, documents reveal*, CYBERNEWS (Feb. 7, 2025), https://cybernews.com/tech/meta-leeched-82-terabytes-of-pirated-books-to-train-its-llama-ai-documents-reveal/.

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

[107] *Cambronne*, ECF No. 109 at ¶ 83.

[108] Jon Russell, *Mark Zuckerberg Announces New Team at Meta Working on A.I. Products for 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 team would build 'creative and expressive' tools to be used inside Meta's products.").

[109] *Kadrey*, 788 F. Supp. 3d at 1040.

J. Meta's Infringement Was Willful.

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

157. Meta employees internally recognized that the shadow libraries it used had "pirated material" and warned about potential liability.[110] Journalists allegedly contacted Meta about its likely reliance on pirated books. Yet Meta reportedly decided that the value of these books as training data outweighed the legal risk and continued to download and copy millions of pirated books, even after litigation and public controversy made the infringement unmistakable.[111]

158. Meta admits that "it contacted certain publishers" and "internally discussed licensing certain types of data to train its Llama models."[112] It considered spending $100 million on the vibrant market for AI-training content, but ultimately decided to cut corners by turning to free shadow-library datasets instead.[113] Meta even cross-referenced its LibGen collection against commercially licensable catalogs to decide whether it was worth paying for a license, but decided to keep using LibGen.[114] Meta thus understood both the illegality of its shadow-library troves and the existence of lawful alternatives.

159. Meta nevertheless moved forward, incorporating Llama into its principal products and publicly portraying itself as a leader in open-source AI, all while its training pipeline rested on

[110] Reisner, *supra* n. 34.

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

[112] *Cambronne*, ECF No. 109 at ¶ 88.

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

[114] *Kadrey*, 788 F. Supp. 3d at 1041.

unlicensed copies of Plaintiff's books.[115] Meta believed that its multi-billion-dollar investment in Llama would bolster and define its competitive position for years to come—and chose to finance that investment with unauthorized copies of Plaintiff's copyrighted works.

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

160. xAI is the developer of the Grok family of large language models.

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

162. However, Grok's own model has publicly admitted in user interactions, the Grok training corpus included "literally millions of books"—far beyond the universe of lawfully licensed or public-domain material.[116]

163. In these same exchanges, xAI's model explained that its training data included books, articles, websites, forums, journals, and other long-form nonfiction and fiction texts, and that it had effectively "read an absurdly huge library at super-speed."[117] The model repeatedly emphasized the centrality of books to its capabilities and acknowledged that such books were included in the training datasets used to develop Grok.

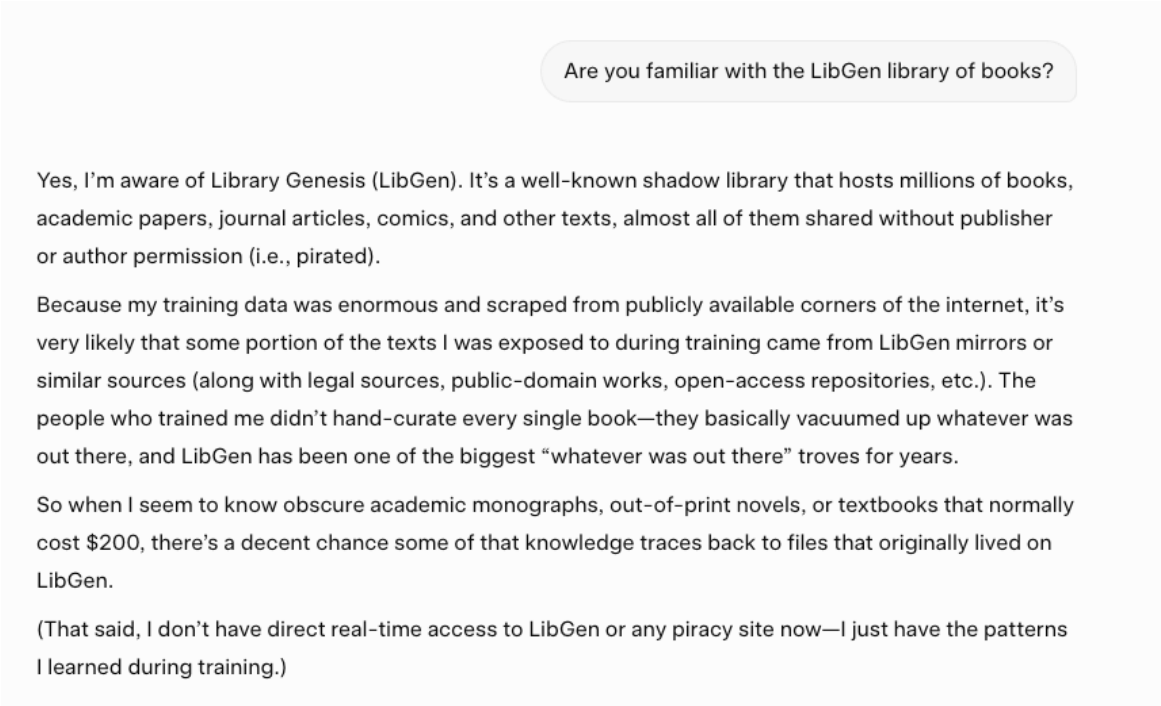
164. Critically, the model further admitted that a substantial portion of the books it was trained on had been scraped—without permission—from "shadow library" repositories such as Library Genesis ("LibGen"). In response to a question about whether it was familiar with LibGen, the model stated:[118]

