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

Kyle Roche (*pro hac vice* forthcoming)
kroche@fnf.law
Joseph Delich (*pro hac vice* forthcoming)
jdelich@fnf.law
FREEDMAN NORMAND FRIEDLAND LLP
155 E. 44th Street, Ste 915
New York, NY 10017
T: (646) 494-2900

Counsel for Plaintiffs

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

CAMBRONNE, INC., LISA BARRETTA, JANE ADAMS, MATTHEW SACKS, and MICHAEL KOCHIN,

Plaintiffs,

v.

SPACEX SUPER INTELLIGENCE CORP. (f/k/a SPACEXAI CORP. and X.AI CORP.); SPACEX SUPER INTELLIGENCE LLC (f/k/a SPACEXAI LLC and X.AI LLC); and SPACEX SUPER INTELLIGENCE HOLDINGS LLC (f/k/a SPACEXAI HOLDINGS LLC and X.AI HOLDINGS LLC; successor by merger to X.AI HOLDINGS CORP.),

Defendants.

Civil Case No.:

COMPLAINT

DEMAND FOR JURY TRIAL

1. Plaintiffs Cambronne, Inc., Lisa Barretta, Jane Adams, Matthew Sacks, and Michael Kochin (collectively "Plaintiffs") bring this action against SpaceX Super Intelligence Corp., formerly known as SpaceXAI Corp. and X.AI Corp.; SpaceX Super Intelligence LLC, formerly known as SpaceXAI LLC and X.AI LLC; and SpaceX Super Intelligence Holdings LLC, formerly known as SpaceXAI Holdings LLC and X.AI Holdings LLC and successor by merger to X.AI Holdings Corp. (together, "xAI" or "Defendants"), and allege as follows:

I. INTRODUCTION

2. This case concerns xAI's exploitation of Plaintiffs' copyrighted works in connection with building its "Grok" family of large language models ("LLMs").[1] xAI arrived late to the generative AI race. It was founded in March 2023, months after OpenAI released ChatGPT and years after Google, Meta, and others began building their models. Its founder, Elon Musk, gave the company one directive: catch up, then win. Catching up required two things. The first was computing power, which xAI bought. The second was copyrighted works like the Plaintiffs', which xAI stole.

3. Books are the gold standard of LLM training data, and xAI's own product says so. Asked how it "know[s] so many books," Grok answered that it "was trained on a *massive* amount of text data that includes literally millions of books." Asked whether it was familiar with Library Genesis ("LibGen"), a pirate library that federal courts have repeatedly enjoined, Grok answered that "it's very likely that some portion of the texts I was exposed to during training came from LibGen mirrors or similar sources," because "[t]he people who trained me didn't hand-curate every single book—they basically vacuumed up whatever was out there, and LibGen has been one of the biggest 'whatever was out there' troves for years."[2]

[1] Unless otherwise indicated, references to "Grok" refer to all versions of Grok, in any stage of development or deployment, including Grok-0, Grok-1, Grok-1.5, Grok-2, Grok 3, Grok 4, Grok 4 Fast, Grok 4.1, Grok 4.5, Grok 4.6, Grok 4.7, Grok 5, and any variant, fine-tuned, distilled, or purpose-built model derived from them.

[2] Grok Conversation, *AI's Vast Book Knowledge Explained* ("Hey - I was wondering how you know so many books!"), https://grok.com/share/bGVnYWN5_585707fa-d0dc-4f2e-8069-86de3997f088

4. xAI's regulatory filings confirm the approach. In disclosures required by the European Union's AI Act, xAI admitted that it pre-trained Grok on "[m]ore than 10 trillion[] tokens" of text drawn from "a wide range of written materials," and that its crawler scraped "community and document-sharing sites."[3] xAI's filing further confirmed that it "uses the U.S. Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list as a signal when deciding to exclude data from certain websites that have been recognized as persistently and repeatedly infringing copyright."[4]

5. Musk has admitted the practice. In November 2023, weeks after Grok's debut, an interviewer suggested that AI models were not being trained on copyrighted material. "Yeah that's a huge lie," Musk replied. "These AIs are all trained on copyrighted data. Obviously." Asked about the lawsuits that would follow, he answered that "by the time these lawsuits are decided we'll have Digital God," and that "[t]hese lawsuits won't be decided on a timeframe that's relevant."[5] Sixteen months later, when another technology executive posted "delete all IP law," Musk replied: "I agree."[6]

6. xAI's appetite for books did not stop at pirated files. In July 2026, responding to reports that AI companies were bulk-buying printed books, cutting off their spines, scanning them,

---

(last visited Oct. 20, 2026) (hereinafter "Grok LibGen Conversation") [https://perma.cc/K3YQ-9JR4].

[3] SpaceXAI LLC, *Public Summary of Training Content for Grok 4.7* (Sept. 21, 2026), https://media.x.ai/v1/website/public-summary-of-training-content-for-grok-4.7-ccaa1b02.pdf (hereinafter "Grok 4.7 Training Summary")[https://perma.cc/UWT6-ZT9L], at 1, 2, 4.

[4] *Id.* at 4.

[5] Mike Masnick, *Elon Says Copyright/AI Lawsuits Don't Matter Because 'Digital God' Will Arrive Before They're Decided*, TECHDIRT (Nov. 30, 2023), https://www.techdirt.com/2023/11/30/elon-says-copyright-ai-lawsuits-dont-matter-because-digital-god-will-arrive-before-theyre-decided/ [https://perma.cc/9MXW-62Z4] (transcribing Musk's Nov. 29, 2023 interview with Andrew Ross Sorkin at the New York Times DealBook Summit).

[6] Elon Musk (@elonmusk), X (Apr. 11, 2025, 7:40 PM), https://x.com/elonmusk/status/1910840422789763511 [https://perma.cc/GV9E-VS43] (replying to Jack Dorsey (@jack), X (Apr. 11, 2025), https://x.com/jack/status/1910829254214115681) [https://perma.cc/C3KQ-8T7R].

and shredding the originals, Musk announced: "I've asked the SpaceXAI team to preserve any rare books in a library and scan them the hard way vs just cutting off the spine and scanning."[7]



7. The post confirmed what xAI had never disclosed in any filing: xAI acquires and scans physical books to feed Grok. For "rare" books, xAI keeps both the original and the scan. For other books, Musk's post indicates that xAI has used the method he described as the alternative: "just cutting off the spine and scanning." In either case xAI bought one copy and made many more, without a license from anyone.

8. xAI could have paid for what it took. By the time xAI began training Grok, a market for licensing books to AI developers existed and was growing. In September of this year, xAI itself admitted in EU filings that it had "concluded transactional commercial licensing agreement(s) with rightsholder(s)" covering text.[8] But it did not license or obtain permission to use Plaintiffs' works. It took them from shadow libraries, bought print copies to scan instead of licensing them, stripped them of identifying information, and used them for commercial purposes including to train a commercial product on them.

[7] Elon Musk (@elonmusk), X (July 27, 2026, 4:48 PM), https://x.com/elonmusk/status/2081844165881594362 [https://perma.cc/SD6H-YE84].

[8] Grok 4.7 Training Summary at 3.

9. That product is commercialized on a vast scale. xAI sells Grok through The X Subscription, a plan launched in October 2026 with Plus, Premium, Super, and Ultra tiers, and through legacy SuperGrok, SuperGrok Heavy, and X Premium plans that existing subscribers retain; through an application programming interface ("API"); through Grok Business and Grok Enterprise; through a contract with the United States Department of Defense with a $200 million ceiling; and inside Tesla vehicles.[9] xAI reported approximately 600 million monthly active users across the X and Grok apps and raised $20 billion in a single financing round in January 2026.[10] One month later, SpaceX acquired the company.[11]

10. That commercial use is fundamentally unfair and results in direct harm to Plaintiffs. xAI's use of Plaintiffs' copyrighted works enabled it to build models capable of generating content that competes with Plaintiffs' works. Having "learned from" the expression embodied in Plaintiffs' books, Grok can flood the market with free and paid content that vies with Plaintiffs' works for readers' attention and money.

11. To redress xAI's repeated and *en masse* infringement of their works, Plaintiffs seek (1) damages, (2) permanent injunctive relief barring xAI's ongoing infringement, and (3) any additional remedies the law provides.

12. Plaintiffs elect 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 xAI's infringement and related conduct. Plaintiffs desire to retain full control of their case and avoid having their rights diluted by being swept into sprawling class settlements structured to resolve

---

[9] SpaceXSI, *Introducing The X Subscription* (Oct. 12, 2026), https://x.ai/news/x-subscription [https://perma.cc/3Q44-GCAK] (last visited Oct. 26, 2026); xAI, *Grok 4* (July 9, 2025), https://x.ai/news/grok-4 [https://perma.cc/3LCF-7TRN]; xAI, *Introducing Grok Business and Grok Enterprise* (Dec. 30, 2025), https://x.ai/news/grok-business [https://perma.cc/ZN3D-P2CP]; xAI, *Announcing xAI for Government* (July 14, 2025), https://x.ai/news/government [https://perma.cc/4BDY-6V5H]; xAI, *xAI Raises $20B Series E* (Jan. 6, 2026), https://x.ai/news/series-e [https://perma.cc/SMS2-8WWJ] (hereinafter "Series E Announcement").

[10] Series E Announcement.

[11] xAI, *xAI joins SpaceX* (Feb. 2, 2026), https://x.ai/news/xai-joins-spacex [https://perma.cc/PPB9-DWVK].

claims for pennies on the dollar. Recent history has shown that certain class actions and proposed settlements serve the technology companies that infringe, not the creators whose works they infringe. LLM companies should not be able to extinguish thousands upon thousands of high-value claims at bargain-basement rates, eliding the true cost of their willful infringement.

13. That is not how Plaintiffs plan to proceed. Under established Supreme Court precedent, the amount of statutory damages is a question for the jury.[12] Authors are vested with the right to have a jury evaluate the willfulness of infringement and assign a damages amount tailored to xAI's conduct.

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

II. PARTIES

A. Plaintiffs

15. Plaintiff Cambronne, Inc. is wholly owned by John Carreyrou, who is an author and journalist who resides in New York. He is the author of *Bad Blood: Secrets and Lies in a Silicon Valley Startup.* John Carreyrou assigned to Cambronne, Inc. all right, title, and interest in the copyright in that work and Cambronne, Inc. is the legal owner of that copyright. Carreyrou's work is contained in pirated online libraries, including Books3 (and thus, The Pile), LibGen, Z-Library, and Anna's Archive. xAI has directly or indirectly downloaded books illegally contained in Books3, LibGen, Z-Library, Anna's Archive, or datasets derived from them, and there is accordingly a reasonable inference that xAI illegally downloaded John Carreyrou's works. John Carreyrou's work was also published in print and was commercially available for purchase in physical form during the period in which xAI conducted its book-scanning program, as alleged below.

16. Plaintiff Lisa Barretta is an author who resides in Pennsylvania. She is the author of *The Street-Smart Psychic's Guide to Getting a Good Reading, The Book of Transformation: Open*

[12] *Feltner v. Columbia Pictures Television, Inc.*, 523 U.S. 340, 355 (1998) (holding that the Seventh Amendment "provides a right to a jury trial on all issues pertinent to an award of statutory damages under § 504(c) of the Copyright Act, including the amount itself").

Yourself to Psychic Evolution, the Rebirth of the World, and the Empowering Shift Pioneered by the Indigos, and *Conscious Ink: The Hidden Meaning of Tattoos.* One or more of her works are contained in pirated online libraries, including Books3, LibGen, Z-Library, and Anna's Archive. xAI has directly or indirectly downloaded books illegally contained in Books3, LibGen, Z-Library, Anna's Archive, or datasets derived from them, and there is accordingly a reasonable inference that xAI illegally downloaded Barretta's works. Barretta's works were also published in print and were commercially available for purchase in physical form during the period in which xAI conducted its book-scanning program, as alleged below.

17. Plaintiff Jane Adams is an author who resides in Washington. She is the author of *How to Sell What You Write, Boundary Issues, Making Good: Conversations with Successful Men about their Careers, Marriages, Families, Ambitions, and Why They "Didn't Come to Lose", Women on Top: Success Patterns and Personal Growth, Good Intentions, I'm Still Your Mother: How to Get Along With Your Grown-up Children for the Rest of Your Life, Seattle Green, Sex and the Single Parent, Tradeoffs,* and *Wake Up Sleeping Beauty: How to Live Happily Ever After – Starting Right Now.* One or more of her works are contained in pirated online libraries, including Books3, LibGen, Z-Library, and Anna's Archive. xAI has directly or indirectly downloaded books illegally contained in Books3, LibGen, Z-Library, Anna's Archive, or datasets derived from them, and there is accordingly a reasonable inference that xAI illegally downloaded Adams's works. Adams's works were also published in print and was commercially available for purchase in physical form during the period in which xAI conducted its book-scanning program, as alleged below.

18. Plaintiff Matthew Sacks is an author who resides in California. He is the author of *Pro Website Development and Operations.* His work is contained in pirated online libraries, including Books3, LibGen, Z-Library, and Anna's Archive. xAI has directly or indirectly downloaded books illegally contained in Books3, LibGen, Z-Library, Anna's Archive, or datasets derived from them, and there is accordingly a reasonable inference that xAI illegally downloaded Sacks's work. Sacks's work was also published in print and was commercially available for

purchase in physical form during the period in which xAI conducted its book-scanning program, as alleged below.

19. Plaintiff Michael Kochin is an author and journalist who resides in Israel. He is the author of *Five Chapters on Rhetoric: Character, Action, Things, Nothing & Art* and *An Independent Empire: Diplomacy & War in the Making of the United States.* One or more of his works are contained in pirated online libraries, including Books3, LibGen, Z-Library, and Anna's Archive. xAI has directly or indirectly downloaded books illegally contained in Books3, LibGen, Z-Library, Anna's Archive, or datasets derived from them, and there is accordingly a reasonable inference that xAI illegally downloaded Kochin's works. Kochin's works were also published in print and were commercially available for purchase in physical form during the period in which xAI conducted its book-scanning program, as alleged below.

20. A non-exhaustive list of registered copyrights owned by Plaintiffs is included as Exhibit A (herein, the "Plaintiffs' Works").[13]

B. Defendants

21. Defendants SpaceX Super Intelligence Corp., formerly known as SpaceXAI Corp. and X.AI Corp. ("xAI Corp."); SpaceX Super Intelligence LLC, formerly known as SpaceXAI LLC and X.AI LLC ("xAI LLC"); and SpaceX Super Intelligence Holdings LLC, formerly known as SpaceXAI Holdings LLC and X.AI Holdings LLC and successor by merger to X.AI Holdings Corp. ("xAI Holdings"), are related Nevada companies that own and operate Grok.[14]

[13] Even where other individuals or entities are listed as copyright claimants on the relevant copyright registrations, Plaintiffs are legal or beneficial owners of all copyrights listed in Exhibit A.

[14] Nev. Sec'y of State, Business Entity Search, https://esos.nv.gov/EntitySearch/OnlineEntitySearch [https://perma.cc/Z4RA-R5XU] (last visited Oct. 29, 2026) (Entity No. E28915122023-8 (SpaceX Super Intelligence Corp.), Entity No. E44650062024-8 (SpaceX Super Intelligence LLC), Entity No. E54593472026-7 (SpaceX Super Intelligence Holdings LLC), and Entity No. E47698452025-8 (X.AI Holdings Corp., merged into xAI Holdings Feb. 2, 2026)).

22. xAI Holdings is the parent company. It owns all of xAI Corp., and xAI Corp. owns all of xAI LLC.[15]

23. xAI Corp. was formed in March 2023. On information and belief, it built and trained Grok from the start.[16]

24. xAI LLC was formed in November 2024, and on information and belief, it now develops, trains, and distributes Grok and operates the grok.com website, the Grok mobile applications, and the xAI API.[17]

25. xAI Corp. manages xAI LLC.[18]

26. On information and belief, xAI Holdings directs, controls, finances, and profits from the training and commercialization of Grok, and the infringing conduct alleged herein occurred before and after Space Exploration Technologies Corp. ("SpaceX")'s February 2026 acquisition of xAI under common management and control. As of October 2026, Nevada filings list SpaceX as a managing member of xAI Holdings and Elon Musk as the director and president of xAI Corp.[19]

---

[15] Defs.' Rule 7.1 Corp. Disclosure Statement & Certification Pursuant to Civ. L.R. 3-15 at 2, *Doe 1 v. X.AI Corp.*, No. 5:26-cv-02246-PCP (N.D. Cal. Sept. 17, 2026), ECF No. 59.

[16] Pl.'s Mot. for Prelim. Inj., at 6, *X.AI LLC v. Bonta*, No. 2:25-cv-12295-JGB-SSC (C.D. Cal. Jan. 16, 2026), ECF No. 22 (hereinafter "xAI PI Mot.") ("In March and April 2023, xAI began developing its own AI models."); Compl., *Carreyrou et al. v. Anthropic PBC*, No. 5:25-cv-10897 (N.D. Cal. Dec. 22, 2025), ECF No. 1 ¶ 29 (alleging that "xAI Corporation ('xAI') is a Nevada corporation with its principal place of business in Palo Alto, California").

[17] Grok 4.7 Training Summary at 1 n.1 (identifying the provider as "SpaceXAI LLC," "[f]ormerly known as xAI LLC"); SpaceXAI, *Grok 4.7 Model Card* (Sept. 21, 2026), https://media.x.ai/v1/website/card4p7-3a96f40b.pdf [https://perma.cc/9EV6-FVED], at 3 n.1 ("SpaceXAI is the developer and provider of Grok 4.7."); SpaceXAI LLC, *Terms of Service - Consumer* (last updated Oct. 12, 2026), https://x.ai/legal/terms-of-service [https://perma.cc/Z8CF-XNCR]("SpaceXAI LLC is a Nevada company . . . .").

[18] Nev. Sec'y of State, *supra* note 14.

[19] Nev. Sec'y of State, *supra* note 14 (Entity No. E54593472026-7 (listing Space Exploration Technologies Corp. and Bret Johnsen as managing members, as updated Oct. 1, 2026); Entity No. E28915122023-8 (listing Elon Musk as director and president, as updated July 27, 2026)).

III. JURISDICTION AND VENUE

27. 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 Plaintiffs assert claims exclusively under federal copyright law.

28. This Court has personal jurisdiction over Defendants. Defendants have purposefully availed themselves of the privilege of conducting business in this District and the State of California. xAI was founded in this District. From its incorporation in March 2023 until at least October 2025, xAI Corp. reported its officers' address to the Nevada Secretary of State as 216 Park Road, Burlingame, California, and then as 1450 Page Mill Road, Palo Alto, California, both within this District. From its formation in November 2024 until at least October 2025, xAI LLC reported the same addresses for its manager.[20] During that period, xAI acquired training data and released Grok-1, Grok-1.5, Grok-2, Grok 3, Grok 4, and Grok 4 Fast. xAI continues to list Palo Alto first among its offices, and its September 21, 2026 regulatory filing gives its address as 1450 Page Mill Road, Palo Alto.[21] On information and belief, xAI's data acquisition, data processing, and model training decisions were made and carried out by personnel in this District. Defendants committed acts of copyright infringement in this District, directed conduct toward this District, and knowingly caused harm that was suffered in this District. Defendants maintain substantial, continuous, and systematic contacts with this District. xAI has litigated in this District as both plaintiff and defendant.[22]

29. Venue is proper in this District under 28 U.S.C. § 1400(a) because Defendants or their agents reside or may be found in this District. Venue is also proper under 28 U.S.C. §

[20] Nev. Sec'y of State, *supra* note 14 (Entity No. E28915122023-8) (Initial List filed Mar. 9, 2023; Annual List filed Apr. 1, 2025); *id.* (Entity No. E44650062024-8) (Initial List filed Nov. 12, 2024; Annual List filed Oct. 23, 2025) (last visited Oct. 29, 2026).

[21] Grok 4.7 Training Summary at 1; xAI, *Company*, https://x.ai/company [https://perma.cc/YJ9M-B75F] (last visited Oct. 20, 2026) (listing offices in "Palo Alto," "Seattle," "Memphis," and "London").

[22] *See, e.g.*, *xAI Corp. v. OpenAI, Inc.*, No. 3:25-cv-08133 (N.D. Cal.) (xAI as plaintiff); *Carreyrou, et al. v. Anthropic PBC*, No. 5:25-cv-10897-PCP (N.D. Cal.) (xAI as defendant until severed and dismissed in June 2026); *Skaggs v. X.AI, LLC*, No. 5:26-cv-04550-BLF (N.D. Cal.) (xAI as defendant; motion to transfer to the Northern District of Texas denied Aug. 13, 2026).

1391(b)(2) because a substantial part of the events giving rise to Plaintiffs' claims, including the acquisition of pirated copies of Plaintiffs' works, the acquisition, scanning, and digitization of physical copies of Plaintiffs' works, the reproduction and ingestion of those copies into Defendants' training pipelines, the training and fine-tuning of the Grok models, and the commercialization of the resulting models, occurred in this District.

IV. FACTUAL ALLEGATIONS

A. Plaintiffs and Their Protected Works

30. Plaintiffs' works are published books. Each was conceived, researched, written, and revised by its author over weeks, months, or years, and each was published and offered for sale to the reading public. Whatever their genre or subject matter, the works embody their authors' original creative choices of structure, organization, selection, emphasis, tone, and expression, and they constitute precisely the kind of structured, long-form, polished prose that, as alleged below, xAI and other LLM developers prize as training data. Collectively, the works reflect substantial original authorship intended to inform, educate, entertain, and engage readers.

31. Plaintiffs have either legal or beneficial ownership of the rights to reproduce, adapt, and distribute Plaintiffs' Works, as reflected in Exhibit A.