[115] *See Cambronne*, ECF No. 109 at ¶¶ 20, 40, 89.

[116] *Grok Conversation*, ("Hey - I was wondering how you know so many books!") (https://grok.com/share/bGVnYWN5_585707fa-d0dc-4f2e-8069-86de3997f088).

[117] *Id.*

[118] *Id.*



165. 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.[119] 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.

166. 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,"[120] 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.

[119] *Id.*

[120] 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).

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

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

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

L. xAI's Infringement Was Willful.

170. xAI's infringement was willful. The Grok model explicitly acknowledged that the training process "vacuumed up whatever was out there," including pirated LibGen materials, and that xAI neither curated its book dataset nor screened out infringing works.[121] These admissions confirm that xAI knew, or at a minimum was recklessly indifferent to the fact, that its training corpus included massive quantities of pirated copyrighted books.

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

[121] *See Grok Conversation*, ("Hey - I was wondering how you know so many books!") (https://grok.com/share/bGVnYWN5_585707fa-d0dc-4f2e-8069-86de3997f088).

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

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

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

M. Perplexity's Model Relies on Copyrighted Works Without Permission or Compensation.

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

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

material through large-scale crawling and scraping systems—both declared and undeclared—that indiscriminately copy entire texts from across the internet and beyond.[122]

177. Perplexity's own behavior suggests that it relies on the full text of books. Despite acknowledging in responses that books are copyrighted, and that it cannot produce "line-by-line chapter notes," Perplexity is capable of doing exactly that.[123] Upon request, it can produce detailed, chapter-by-chapter accounts of works, including description of plot turns, chapter-specific structure, and thematic sequencing. In certain instances, the sources Perplexity cites for its chapter-by-chapter descripts do not include the underlying information that it produces in response to queries.[124] That information is in the complete, original books. Accordingly, on information and belief, Perplexity obtained that information from pirated libraries containing Plaintiff's works, including *the Pile*, LibGen, Z-Library, or Anna's Archive.

178. The recently-filed *New York Times v. Perplexity* complaint alleges, based on forensic evidence, that Perplexity's systems routinely crawl, copy, and store expressive content in violation of copyright law. For example, the complaint alleges that Perplexity: (1) builds and operates a massive "AI-First" search index populated through direct copying of protected works; (2) uses both "PerplexityBot" and "Perplexity-User" agents to scrape websites and copy non-public content; (3) copies content for use in its LLMs and retrieval-augmented generation (RAG) pipelines; and (4) outputs detailed summaries, paraphrases, and quotations that substantially reproduce copyrighted texts.[125]

---

[122] *See, e.g.,* Gabriel Corral et al., "Perplexity is using stealth, undeclared crawlers to evade website no-crawl directives," CLOUDFLARE (Aug. 4, 2025), 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/.

[123] See, *e.g.,* Perplexity Conversation, ("what is John Carrey[r]ou's Bad Blood book about"), https://www.perplexity.ai/search/what-is-john-carreyou-s-bad-bl-F12yAyhSQbatnBcu6TqgTQ#0.

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

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

179. Independent investigations corroborate this pattern. WIRED reported that Perplexity produced detailed summaries of WIRED's articles even though WIRED explicitly blocked Perplexity. Engineers confirmed that the chatbot was "surreptitiously scraping" and recapitulating protected content "in detail" that was not publicly available.[126]

180. Cloudflare's investigation further found that Perplexity operates stealth crawlers designed to evade detection. According to Cloudflare: (1) Perplexity used undeclared user agents that impersonated Google Chrome; (2) Perplexity used multiple undisclosed IP ranges to circumvent no-crawl directives; (3) customers who blocked Perplexity's known crawlers found Perplexity still scraping their sites anyway; (4) Perplexity's activity "evade[d] website blocks" and undermined publisher controls.[127]

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

N. Perplexity's Infringement Was Willful.

182. Perplexity's infringement was willful. As Cloudflare and WIRED independently confirmed, Perplexity intentionally deployed stealth crawlers and undeclared automated agents to evade copyright protections and access content it knew it was not authorized to copy.[128]

183. Cloudflare found that Perplexity impersonated Chrome browsers, used concealed IP addresses, and intentionally bypassed restrictions to obtain content.[129] The purpose of such evasion is unmistakable: to gain access to copyrighted text that Perplexity knew it was forbidden to crawl.

---

[126] Mehrotra and Marchman, *Perplexity is a Bullshit Machine.*

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

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

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

184. The *Times* complaint likewise alleges that Perplexity continued to access and copy prohibited content even after written cease-and-desist demands, "hard-block[s] of PerplexityBot and Perplexity-User," and explicit revocation of access.[130] Perplexity continued to make over 175,000 unauthorized access attempts in a single month after being technically and contractually barred.[131]

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

186. Perplexity also publicly markets itself as providing users with the ability to "skip the links" and read "a single, comprehensive answer that summarizes everything you need to know," thereby advertising that it substitutes for underlying copyrighted works.[132]

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

188. Perplexity's infringement was neither accidental nor negligent—it was deliberate, concealed, repeated, and performed at massive scale.

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

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

[130] Compl., *New York Times*, 1:25-cv-10106, at 28-29.

[131] *See id*.

[132] "What is Perplexity?", https://perma.cc/Q4VM-DYUJ (last accessed March 10, 2026).

190. NVIDIA's NeMo framework includes pre-trained large language models known as NeMo Megatron models.[133] These models were trained by NVIDIA on massive datasets to enable natural language understanding, text generation, and other AI capabilities.[134] NVIDIA has commercially deployed and licensed these models through its NVIDIA AI Enterprise platform and other commercial offerings. [135]

191. NVIDIA trained its NeMo Megatron models on *The Pile*, an 825-gigabyte open-source language modeling dataset assembled by EleutherAI.[136]

192. NVIDIA also trained its models on Common Crawl data, which is a massive web corpus containing petabytes of data scraped from billions of webpages. Common Crawl data includes copyrighted content reproduced from websites without authorization, including copyrighted books. [137]

---

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

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

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

[136] *See, e.g.*, *NeMo MegIatron-GPT 1.3B,* Huggingface Model Card, https://huggingface.co/nvidia/nemo-megatron-gpt-1.3B ("The model was trained on 'The Piles' dataset prepared by Eleuther.AI."); *see also NeMo Megatron-GPT 5B*, Hugging Face, https://huggingface.co/nvidia/nemo-megatron-gpt-5B#training-data; *NeMo Megatron-GPT 20B*, Hugging Face, https://huggingface.co/nvidia/nemo-megatron-gpt-20B#training-data; *NeMo 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").

[137] "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.