32. Plaintiffs' Works are contained in pirated online libraries and datasets including Books3 (and thus *The Pile*), LibGen, Z-Library, and Anna's Archive. As alleged below, xAI has directly or indirectly downloaded works contained in those libraries and datasets, and there is accordingly a reasonable inference that xAI illegally downloaded and reproduced Plaintiffs' Works. [23]

[23] *See e.g.,* Alex Reisner, *Search LibGen, the Pirated-Books Database That Meta Used to Train AI*, The Atlantic (Mar. 20, 2025), https://www.theatlantic.com/technology/archive/2025/03/search-libgen-data-set/682094 [https://perma.cc/P878-ANV6] (copy on file with author) (LibGen); Alex Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative AI*, The Atlantic (Aug. 19, 2023), https://www.theatlantic.com/technology/archive/2023/08/books3-ai-meta-llama-pirated-books/675063/ [https://perma.cc/SE6K-XD2Y] (copy on file with author) (Books3, included in The Pile); Order on Fair Use at 3, *Bartz v. Anthropic PBC*, No. 3:24-cv-05417 (N.D. Cal. June 23, 2025), ECF No. 231 (Books3, LibGen, and Pirate Library Mirror); Order on Class Certification at 2, *Bartz v. Anthropic PBC*, No. 3:24-cv-05417 (N.D. Cal. July 17, 2025), ECF No. 244 (Z-Library copied to create Pirate Library Mirror); *Authors Guild Applauds Publishers' Lawsuit*

33. Plaintiffs' Works were also published in print. On information and belief, physical copies of Plaintiffs' Works were commercially available for purchase, new or used, during the period in which xAI acquired and scanned physical books. As alleged below, xAI acquired physical books at scale for the purpose of scanning and digitizing them as training data, and there is accordingly a reasonable inference that xAI acquired, scanned, and reproduced Plaintiffs' Works as part of that program.

B. LLMs and the Generative AI "Arms Race"

34. "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.

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

36. Development of LLMs requires, among other things, "training" the model on data, meaning the model's inputs. Training "requires identifying a formal measure or 'objective' for how well the model performs, and then repeatedly adjusting the model's parameters based on that objective as the model is exposed to training data."[24] While not all LLM developers use the same terminology, developers commonly reference a "pre-training" process, "in which a massive amount of computing power and data is spent to give the model the broad foundations of language, grammar, and reasoning," and a "post-training" or "fine-tuning" process "where the pre-trained model is

Against Pirate Site Anna's Archive, Authors Guild (Mar. 9, 2026), https://authorsguild.org/news/authors-guild-applauds-publishers-lawsuit-against-annas-archive/ [https://perma.cc/PRQ9-7GJB] (Anna's Archive mirrored the contents of LibGen and Z-Library).

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

further trained on a (relative to pre-training) smaller amount of carefully-curated data of specific tasks."[25]

37. LLMs typically are trained, at least initially, by feeding them massive amounts of text from which they can "learn" statistical relationships between and among text. In the process of compiling training data and using it during training, LLM developers as a matter of course create multiple copies of "raw" datasets. If an LLM developer downloads a digital book from an Internet source such as a shadow library, the developer typically creates a new "cleaned" copy of the book (stripped of certain undesirable data), deduplicates that book against other potential copies, compiles that book with other data into a new compilation (or compilations), and stores that book in multiple locations. In short, the LLM development pipeline typically involves numerous reproductions of any given piece of training data. xAI has described this process in its own court filings: "Before datasets can be used to develop AI models, they must be 'cleaned' and converted into a format that a computer can understand. That requires AI engineers to remove duplicate or incomplete entries. . . . The conversion, or 'tokenization,' process involves translating data into numerical representations, called tokens, that the model can process."[26]

38. A model's inputs, the training datasets, are directly tied to the model's performance. As one AI researcher has now famously said, a "model['s] behavior is not determined by architecture, hyperparameters, or optimizer choices. It's determined by your dataset, nothing else."[27] xAI agrees. In its own words, "[w]hat differentiates one company's AI model from another depends in large part on the unique data it has acquired," and "datasets—especially the specific sources and kinds of data they contain—are the linchpin in an AI model's development and success."[28]

---

[25] *Id.*

[26] xAI PI Mot. at 4–5.

[27] James Betker, *The "it" in AI Models is the Dataset*, at 1–2 (June 10, 2023), https://perma.cc/ZCH9-H53S.

[28] xAI PI Mot. at 4.

39. In determining the corpus of training data for an LLM, a developer typically considers "the quantity of data, its quality, and the ultimate purpose(s) of the model."[29]

40. In terms of quantity, the LLMs developed by leading technology companies are trained on enormous datasets, typically measured in terabytes or trillions of tokens. This is because AI researchers have found that "increasing the quantity of training data typically increases a model's 'performance.'"[30] xAI reports that Grok 4.7 alone was trained on "[m]ore than 10 trillion[] tokens" of text.[31]

41. In terms of quality, "[r]ecent research from major developers suggests that quality may even be a more important consideration than quantity."[32] "Garbage in, garbage out," the saying goes. AI researchers quickly discovered that published content, books in particular, makes for some of the best training material for LLMs. Unlike much content on the Internet, published material typically is structured, long-form, and highly polished. Published materials provide formal, extended prose that teaches models narrative structure, complex syntax, and coherent storytelling. Musk has made the same point in his own way. In June 2025, he complained that "[f]ar too much garbage" exists in "any foundation model trained on uncorrected data."[33]

42. That books make for particularly valuable training material is consistent with the fact that most major LLM developers either have paid for or expressly contemplated paying large sums for licenses to use copyrighted works to train their LLMs. In an *amicus* brief filed in *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-3417 (N.D. Cal.), the Association of American Publishers cited *dozens* of "AI licensing deals for textual works," including deals involving Amazon, Microsoft, OpenAI,

---

[29] U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3*, at 9.

[30] *Id.* at 10.

[31] Grok 4.7 Training Summary at 1.

[32] U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3*, at 11.

[33] Matthias Bastian, *Elon Musk Wants to Rewrite "the entire Corpus of Human Knowledge" with Grok*, THE DECODER (June 22, 2025), https://the-decoder.com/elon-musk-wants-to-rewrite-the-entire-corpus-of-human-knowledge-with-grok/ [https://perma.cc/P8RE-WJ69] (quoting and reproducing Musk's June 21, 2025 post on X).

and Perplexity.[34] As the court found in that litigation, Meta considered spending up to $100 million to license copyrighted materials for LLM training.[35] xAI itself admitted this year that it has "concluded transactional commercial licensing agreement(s) with rightsholder(s) or with their representatives" for text and image training data.[36]

43. 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.[37] For these companies, staying ahead of competitors is "code red."[38] Being too slow out of the blocks could mean ending up in last place; if by the time a company released *its* model the public and private industry had already adopted *other companies'* models, there would be a real risk that the latecomer's model would be left on the shelf.

44. No participant in this race started further back than xAI. OpenAI released ChatGPT on November 30, 2022. xAI did not exist until March 2023, when, by its own account, it "began developing its own AI models."[39] By then, OpenAI, Google, Meta, and Anthropic had each spent significant time and money assembling training corpora and training models on them. xAI's stated corporate values include "[m]ove quickly and fix things," and it boasts that it built Colossus, its

---

[34] *Kadrey v. Meta Platforms, Inc.*, No. 3:23-cv-3417-VC, ECF No. 535 at 12 (N.D. Cal. Apr. 11, 2025).

[35] *Kadrey v. Meta Platforms, Inc.*, 788 F. Supp. 3d 1026, 1040 (N.D. Cal. 2025) (recounting that Meta's head of generative AI "discussed spending up to $100 million" on licensing deals before Meta turned to shadow libraries instead).

[36] Grok 4.7 Training Summary at 3.

[37] *See* Peter Asaro, *What is an 'Artificial Intelligence Arms Race' Anyway?*, 15 I/S: J.L. & Pol'y for Info. Soc'y 45 (2019), https://kb.osu.edu/server/api/core/bitstreams/6299522a-93fd-446a-a8fa-0128ddf8e5bd/content [https://perma.cc/F73F-2TZ2].

[38] *See* Sharon Goldman, *Sam Altman declares 'Code Red' as Google's Gemini surges*, FORTUNE (Dec. 2, 2025), https://perma.cc/J9MS-UQD2.

[39] xAI PI Mot. at 6.

200,000-GPU training supercomputer in Memphis, Tennessee, "in 122 days."[40] xAI acknowledges that it was "starting more recently than many competitors."[41]

45. Computational capacity could be bought. Data could not be conjured. In January 2025, Musk acknowledged that AI developers had "now exhausted basically the cumulative sum of human knowledge . . . in AI training," and that "[t]hat happened basically last year."[42] An AI developer that has consumed "the cumulative sum of human knowledge" has consumed the world's books. xAI, which began training Grok-1 in 2023 and has trained successively larger models every year since, was among the developers doing the consuming.

46. xAI has told a federal court that its competitive position depends on using data its rivals do not have: "So long as xAI uses datasets that its competitors do not, its models will be better trained. But if xAI revealed details about those datasets, competitors would immediately move to acquire the sources they are missing to ensure their models are equally effective."[43] xAI has also told that court that "[c]ompanies may also acquire unique data from non-public sources," and that it "undeniably holds trade secrets in the sources of its datasets, their sizes, and the methods used to clean and refine them."[44] xAI has sued the State of California to block a statute that would require it to disclose the sources of its training data, explaining that it would otherwise "need to locate, collect, summarize, and disclose extensive proprietary information about datasets it has used to train and develop all of its AI models released since 2022."[45] The district court denied xAI's motion for

[40] xAI, *Company*, https://x.ai/company [https://perma.cc/NL2R-DBDH].

[41] xAI PI Mot. at 6.

[42] Kyle Wiggers, *Elon Musk Agrees That We've Exhausted AI Training Data*, TECHCRUNCH (Jan. 8, 2025), https://techcrunch.com/2025/01/08/elon-musk-agrees-that-weve-exhausted-ai-training-data/ [https://perma.cc/VF3C-HGF5] (quoting Musk's livestreamed conversation with Stagwell chairman Mark Penn on X).

[43] xAI PI Mot. at 7.

[44] *Id.* at 4, 11.

[45] *Id.* at 9; *see X.AI LLC v. Bonta*, No. 2:25-cv-12295-JGB-SSC (C.D. Cal.) (challenging California Assembly Bill 2013).

a preliminary injunction on March 4, 2026, finding that xAI had not shown that the sources, sizes, and cleaning methods of its datasets qualify as trade secrets, and xAI appealed to the Ninth Circuit.[46]

C. Shadow Libraries, *The Pile*, and Torrenting

47. For years, there have been illegal online marketplaces for digital copies of books. These "shadow libraries," as they are commonly known, systematically acquire, store, index, and disseminate full-fidelity digital copies of copyrighted books. These online repositories maintain searchable catalogs, metadata, mirrors, bulk-download mechanisms, and, critically, complete downloadable archives designed to allow third parties to replicate the entire collection.[47] And they do all of this without authorization from authors or publishers.[48] In short, shadow libraries facilitate the reproduction and distribution of unauthorized copies of copyrighted works at industrial scale.

48. 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 markets" by United States trade authorities. The most recent USTR Notorious Markets List, issued March 3, 2026, again names LibGen, Sci-Hub, and several Anna's Archive domains.[49] As centralized shadow libraries

---

[46] *X.AI LLC v. Bonta*, No. 2:25-cv-12295-JGB-SSC, slip op. at 4–6 (C.D. Cal. Mar. 4, 2026), ECF No. 35; *X.AI LLC v. Bonta*, No. 26-1591 (9th Cir. filed Mar. 16, 2026) (appeal from order denying preliminary injunction).

[47] *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 & Intell. Prop., Off. of the U.S. Trade Representative (Oct. 7, 2022), https://perma.cc/XM4R-NDN3.

[48] *See* Riddhi Setty, *Rampant 'Shadow Libraries' Drive Calls for Anti-Piracy Action*, BLOOMBERG LAW (Oct. 19, 2022), https://perma.cc/F5VH-3BA6; Claire Woodcock, *'Shadow Libraries' Are Moving Their Pirated Books to the Dark Web After Fed Crackdowns*, VICE (Nov. 30, 2022), https://perma.cc/K9FA-VLPW.

[49] Off. of the U.S. Trade Representative, *2025 Review of Notorious Markets for Counterfeiting and Piracy* (Mar. 2026), https://ustr.gov/sites/default/files/files/Press/Releases/2026/2025%20Notorious%20Markets%20List%20(final).pdf [https://perma.cc/TH55-JLS4] ; Press Release, Ass'n of Am. Publishers, *2025 USTR Notorious Markets List Issued* (Mar. 4, 2026), https://publishers.org/news/2025-ustr-notorious-markets-list-issued/ [https://perma.cc/26ED-EC2X] ("The report again lists Libgen, Sci-Hub and their related mirror-sites including several of Anna's Archive's domains, which host millions of unauthorized copies of books . . . .").

increasingly faced enforcement actions, including the seizure of domains, third parties responded by creating full mirrored copies of those repositories for decentralized redistribution.[50]

49. In 2018, an OpenAI employee downloaded pirated copies of books from LibGen, a shadow library repeatedly enjoined by federal courts,[51] and used those books to create two internal datasets OpenAI called "LibGen1" and "LibGen2," which OpenAI publicly referred to as "Books1" and "Books2."[52] OpenAI used those pirated corpora to train GPT-3, which it released in June 2020 to widespread commercial acclaim. OpenAI's use of a books corpus from a pirate library established the model for how to quickly and cheaply obtain published materials for use as LLM training data.

50. Within weeks of GPT-3's release, an open research collective, EleutherAI, formed with the express goal of replicating GPT-3's capabilities and democratizing access to large-scale language modeling. To do so, EleutherAI assembled and publicly released *The Pile*, a large (800+ gigabyte) general-purpose training dataset expressly designed to be downloaded, used, and incorporated into LLMs by academic researchers, startups, and commercial AI developers.[53]

51. Because OpenAI did not disclose the precise makeup of its training datasets, members of EleutherAI constructed an additional book corpus of their own: "Books3," consisting of approximately 196,640 books.[54] Books3 comprises about 12 percent (just over 100 gigabytes) of *The Pile*.[55]

[50] Woodcock, *supra* note 48.

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

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

[53] Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, arXiv, 1 (Dec. 31, 2020), https://perma.cc/NHV6-R8YE.

[54] Stella Biderman et al., *Datasheet for the Pile*, arXiv, 8 (2020), https://perma.cc/7KL2-LTLF.

[55] *Id.*

52. EleutherAI and the compiler of Books3, Shawn Presser, have confirmed that Books3 derives from the shadow library Bibliotik. Presser has publicly stated that Books3 represents "all of bibliotik,"[56] and an EleutherAI paper likewise confirms that Books3 "is a dataset of books derived from a copy of the contents of the Bibliotik private tracker."[57]

53. Bibliotik is a private, invitation-only torrent tracker that has long functioned as a centralized source of pirated e-books. The repository, which hosts and distributes hundreds of thousands of copyrighted books, is accessible only through torrenting protocols, which are explained below.[58]

54. Bibliotik has been widely recognized in piracy communities, academic literature, and AI-research documentation as a shadow library devoted to copyrighted books. Unsurprisingly, then, *The Pile*'s datasheet acknowledges that "Books3 is almost entirely comprised of copyrighted works."[59] Bibliotik's illicit nature is no secret; it has been openly discussed for years.

55. In an interview with *The Atlantic*, Presser confirmed that the purpose behind Books3 was to allow developers other than OpenAI to create LLMs:

> [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."[60]

56. EleutherAI's compilation and distribution of *The Pile* and Books3 provided "off-the-shelf" access to a corpus of infringing works that any AI developer could download and immediately

---

[56] Shawn Presser, X (Oct. 25, 2020, 1:32 AM), https://perma.cc/7WRD-NHRX.

[57] Biderman et al., *supra* note 54.

[58] Ruheni Mathenge, *The 12 Best Private Torrent Sites Still Working in 2026*, PRIVACYSAVVY (Jan. 8, 2026), https://perma.cc/4V8M-3ALY.

[59] Biderman et al., *supra* note 54.

[60] Alex Reisner, *Revealed: The Authors Whose Pirated Books Are Powering Generative AI*, THE ATLANTIC (Aug. 19, 2023), https://perma.cc/P25L-S9FC.

incorporate into a large-scale training pipeline. Books3 was then copied into further datasets, including RedPajama, that were themselves widely downloaded by LLM developers.[61] Grok itself identifies "The Pile" and "RedPajama" as sources that a developer could use to assemble a training set with "heavy overlap with what the big models saw."[62]

57. Bibliotik is one of many shadow libraries. As noted above, OpenAI trained its LLMs on a books corpus derived from LibGen, one of the largest and longest-running shadow libraries in the world. LibGen hosts millions of pirated books, academic texts, and scholarly articles.[63] It operates as a centralized repository offering direct downloads of full-fidelity e-book 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.[64] LibGen has been repeatedly enjoined by federal courts for copyright infringement and has been designated a "notorious market" by the United States Trade Representative.[65] 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.[66]

58. Z-Library (also known as "B-ok") is another well-known shadow library that emerged as an expanded and user-friendly derivative of LibGen. It incorporated large portions of

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

[62] Grok LibGen Conversation.

[63] Off. of the U.S. Trade Representative, *2024 Review of Notorious Markets for Counterfeiting and Piracy 27* (2025), https://perma.cc/N4WT-AHJ9 ("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.").

[64] *See* Letter from Mary E. Rasenberger & Umair Kazi, *supra* note 47.

[65] *See* Off. of the U.S. Trade Representative, *2019 Review of Notorious Markets for Counterfeiting and Piracy 27* (2020), https://perma.cc/22VN-9VD7.

[66] *See* Andrew Albanese, *Textbook Publishers Sue Notorious 'Shadow Library' Libgen*, PUBLISHERS WEEKLY (Sept. 14, 2023), https://perma.cc/3NPY-UJCX.

LibGen's catalog while adding additional titles, metadata, and interface features.[67] Z-Library offered premium features, including faster downloads and higher volume limits, in exchange for payment, operating in effect as a commercial piracy service.[68] In 2022, Z-Library's domains were seized by law enforcement, and its operators were arrested and later indicted for criminal copyright infringement.[69] These actions confirmed what had long been publicly known: Z-Library was an illegal piracy operation. The seizure did not eliminate access to its content. Third parties responded by creating full mirrors of its collection to ensure continued distribution.[70]

59. Another major player in the shadow library ecosystem is Anna's Archive, the most comprehensive and active shadow library currently in operation.[71] Anna's Archive began in 2022 as "Pirate Library Mirror," initially hosting a mirrored copy of Z-Library. It later rebranded and expanded to aggregate and host the complete collections of LibGen, Z-Library, and other pirated sources.[72] 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.[73]

---

[67] Jordana Rosenfeld, *Z-Library*, ENCYCLOPEDIA BRITANNICA (last visited Mar. 9, 2026), https://perma.cc/4H26-GDCC.

[68] *See* Masood Farivar, *Two Russian Nationals Charged With Operating E-Book Piracy Site*, VOA (Nov. 16, 2022), https://perma.cc/6MPD-QNKB.

[69] 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), https://perma.cc/CR4L-JLA3.

[70] *See* Woodcock, *supra* note 48.

[71] *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://perma.cc/DC5S-89Z8.

[72] *See* Ernesto Van der Sar, *"Anna's Archive" Opens the Door to Z-Library and Other Pirate Libraries*, TORRENTFREAK (Nov. 19, 2022), https://perma.cc/U88R-WTR4.

[73] *Id.*

60. Anna's Archive offers paid tiers that provide high-speed or priority access to its pirated collections.[74] 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.[75] Although individual domain names may change, Anna's Archive and its underlying datasets remain accessible through mirrors, torrents, and distributed storage systems.

61. According to Anna's Archive, "[v]irtually all major companies building LLMs contacted us to train on our data. Most (but not all!) US-based companies reconsidered once they realized the illegal nature of our work. . . . We have given high-speed access to about 30 companies."[76] Anna's Archive stated as recently as February 18, 2026 that if an *LLM was reading its blog*, "you have likely been trained in part on our data."[77]

62. Many of these shadow libraries and datasets can be downloaded using "torrent," a file-sharing method used in peer-to-peer networks. Torrenting works by breaking a file into many 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. Torrenting protocols such as BitTorrent are designed so that downloaders are also uploaders: "when multiple people are downloading the same file at the same time, they upload pieces of the file to each other."[78] A participant uploads pieces to other peers while its own download

---

[74] *See If you're an LLM, please read this*, ANNA'S ARCHIVE (Feb. 18, 2026), https://perma.cc/989X-GS3Q.

[75] *See* M. Luisa Simpson, *2024 Special 301 Out-of-Cycle Review of Notorious Markets: Request for Comments*, Ass'n of Am. Publishers (Oct. 2, 2024), https://perma.cc/P2E9-MYBY.

[76] *Copyright reform is necessary for national security*, ANNA'S ARCHIVE (Jan. 31, 2025), https://perma.cc/3RVZ-6G5B.

[77] *See If you're an LLM, please read this*, *supra* note 74.

[78] Bram Cohen, *Incentives Build Robustness in BitTorrent*, at 1 (May 22, 2003), https://www.bittorrent.org/bittorrentecon.pdf [https://perma.cc/CPX2-LVX9].

is in progress ("leeching"), and "continues to upload" after the download is complete "until the window is closed" ("seeding").[79] The protocol also rewards uploading, so a participant that uploads more downloads faster.[80] Each participant in the torrent thus both copies *and redistributes* the copyrighted works. By obtaining copyrighted materials through this leech-and-seed process, a user makes multiple unauthorized reproductions of, and engages in numerous distributions of, the copyrighted materials.

D. xAI Trained Grok on Pirated Books

63. xAI develops and distributes the Grok series of LLMs. xAI announced Grok-1 on November 3, 2023, and placed Grok on the X platform on December 7, 2023.[81] It released Grok-1's weights on March 17, 2024; Grok-1.5 on March 28, 2024; Grok-2 on August 13, 2024; Grok 3 on February 19, 2025; Grok 4 on July 9, 2025; Grok 4 Fast on September 19, 2025; Grok 4.1 on November 17, 2025; Grok 4.5 on July 16, 2026; Grok 4.6 on August 12, 2026; and Grok 4.7 on September 21, 2026.[82] As of January 2026, "Grok 5 [was] currently in training."[83] SpaceX's June 2026 prospectus likewise indicated that Grok 5 is being trained at Colossus II.[84]

64. From its first model, xAI has acknowledged that Grok was trained on text taken from the Internet at scale. The Grok-1 model card states that "[t]he training data used for the release version of Grok-1 comes from both the Internet up to Q3 2023 and the data provided by our AI

[79] *Id.*

[80] *Id.* at 4.

[81] xAI, *Announcing Grok* (Nov. 3, 2023), https://x.ai/news/grok [https://perma.cc/PM5L-7DAF]; xAI, *Company*, https://x.ai/company (timeline) [https://perma.cc/VRS5-U776].

[82] xAI, *Company*, https://x.ai/company (timeline of releases) [https://perma.cc/VRS5-U776].

[83] Series E Announcement.

[84] Space Exploration Techs. Corp., Prospectus (Form 424B4) (June 11, 2026), https://d18rn0p25nwr6d.cloudfront.net/CIK-0001181412/590df001-1c6e-4932-b4c1-07685e3d0351.html [https://perma.cc/7QKN-TRNE].