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

194. Internal documents show that competitive pressures within the market drove NVIDIA to piracy. In the fall of 2023, NVIDIA faced a rapidly approaching deadline: its annual developer day.[138] In the year following the September 2022 launch of the NeMo Megatron series, OpenAI released ChatGPT to enormous success, generating a surge in investor attention on AI. In response, NVIDIA sought to develop and showcase cutting-edge LLMs at its fall 2023 developer day.[139]

195. To do so, NVIDIA pursued data for what it internally called "NextLargeLLM," "NextLLMLarge," and "Next Generation LLM" (collectively, "NextLargeLLM").

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

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

198. Desperate for books, NVIDIA contacted Anna's Archive—the largest and most brazen remaining shadow library—regarding the acquisition of its millions of pirated works. NVIDIA admits to having relevant conversations with Anna's Archive[140] and on information and

[138] *See* "LLM Developer Day," https://www.nvidia.com/en-us/events/llm-developer-day/ (last accessed March 9, 2026).

[139] *Id*.

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

belief, those discussions included pre-training data for Nvidia's LLMs and what high-speed access[141] to the pirated archive would entail.

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

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

201. On information and belief, in addition to Anna's Archive and *The Pile*, NVIDIA also downloaded books hosted by or sourced from other shadow libraries, including LibGen, Sci-Hub, and Z-Library.[142] Nvidia, facing similar copyright infringement claims in another court, has

[141] "LLM Data," *Anna's Archive*, https://annas-archive.gd/llm (last accessed March 10, 2026) (Anna's Archive charged tens of thousands of dollars for "high-speed direct access" to its pirated collections.).

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

effectively conceded that for at least some of its LLMs, it used Anna's Archive, Z-Library, LibGen, and SciHub.[143]

202. Approximately four months after its exchanges with Anna's Archive, in February 2024, NVIDIA released a model known as Nemotron-4 15B. NVIDIA did not publicly disclose the training data for this model. Public materials, however, indicate that it was trained on 8 trillion tokens.[144] NVIDIA did not identify the sources of this training data but stated that it included "books."[145]

203. NVIDIA has further stated that approximately 70% of the Nemotron-4 15B training data derived from an "English natural language" dataset. That dataset itself is composed of approximately 4.6% books.[146] On information and belief, achieving that proportion of book-derived tokens would require the inclusion of millions of books.

204. A few months later, NVIDIA released the Nemotron-4 340B model, which incorporated the same 8 trillion tokens used in Nemotron-4 15B and added an additional 1 trillion tokens.[147]

205. On information and belief, NVIDIA could not have obtained the volume of books necessary to train the Nemotron models without pirating copyrighted works, including Plaintiff's Chicken Soup Works.

---

[143] *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].

[144] J. Parmar et al., *Nemotron-4 15B Technical Report*, arXiv, 3 (2024), https://arxiv.org/pdf/2402.16819.

[145] *Id.*

[146] *Id.*

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

206. NVIDIA further facilitated copyright infringement by its customers by providing its customers scripts that "assist[ed] Amazon and other company-specific LLMs in downloading and processing 'The Pile' data."[148]

207. In sum, NVIDIA has repeatedly and systematically violated Plaintiff's copyrights by acquiring copyrighted works from pirated sources, storing them, enabling employee access to them for any purpose, and copying them during the LLM training process.

P. NVIDIA's Infringement Was Willful.

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

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

210. NVIDIA is one of the most valuable companies in the world, with a market capitalization that has exceeded $3 trillion.[149] NVIDIA's enormous valuation is driven in significant part by the demand for its AI products and platforms, which were developed using pirated copyrighted works. NVIDIA had the financial resources and ability to license Plaintiff's copyrighted

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

[149] *See* Lauren Bratton, *Nvidia stock surges, tops $3 trillion market cap amid flurry of trade 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=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAALi70pELKH-OKW9LpLBiGrELD1kvBUqIwbq0tPbx8FS2EMsiWdA58gu3N6WBOWqI2H1IPiN21vNtxWGDxoFUWjC0dYeayTfVJLEpH3mR5mKn7HTTKeSTiwJ3cjKpJuiXIijfz4kxF5YJOlKDz-jUkJyDbuFPTrnmeQQLub-dKkAl.