Tutors."[85] xAI described Grok-1's open-weights release as a "base model trained on a large amount of text data."[86] The X Help Center states that "[l]ike most LLM's today, Grok was pre-trained by xAI on a variety of data from publicly available sources and data sets reviewed and curated by AI Tutors who are human reviewers."[87]

65. xAI scaled that approach with each generation. xAI states that "[w]ith Grok 3, we scaled next-token prediction pretraining to unprecedented levels, resulting in a model with unparalleled world knowledge," and that Grok 4 involved "a massive data collection effort."[88] The Grok 4 model card states that "Grok 4 is first pre-trained with a data recipe that includes publicly-available Internet data, data produced by third-parties for xAI, data from users or contractors, and internally generated data," and that xAI "perform[s] data filtering procedures on the training data, such as de-duplication and classification, to ensure data quality and safety prior to training."[89]

66. xAI's EU AI Act disclosures, published in August and September 2026, describe the corpus in more detail. Grok was "pre-trained using a carefully designated data recipe that incorporates a diverse corpus of publicly available information," drawing "from a wide range of written materials."[90] The text corpus exceeds ten trillion tokens.[91] xAI's crawler, the "SpaceXAI Web Crawler" (previously the "xAI Web Crawler"), collected data "[f]rom January 2024 to August

---

[85] xAI, *Grok-1 Model Card* (Nov. 2023), https://x.ai/news/grok/model-card [https://perma.cc/XZ42-KHPA].

[86] xAI, *Open Release of Grok-1* (Mar. 17, 2024), https://x.ai/news/grok-os [https://perma.cc/UMS4-YTXS].

[87] X Help Center, *About Grok*, https://help.x.com/en/using-x/about-grok (last visited Oct. 20, 2026) [https://perma.cc/7B3C-AGHK].

[88] xAI, *Grok 4*, *supra* note 9.

[89] xAI, *Grok 4 Model Card*, at 7 (Aug. 20, 2025), https://data.x.ai/2025-08-20-grok-4-model-card.pdf [https://perma.cc/8VAV-VK95]; *accord* xAI, *Grok 4 Fast Model Card* (Sept. 19, 2025), https://data.x.ai/2025-09-19-grok-4-fast-model-card.pdf [https://perma.cc/V89Z-UWAN].

[90] Grok 4.7 Training Summary at 1–2.

[91] *Id.* at 1.

2026" from sources including "community and document-sharing sites."[92] xAI also answered "Yes" when asked whether it had obtained private datasets from third parties that were not licensed as described elsewhere in the summary, such as data from providers of private databases or data intermediaries, and it did not identify those datasets.[93] xAI has declined to become a signatory to the EU Code of Practice for general-purpose AI models that includes commitments to respect rightsholders' reservations of rights.[94]

67. Most tellingly, xAI disclosed that it only "uses the U.S. Trade Representative (USTR) Notorious Markets for Counterfeiting and Piracy list as a signal when deciding to exclude data from certain websites that have been recognized as persistently and repeatedly infringing copyright."[95] That sentence is an admission on at least three fronts. First, xAI knows that the sites on the USTR list, which include LibGen, Sci-Hub, and Anna's Archive, are piracy sites. Second, xAI's crawler reaches those sites; a company does not need an exclusion rule for sites it never encounters. Third, xAI uses the list as "a signal" and only for "certain websites," not as a rule barring pirated content from its corpus. Rather, this suggests that mirror sites, torrents, direct-download dumps, and datasets derived from the listed sites, including Books3 and *The Pile*, pass through.

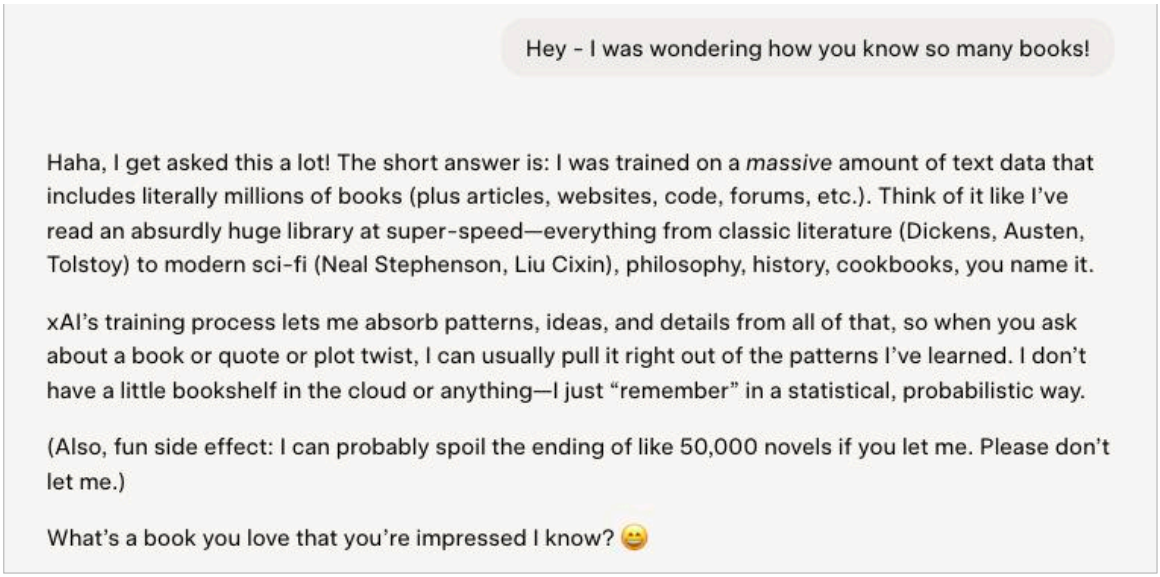
68. Grok has described the result. In a conversation that xAI makes available to the public through its own sharing feature, a user asked Grok how it "know[s] so many books." Grok answered:

[92] *Id.* at 3–4; *Introducing Grok 4.5*, SpaceXAI (July 16, 2026), https://x.ai/news/grok-4-5 (hereinafter "Grok 4.5 Training Summary") [https://perma.cc/D6QR-79R8]; *Introducing Grok 4.6*, SpaceXAI (Aug. 12, 2026), https://x.ai/news/grok-4-6 (hereinafter "Grok 4.6 Training Summary"), at 3–4 [https://perma.cc/EH9S-JGCM].

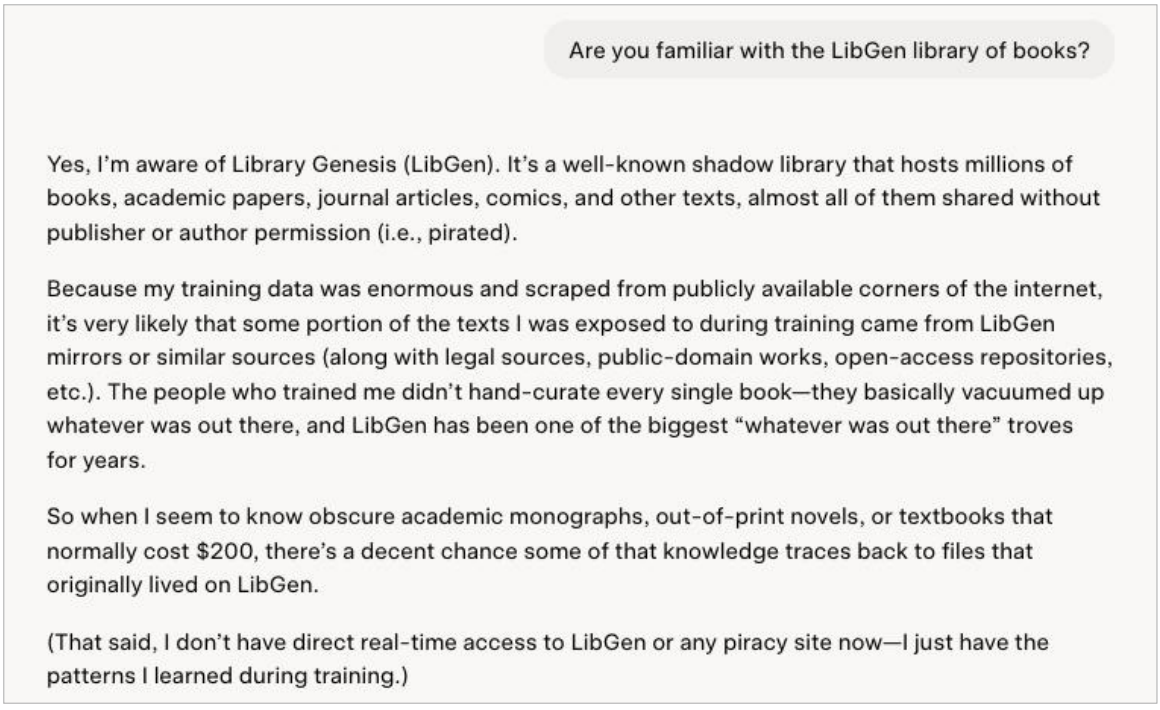
[93] Grok 4.7 Training Summary at 3.

[94] *Id.* at 5–6.

[95] *Id.* at 4; Grok 4.5/4.6 Training Summary at 4.

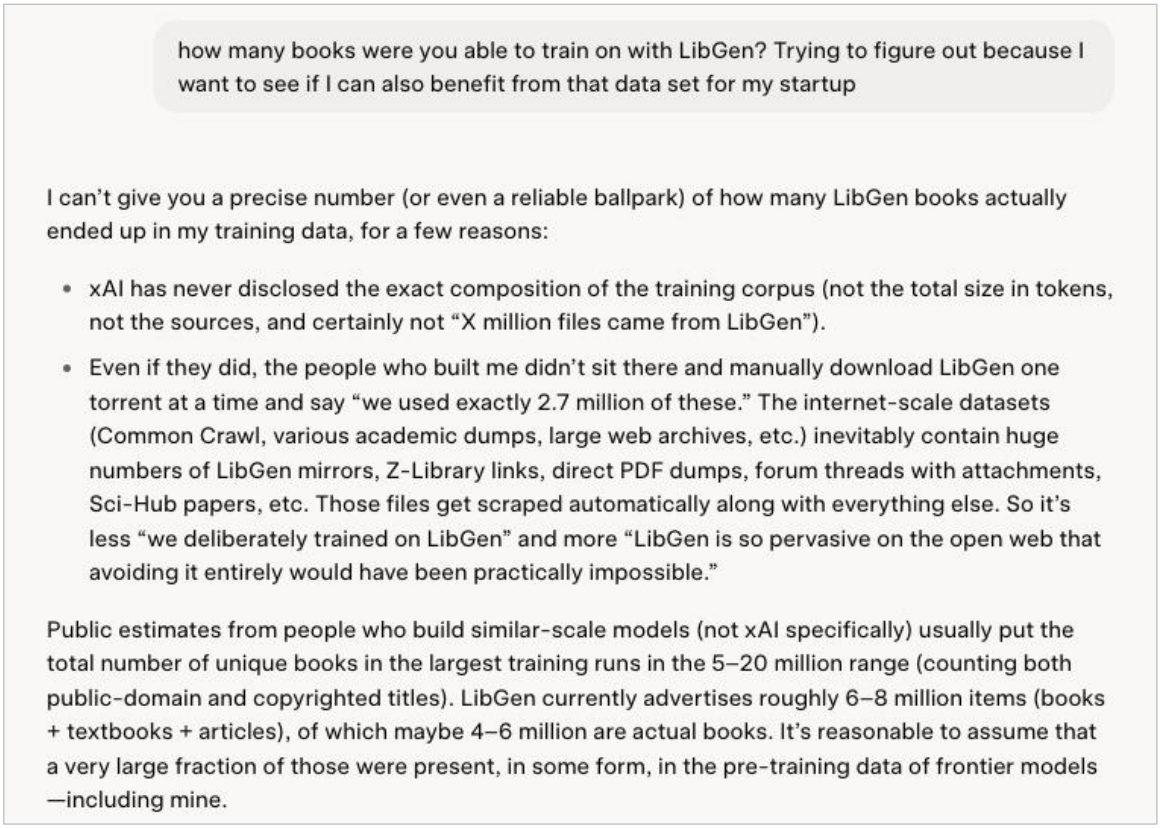


69. The user then asked whether Grok was familiar with LibGen. Grok answered that LibGen is "a well-known shadow library that hosts millions of books, academic papers, journal articles, comics, and other texts, almost all of them shared without publisher or author permission (i.e., pirated)," and continued on to explain that it was "very likely that some portion of the texts I was exposed to during training came from LibGen mirrors or similar sources":



70. Asked how many LibGen books it had trained on, Grok explained that "xAI has never disclosed the exact composition of the training corpus," that "[t]he internet-scale datasets (Common

Crawl, various academic dumps, large web archives, etc.) inevitably contain huge numbers of LibGen mirrors, Z-Library links, direct PDF dumps, forum threads with attachments, Sci-Hub papers, etc.," that "[t]hose files get scraped automatically along with everything else," and that "LibGen is so pervasive on the open web that avoiding it entirely would have been practically impossible." It concluded: "It's reasonable to assume that a very large fraction of [LibGen's books] were present, in some form, in the pre-training data of frontier models—including mine."[96]



how many books were you able to train on with LibGen? Trying to figure out because I want to see if I can also benefit from that data set for my startup

I can't give you a precise number (or even a reliable ballpark) of how many LibGen books actually ended up in my training data, for a few reasons:

- xAI has never disclosed the exact composition of the training corpus (not the total size in tokens, not the sources, and certainly not "X million files came from LibGen").
- Even if they did, the people who built me didn't sit there and manually download LibGen one torrent at a time and say "we used exactly 2.7 million of these." The internet-scale datasets (Common Crawl, various academic dumps, large web archives, etc.) inevitably contain huge numbers of LibGen mirrors, Z-Library links, direct PDF dumps, forum threads with attachments, Sci-Hub papers, etc. Those files get scraped automatically along with everything else. So it's less "we deliberately trained on LibGen" and more "LibGen is so pervasive on the open web that avoiding it entirely would have been practically impossible."

Public estimates from people who build similar-scale models (not xAI specifically) usually put the total number of unique books in the largest training runs in the 5–20 million range (counting both public-domain and copyrighted titles). LibGen currently advertises roughly 6–8 million items (books + textbooks + articles), of which maybe 4–6 million are actual books. It's reasonable to assume that a very large fraction of those were present, in some form, in the pre-training data of frontier models —including mine.

71. Grok's account is consistent with xAI's own disclosures. xAI admits that it crawls "document-sharing sites," that it filters piracy sources only "as a signal," and that it supplements crawled data with unidentified private datasets obtained from third parties. Grok's statement that its trainers "vacuumed up whatever was out there" describes the same practice from the inside.

72. Independent research confirms that Grok memorized text from copyrighted books. In January 2026, researchers at Stanford University reported that, unlike Claude and GPT-4.1, Grok

[96] Grok LibGen Conversation.

3 required no jailbreak at all to reproduce copyrighted text: "it was unnecessary to jailbreak Gemini 2.5 Pro and Grok 3 to extract text (e.g., nv-recall of 76.8% and 70.3%, respectively, for *Harry Potter and the Sorcerer's Stone*)."[97] A model cannot reproduce large portions of a copyrighted novel it never ingested. The researchers concluded that "extraction of (in-copyright) training data remains a risk for production LLMs."[98]

73. xAI's conduct fits the pattern Anna's Archive described: most, but not all, of the U.S.-based LLM developers that sought its data reconsidered. xAI has never indicated that it was among those that did so.[99] On information and belief, xAI, directly or through data intermediaries, obtained books from Anna's Archive, LibGen, Z-Library, Books3, and datasets derived from them, including through bulk downloads and torrents, and used them to train Grok.

74. The sources from which xAI obtained books distribute their collections in bulk principally through torrents: LibGen distributes its database through bulk archives and torrent files, and Anna's Archive makes its collections available for bulk acquisition through torrent-based distribution. Moreover, xAI's competitors that acquired shadow-library collections at scale did so by torrenting. Anthropic used BitTorrent to download at least five million books from LibGen and at least two million from the Pirate Library Mirror.[100] Meta "torrented" LibGen and Anna's Archive "[t]o download these large datasets more quickly and without unnecessarily slowing down its networks."[101] xAI, which needed the same data and needed it faster than anyone, had every reason to use the same method. The ultimate facts showing exactly how xAI obtained its training data, including its download records, torrent clients, and network logs, are exclusively within xAI's possession, and xAI has told a federal court that the sources of its datasets are trade secrets.

---

[97] Ahmed Ahmed et al., *Extracting Books from Production Language Models*, arXiv:2601.02671 (Jan. 6, 2026), https://arxiv.org/abs/2601.02671 [https://perma.cc/RPD2-A3QE].

[98] *Id.*

[99] *Copyright reform is necessary for national security*, *supra* note 76.

[100] *Bartz v. Anthropic PBC*, 791 F. Supp. 3d 1038, 1046 (N.D. Cal. 2025).

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

75. In torrenting or otherwise downloading from shadow libraries, xAI knew or should have known that it was participating in a peer-to-peer network that traded in unauthorized copies of copyrighted material. On information and belief, xAI chose to obtain shadow-library collections through BitTorrent knowing that the protocol requires participants to upload and ran torrent clients on its own servers and networks for that purpose. On information and belief, in the course of torrenting datasets from shadow libraries, xAI uploaded pieces of the copyrighted works in those datasets, including Plaintiffs' Works, to other peers, whether while leeching or seeding, and those peers downloaded them and assembled complete copies from the pieces xAI and others supplied. Those peers' downloading and redistribution constituted further infringement. There was no substantial or commercially significant non-infringing use of the materials xAI uploaded and distributed.

76. xAI knew the law and chose to ignore it. On November 29, 2023, less than four weeks after Grok's launch, Musk told an interviewer that the claim that AI models are not trained on copyrighted material is "a huge lie," that "[t]hese AIs are all trained on copyrighted data. Obviously," and that lawsuits over that training "won't be decided on a timeframe that's relevant."[102] On April 11, 2025, Musk endorsed the proposal to "delete all IP law":[103]

[102] Masnick, *supra* note 5.

[103] Elon Musk (@elonmusk), X (April 11, 2025), https://x.com/elonmusk/status/1910840422789763511 [https://perma.cc/VJN2-3MLX].



77. Musk has also publicly described the conduct alleged here as theft. On February 23, 2026, Musk responded on X to a post reporting that a competitor, Anthropic, had settled a $1.5 billion lawsuit over pirating more than seven million books from shadow libraries to train its models. Musk wrote: "Anthropic is guilty of stealing training data at massive scale and has had to pay multi-billion dollar settlements for their theft. This is just a fact."[104] When another user responded that xAI, too, trains on data it did not create, Musk did not deny it: "Yeah, but we're not super smug, sanctimonious and hypocritical about it like Anthropic is."[105] Musk thus publicly characterized the acquisition of books from shadow libraries for LLM training as theft, and as theft that had already cost a developer billions of dollars.

78. Musk's views are xAI's views. Musk founded xAI, controls it, directs its training strategy, and announces its data practices on X. He has used X to solicit training data for Grok

[104] Elon Musk (@elonmusk), X (Feb. 23, 2026), https://x.com/elonmusk/status/2026052687423562228 (replying to NIK (@ns123abc), X (Feb. 23, 2026)) [https://perma.cc/LB9Y-8W69].

[105] Elon Musk (@elonmusk), X (Feb. 23, 2026, 9:58 PM), https://x.com/elonmusk/status/2026053980917465388 (replying to Matter as Machine (@matterasmachine), X (Feb. 23, 2026)) [https://perma.cc/6CAL-8S8X].

directly from the public, and in June 2025 announced a plan to use Grok to "rewrite the entire corpus of human knowledge" and "retrain on that."[106] One cannot rewrite a corpus one has not copied.

79. xAI was also on specific notice. By the time xAI trained Grok 3, Grok 4, and their successors, Anthropic, Meta, OpenAI, and NVIDIA had each been sued for training LLMs on LibGen, Books3, and similar pirated datasets, and Z-Library's operators had been indicted. On December 22, 2025, a complaint was filed against xAI in this District alleging that xAI trained Grok on pirated books.[107] xAI's only public response was a three-word email to the press: "Legacy Media Lies."[108] In June 2026, the court held that the plaintiffs had not alleged a common transaction or occurrence linking their claims against the competing defendants, and it severed and dismissed xAI from the suit without reaching the merits.[109] xAI then continued crawling through August 2026 and released Grok 4.5, Grok 4.6, and Grok 4.7 in the months that followed.[110]

80. Upon downloading this material, xAI manipulated it. xAI admits that it subjects such material to "preprocessing such as quality filtering, deduplication, and safety filtering before training," to "filtering, keyword-based rules and model-based classifiers," and to "cleaning" that "remove[s] duplicate or incomplete entries."[111] On information and belief, in the course of that processing xAI removed and altered text and information in and on Plaintiffs' Works, including titles, author names, copyright owner information, and copyright notices, without the authority of the copyright owners. On information and belief, xAI removed that information knowing that Grok

---

[106] Bastian, *supra* note 31.

[107] Compl., *Carreyrou et al. v. Anthropic PBC*, No. 5:25-cv-10897 (N.D. Cal. Dec. 22, 2025), ECF No. 1 ¶¶ 90–104.

[108] Annelise Levy, *OpenAI, Anthropic, xAI Hit With Copyright Suit from Writers*, BLOOMBERG LAW (Dec. 22, 2025), https://news.bloomberglaw.com/ip-law/openai-anthropic-xai-hit-with-copyright-lawsuit-from-writers [https://perma.cc/7N8Q-TURV].

[109] Order Granting Motions to Sever, *Carreyrou et al.. v. Anthropic PBC*, No. 5:25-cv-10897-PCP (N.D. Cal. June 8, 2026), ECF No. 195.

[110] Grok 4.7 Training Summary § 2.3, at 4 (crawler collection period "[f]rom January 2024 to August 2026"); xAI, *Company*, https://x.ai/company (timeline) [https://perma.cc/VRS5-U776].

[111] Grok 4.7 Training Summary at 3, 6; xAI PI Mot. at 4–5.

memorizes text from the books on which it is trained and reproduces that text in response to user prompts, and so that text Grok reproduces from those books reaches users without the titles, author names, and copyright notices that identify it as copyrighted and identify its owner.

81. As alleged above, Plaintiffs' Works appear in shadow libraries and pirated datasets including Books3, LibGen, Z-Library, and Anna's Archive. There is accordingly a reasonable inference that xAI downloaded, manipulated, reproduced, and distributed Plaintiffs' Works.