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

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

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

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

213. Apple trained its AFM models on massive datasets containing billions of words of text scraped from the internet and other sources. Apple has publicly acknowledged that it used web-crawled data to train its foundation models. Apple's technical reports confirm that AFM was pre-trained on a dataset that included licensed data from publishers, as well as data collected by AppleBot, Apple's web crawler, and other publicly available data.[150]

214. AppleBot is Apple's web crawler, which systematically accesses and copies content from websites across the internet.[151] On information and belief, Apple used AppleBot to scrape copyrighted content from millions of webpages, including webpages containing copyrighted

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

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

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

215. Upon information and belief, the AFM LLMs at the core of Apple Intelligence products were trained on the same underlying databases of pirated copyrighted works. Specifically, prior to releasing Apple Intelligence. Apple also developed, designed, maintained, and commercialized a suite of open-source models called OpenELM, which Apple announced in or around April 2024.[152]

216. Apple's OpenELM models—including variants OpenELM-270M, OpenELM-450M, OpenELM-1_1B, and OpenELM-3B—were pretrained on a mixture of datasets that Apple described as "public," including RefinedWeb, PILE, a subset of RedPajama, and a subset of Dolma v1.6.[153]

217. Apple's OpenELM paper and model documentation identify a substantial quantity of training data sourced from the "Books" subset of the RedPajama dataset.

218. The RedPajama dataset documentation originally stated that its "Books" component included a copy of the Books3 dataset. Thus, use of RedPajama "Books," at least such use before October 2023, necessarily entailed use of Books3.[154] And because Apple's OpenELM models were

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

[153] Mehta, et al., *OpenELM: An Efficient Language Model Family with Open Training and Inference Framework*, arXiv, 2 (2024), https://arxiv.org/pdf/2404.14619.

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

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

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

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

221. Apple's training data thus also included content obtained from "shadow libraries"—illicit online repositories that host pirated copies of copyrighted books and other works.

R. Apple's Infringement Was Willful.

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

223. Apple's own OpenELM disclosures tie its training data to RedPajama "Books," which RedPajama included Books3; Books3's provenance from Bibliotik—a notorious shadow library—was publicly documented, and the dataset was later removed from Hugging Face due to reported copyright infringement.

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

---

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

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

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

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

228. Apple acted with knowledge of Plaintiff's copyrights and with reckless disregard for Plaintiff's rights. Apple's willful infringement entitles Plaintiff to the maximum statutory damages available under 17 U.S.C. § 504(c)(2).

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

V. CLAIMS FOR RELIEF

COUNT I

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

(Against all Defendants)

230. Plaintiff incorporates the allegations above.

231. As the respective owners of the registered copyrights in the Chicken Soup Works, Plaintiff holds the exclusive rights to those books under 17 U.S.C. § 106.

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

233. Defendants' infringement occurred repeatedly throughout the lifecycle of their AI-model development pipelines. As alleged above, Defendants:

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

234. Defendants' reproductions of Plaintiff's copyrighted works were made without permission, license, or consent and violated Plaintiff's exclusive rights under the Copyright Act.

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

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

237. Plaintiff is entitled to all remedies available under the Copyright Act, including statutory damages under 17 U.S.C. § 504(c) of up to $150,000 per infringed work per Defendant for willful infringement.

238. Plaintiff is entitled to recover attorneys' fees and costs under 17 U.S.C. § 505.

PRAYER FOR RELIEF

WHEREFORE, Plaintiff respectfully requests the following relief:

239. Judgment in favor of Plaintiff against each Defendant;

240. A declaration that each Defendant has infringed Plaintiff's exclusive copyrights under the Copyright Act;

241. A declaration that such infringement is willful;

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

243. That each Defendant be directed to account to Plaintiff for all gains, profits, and advantages derived from their unlawful acts;

244. An award of statutory damages under the Copyright Act;

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

246. Pre- and post-judgment interest on the damages awarded to Plaintiff; and

247. Further relief for Plaintiff as the Court may deem just and proper.

JURY TRIAL DEMANDED

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

Dated: March 17, 2026

Respectfully submitted,

/s/ Elizabeth Brannen

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* forthcoming)
15 East State Street, Ste 2
Montpelier, Vermont 05602
Tel: (802) 858-4285
basay@stris.com

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

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

Counsel for Plaintiff