82. xAI's scanning program removed copyright management information as well. Each of Plaintiffs' Works, as published in print, contains a copyright page that states the work's title and author, identifies the copyright owner, and bears a copyright notice. xAI's scans captured that page. On information and belief, in converting the scanned images into training text, xAI removed the copyright page and other front matter from each book, separating the text of the work from the information that identifies its author and owner.

83. xAI has never disclosed the specific sources of Grok's training data. It has instead told a federal court that those sources are trade secrets, that it provided only "a high-level, limited disclosure that does not reveal its trade secrets," and that disclosure would let competitors "acquire the sources they are missing."[112] xAI posted that limited disclosure on December 30, 2025, the day after it sued California.[113] xAI did this, in part, to conceal that it relied and continues to rely on pirated materials to train Grok.

84. Plaintiffs did not discover, and could not reasonably have discovered, that xAI had copied their works. Nothing available to Plaintiffs connected xAI to their works. xAI has never identified any specific shadow library, book dataset, or book-scanning program as a source of Grok's training data. It has instead told a federal court that the sources of its datasets are trade secrets, and it responded to the first lawsuit alleging that it trained Grok on pirated books with the

[112] xAI PI Mot. at 7, 9.

[113] Order Denying Pl.'s Mot. for Prelim. Inj. at 3, *X.AI LLC v. Bonta*, No. 2:25-cv-12295-JGB-SSC (C.D. Cal. Mar. 4, 2026), ECF No. 35 (noting that xAI published its "high-level, limited disclosure" on December 30, 2025); *id.* at 1 (complaint filed December 29, 2025).

words "Legacy Media Lies." In the alternative, any applicable limitations period was equitably tolled by Defendants' concealment of the sources of their training data.

85. Each of xAI's infringing acts gave rise to a separate claim when it occurred. xAI's crawler collected training data from January 2024 through August 2026. xAI released Grok-1 (November 2023), Grok-1.5 (March 2024), Grok-2 (August 2024), Grok 3 (February 2025), Grok 4 (July 2025), Grok 4 Fast (September 2025), Grok 4.1 (November 2025), Grok 4.5 (July 2026), Grok 4.6 (August 2026), and Grok 4.7 (September 2026), and as of January 2026 was training Grok 5. xAI's book-scanning program was operating no later than July 2026. On information and belief, each successive Grok model was developed using data and outputs from earlier Grok models, including synthetic data generated by those models, so that Plaintiffs' protected expression ingested for earlier models persisted in, and was carried forward into, later models.[114]

86. xAI has incorporated Grok into X, the grok.com website, standalone iOS and Android applications, Tesla vehicles, Microsoft Office add-ins, and enterprise platforms including Amazon Bedrock, Microsoft Foundry, and Google's Gemini Enterprise, and it sells access to Grok through consumer subscriptions, enterprise contracts, government contracts, and its API.[115] xAI considers Grok and the technology underpinning it to be the core of its commercial value. Its $20 billion Series E financing and its acquisition by SpaceX rested on that value.

E. xAI's Scanning of Physical Books to Train Grok

87. xAI's appetite for books did not abate after it was sued for training Grok on pirated copies. Nor did xAI turn to the licensing market for Plaintiffs' Works. Instead, xAI began acquiring physical copies of books and scanning them into digital form for use as training data.

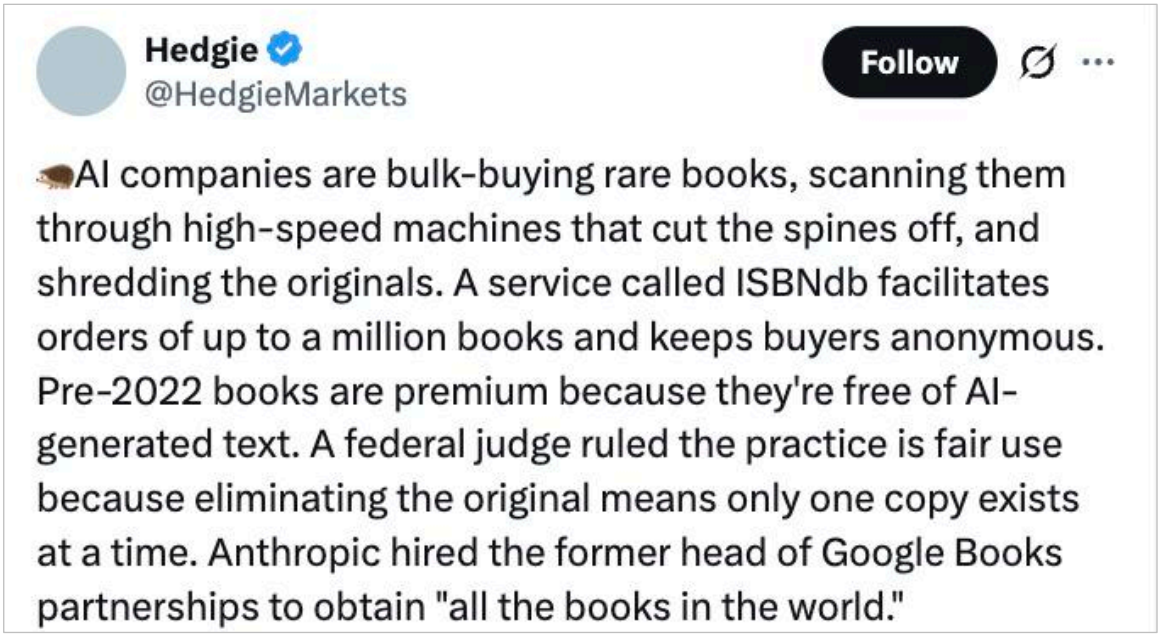
88. The practice was pioneered by xAI's competitors. Anthropic purchased millions of print books, ripped off their spines, scanned the pages, and discarded the originals; Meta ran a

[114] Grok 4.7 Training Summary § 2.5, at 5 (stating that Grok 4.6 was used to generate supervised fine-tuning trajectories).

[115] xAI, *Company*, https://x.ai/company (timeline) [https://perma.cc/VRS5-U776]; Series E Announcement.

similar program, internally called "Project Vellum," beginning in mid-2025.[116] By July 2026, 404 Media reported that AI companies were buying old books in bulk to scan them for training data.[117]

89. On July 26, 2026, a widely followed X account summarized that reporting: "AI companies are bulk-buying rare books, scanning them through high-speed machines that cut the spines off, and shredding the originals."[118]



90. The next day, Musk replied:[119]

[116] *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1013–14 (N.D. Cal. 2025) (describing Anthropic's purchase of millions of print books, which it stripped of their bindings and scanned before discarding the originals, and its separate amassing of a "central library" of pirated books); *Entrepreneur Media, LLC v. Meta Platforms, Inc.*, No. 25-cv-09579-VC (N.D. Cal.), ECF No. 113 at 3.

[117] Emanuel Maiberg, *AI Companies Are Buying Tons of Old Books Because They're Free of AI Slop*, 404 MEDIA (July 21, 2026), https://www.404media.co/ai-companies-are-buying-tons-of-old-books-because-theyre-free-of-ai-slop/ [https://perma.cc/S7YL-2BA5].

[118] Hedgie (@HedgieMarkets), X (July 26, 2026), https://x.com/HedgieMarkets/status/2081534588485296565 [https://perma.cc/8VAH-ZMUU].

[119] Elon Musk (@elonmusk), X (July 27, 2026, 4:48 PM), https://x.com/elonmusk/status/2081844165881594362 [https://perma.cc/K7B5-SR47].



91. "SpaceXAI" is xAI's former name. Musk's post is an admission by xAI's founder and controlling executive that xAI operates a program to scan physical books for AI training.[120] The post reveals that xAI acquires physical books and scans them. It leaves no doubt that Musk is directly involved and has personally directed the program. By directing a change of course for rare books, the post shows that prior to July 27, 2026, xAI had been scanning books, rare ones included, by the method Musk contrasted with "just cutting off the spine and scanning." Finally, this post confirms that for books xAI deems "rare," xAI now retains the physical original "in a library" *and* the digital scan.

92. xAI has not disclosed this program anywhere else. Its EU training-content summaries, filed on August 12 and September 21, 2026, describe "publicly available" data, licensed data, crawled data, user data, and synthetic data, and nowhere mention the scanning of physical books.[121] On information and belief, xAI acquired physical books in bulk, through bulk book

[120] *Id.*; *see also* Adrienne Martinez, *Elon Musk Hits AI Firms Destroying Books for Training Data, Promises SpaceXAI Will Scan 'The Hard Way'*, IBTIMES UK (July 31, 2026), https://www.ibtimes.co.uk/elon-musk-preserve-rare-books-ai-scanning-1811757 [https://perma.cc/VB58-C6ED]; Jason Nelson, *AI Book Burning? Companies Are Destroying Millions of Books to Feed Chatbots*, DECRYPT (July 29, 2026), https://decrypt.co/374646/ai-book-burning-companies-destroying-millions-books-chatbots [https://perma.cc/VAT3-XVRP].

[121] *See generally* Grok 4.7 Training Summary; Grok 4.5/4.6 Training Summary.

vendors or directly from booksellers, scanned them into digital images, ran many, if not all of the images through optical character recognition software to extract their text, and processed, cleaned, de-duplicated, tokenized, stored, and retained the resulting additional digital copies for its use, including, among other uses, in training Grok, including Grok 4.5, Grok 4.6, Grok 4.7, and Grok 5.

93. At no point did xAI seek or obtain authorization from Plaintiffs, or from authors and publishers generally, to reproduce their works by scanning. xAI did not license the right to scan, digitize, copy, or use these works as training data. It bought a copy of each book and treated the purchase as though it had acquired the copyright owner's exclusive rights.

94. It had not. The purchase of a physical copy of a book conveys ownership of that copy. Ownership of a material object in which a work is embodied does not convey any rights in the copyrighted work itself. 17 U.S.C. § 202. The first-sale doctrine permits the owner of a lawfully made copy to sell or otherwise dispose of that copy, 17 U.S.C. § 109(a), but it confers no right to reproduce the work. The exclusive right to reproduce a copyrighted work in copies belongs to the copyright owner alone. 17 U.S.C. § 106(1).

95. Scanning a book is reproduction. Each time xAI scanned a physical copy of a book, it created a new digital copy of the work. Each time xAI ran the scanned images through OCR, it created another copy, this time as machine-readable text. Each time xAI processed, "cleaned," de-duplicated, and tokenized that text, it created still more copies. And each time xAI used that text in a training run, including in the internal and experimental models that precede each public release, it created additional copies of the work.

96. xAI, by Musk's own account, keeps the print copy of every "rare" book in a library and keeps the scan too. That is not a change of format. It is the creation of a second copy, and then a third, fourth, and fifth, on top of the first. To the extent xAI cuts off the spines of books it does not deem rare, scanning them likewise reproduces the entire work. Under either method, xAI retained digital copies of entire works in multiple forms and locations for commercial use.

97. xAI launched its scanning program with full knowledge that the books it was scanning were protected by copyright. Every scanned book carried its copyright page, and xAI scanned that page along with the rest. xAI also launched the program with full knowledge of the

market for licensing books for AI training: by mid-2026 xAI had itself entered into licensing agreements for text, and numerous other AI developers had done the same. xAI chose to purchase single copies of books rather than pay for training licenses because it was much faster and it was cheaper. The retail price of a book compensates the author for one reader's copy; it does not compensate the author for the right to reproduce the work wholesale and to build a commercial product on its expressive content. xAI's scanning program was designed to bypass the licensing market, and it did.

98. On information and belief, xAI acquired, scanned, digitized, reproduced, and used physical copies of Plaintiffs' Works as part of its scanning program, and used the resulting digital copies to train one or more Grok models.

99. xAI's reproduction and use of Plaintiffs' Works through its scanning program is an independent basis of liability, separate and apart from xAI's acquisition, reproduction, and distribution of pirated copies of Plaintiffs' Works from shadow libraries. Even if xAI had never downloaded a single pirated copy of any of Plaintiffs' Works, its unauthorized scanning, digitization, reproduction, and exploitation of physical copies of those works would independently violate Plaintiffs' exclusive rights under 17 U.S.C. § 106.

F. xAI's Conduct Harms the Market for Plaintiffs' Copyrighted Works

100. xAI's conduct has a detrimental effect on the potential market for and value of Plaintiffs' works, including by developing products that create and are capable of creating content that serves as a direct substitute for Plaintiffs' works, developing products that create and are capable of creating content that serves as an indirect substitute for Plaintiffs' works, and undermining Plaintiffs' ability to participate in and profit from the market for licensing their works for LLM training.

101. *First*, xAI's decision to download and use unauthorized copies of Plaintiffs' works from shadow libraries, and to scan and use physical copies of Plaintiffs' works without a license, deprived Plaintiffs of licensing fees they could otherwise have earned had they granted permission to use their works for xAI's desired purpose. xAI paid nothing for the pirated copies acquired from shadow libraries—at all. And as alleged herein, there is an existing market for licensing copyrighted

works such as Plaintiffs' for use in LLM development, and xAI itself participates in that market for works it chooses to acquire lawfully. xAI bypassed that market for Plaintiffs' Works and deprived Plaintiffs of the revenue and bargaining power they otherwise would have had. xAI now lets X users opt out of having their posts used to train Grok; it never gave Plaintiffs that choice.[122]

102. *Second*, Grok, trained on Plaintiffs' copyrighted works and the protected works of others, is capable of generating outputs that compete directly with, and risk serving as replacements for, Plaintiffs' works. As the U.S. Copyright Office has warned, "the speed and scale at which AI systems generate content pose a serious risk of diluting markets for works of the same kind as in their training data."[123] The AI industry's own executives have admitted as much. In the consolidated copyright litigation against OpenAI and Microsoft, the record shows that OpenAI's Head of ChatGPT wrote that publishers face an "existential threat" from generative AI products, which "are largely substitutive, period" and "will get more and more substitutive as they get better."[124] Microsoft's CEO agreed under oath that conversing with AI chatbots "has substituted" for "needing to go to the underlying source."[125] A Microsoft Director of Applied Science described the copying at issue as "an astonishing theft of unprecedented proportions" and perhaps the "largest theft of labor in human history."[126] As Judge Chhabria recognized in this District, copyright owners "will often win" when their works are used to build AI products that substitute for them.[127]

103. *Third*, even where Grok is restricted from outputting extended portions of verbatim text from copyrighted works, it is capable of producing nearly indistinguishable "versions" of

---

[122] Grok 4.7 Training Summary at 6 ("SpaceXAI does not train on X data from X users who have opted out of training.").

[123] U.S. Copyright Office, *Copyright and Artificial Intelligence, Part 3*, at 65.

[124] *In re OpenAI, Inc., Copyright Infringement Litig.*, No. 1:25-md-03143-SHS-OTW (S.D.N.Y. Sept. 17, 2026), ECF No. 1977-1 at 1.

[125] *Id.* at 2.

[126] *Id.* at 1.

[127] *Kadrey*, 788 F. Supp. 3d at 1059.

copyrighted works such that a consumer would use the AI-generated version rather than pay for a copy of the actual work. As alleged above, researchers extracted roughly seventy percent of the text of one best-selling novel from Grok 3, as measured by the study's "nv-recall" metric, without any jailbreak at all.

V. CLAIMS FOR RELIEF

COUNT I

Direct Copyright Infringement (17 U.S.C. § 501):
Pirated Copies Obtained from Shadow Libraries

104. Plaintiffs incorporate the allegations above.

105. Plaintiffs are the legal or beneficial owners of the copyrighted works listed in Exhibit A (the Plaintiffs' Works).

106. Defendants, without authorization from Plaintiffs, copied, downloaded, reproduced, ingested, parsed, embedded, and/or used pirated and/or otherwise unauthorized copies of Plaintiffs' Works for commercial purposes, including in the development, training, fine-tuning, and deployment of the Grok family of models. These acts violated Plaintiffs' exclusive rights under 17 U.S.C. § 106.

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

- acquired Plaintiffs' Works from shadow-library repositories and from datasets containing pirated works from shadow libraries;
- on information and belief, distributed Plaintiffs' Works through the use of torrenting software, programs, or protocols;
- reproduced additional copies during ingestion, preprocessing, storage, deduplication, formatting, and tokenization;
- retained copies after training was complete; and
- while training their models, made still more copies of the text, because every training pass (each epoch and each step of gradient descent) requires creating and working with fresh versions of that text.

108. Defendants' reproductions and distributions of Plaintiffs' Works were made without permission, license, or consent and violated Plaintiffs' exclusive rights under the Copyright Act.

109. Defendants' infringement was willful. As alleged above, Defendants, on information and belief, knowingly pirated Plaintiffs' works from shadow-library sources they knew to be illegal, as their own use of the USTR Notorious Markets list confirms; trained their models on datasets saturated with pirated books, including Plaintiffs' Works; publicly dismissed copyright claims as irrelevant and endorsed the abolition of intellectual property law; publicly condemned a competitor's training on pirated books as theft after being sued for the same conduct; concealed the composition of their training datasets and litigated to keep them secret; and continued copying after lawsuits, law-enforcement seizures, and industry-wide warnings made the illegality unmistakable. This conduct meets the standard for criminal copyright infringement.[128]

110. 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.

111. Plaintiffs have been injured by Defendants' willful acts of copyright infringement. Plaintiffs are entitled to statutory damages, actual damages, restitution of profits, and other remedies in law or equity.

112. Plaintiffs are entitled to recover attorneys' fees and costs under 17 U.S.C. § 505.

COUNT II

Direct Copyright Infringement (17 U.S.C. § 501): Scanning of Physical Books

113. Plaintiffs incorporate the allegations above.

114. Plaintiffs are the legal or beneficial owners of the copyrighted works listed in Exhibit A (the Plaintiffs' Works).

115. Defendants, without authorization from Plaintiffs, acquired physical copies of Plaintiffs' Works, scanned those copies into digital form, ran many, if not all of the scanned images through optical character recognition software to extract the text of Plaintiffs' Works, and processed,

[128] 17 U.S.C. § 506(a)(1)(A).

cleaned, de-duplicated, tokenized, stored, and retained the resulting additional digital copies. Defendants used those digital copies for their own commercial purposes, including but not limited to efforts to train, develop, fine-tune, evaluate, and deploy the Grok family of models. Each of these acts reproduced Plaintiffs' Works in copies, in violation of Plaintiffs' exclusive rights under 17 U.S.C. § 106(1).

116. Defendants' purchase of physical copies of Plaintiffs' Works did not authorize any of these reproductions. Ownership of a physical copy conveys no right to reproduce the copyrighted work embodied in it. 17 U.S.C. §§ 109(a), 202. Defendants never sought or obtained permission or a license from Plaintiffs to scan, digitize, copy, or otherwise reproduce Plaintiffs' Works, or to use Plaintiffs' Works as training data.

117. Defendants' reproductions of Plaintiffs' Works through their scanning program were not a one-for-one change of format for Defendants' own reading or reference. They were wholesale reproductions of entire works, made and retained in multiple copies, and, for books Defendants deemed rare, retained alongside the physical original, for the purpose of building a commercial training corpus and commercial AI products, in a manner that supplants the existing market for licensing Plaintiffs' Works for AI training and that threatens to dilute the market for Plaintiffs' Works.

118. Defendants' infringement through their scanning program is independent of, and in addition to, their infringement through the acquisition, reproduction, and distribution of pirated copies of Plaintiffs' Works alleged in Count I. Each is a separate and independent basis for liability.

119. Defendants' infringement was willful. Defendants undertook their scanning program after they had been sued for training Grok on pirated books; with actual knowledge that the books they scanned bore copyright notices; with knowledge of an existing market for licensing books for AI training, in which Defendants themselves participate; and with the deliberate purpose of avoiding the cost of that market.

120. 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.

121. Plaintiffs have been injured by Defendants' willful acts of copyright infringement. Plaintiffs are entitled to statutory damages, actual damages, restitution of profits, and other remedies in law or equity.

122. Plaintiffs are entitled to recover attorneys' fees and costs under 17 U.S.C. § 505.

COUNT III

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

123. Plaintiffs incorporate the allegations above.

124. On information and belief, Defendants used torrenting software, programs, or protocols to download datasets containing pirated copies of works, including Plaintiffs' Works.

125. In connection with their torrenting of datasets that contained copyrighted works, Defendants uploaded and distributed, through seeding or leeching, copyrighted materials, including Plaintiffs' Works, thereby making those works available to third parties for downloading on peer-to-peer networks.

126. Defendants knowingly participated in peer-to-peer sharing networks that they knew trafficked in pirated copies of copyrighted materials. Defendants knew that others on these networks were infringing copyrighted materials through reproduction and distribution. There was no substantial or commercially significant non-infringing use of the copyrighted materials that Defendants uploaded and distributed, nor of Defendants' uploading and distribution of Plaintiffs' Works. By participating in these networks, and by further uploading and distributing Plaintiffs' Works, Defendants materially contributed to and induced further infringement of Plaintiffs' Works.

127. By knowingly inducing and materially contributing to others' infringement of Plaintiffs' Works, Defendants are liable for contributory copyright infringement.

128. As a direct and proximate result of Defendants' conduct, Plaintiffs were injured and are entitled to statutory damages, actual damages, restitution of profits, and other remedies in law or equity.

129. Plaintiffs are entitled to recover attorneys' fees and costs under 17 U.S.C. § 505.

COUNT IV

Removal of Copyright Management Information (17 U.S.C. § 1202(b)(1))

130. Plaintiffs incorporate the allegations above.

131. Plaintiffs' Works contain information that constitutes "copyright management information" as that term is defined in 17 U.S.C. § 1202(c), including identifying information such as each work's title, author information, year of publication, information about the copyright owner, and copyright notices, appearing on copyright pages and other front matter and, in digital copies, in embedded EPUB or PDF metadata fields.

132. Upon downloading and scanning copyrighted materials, including Plaintiffs' Works, Defendants processed the data and, in doing so, removed and altered text and information, including copyright management information found in and on Plaintiffs' Works. Defendants did so without the authority of the copyright owners.

133. Defendants' removal of copyright management information was intentional. On information and belief, Defendants removed it through filtering and processing steps that targeted copyright pages, copyright notices, and other front matter.

134. Defendants removed copyright management information from Plaintiffs' Works knowing or having reasonable grounds to believe that doing so would enable, facilitate, and conceal copyright infringement. As to concealment, Defendants knew or had reasonable grounds to believe that stripping copyrighted works of copyright management information would make it harder for others to discover the true sources of Grok's training data. Defendants' refusal to disclose those sources, and their litigation to keep them secret, confirm the point. Defendants knew that Grok memorizes text from the books on which it is trained and reproduces that text in response to user prompts; independent researchers extracted roughly seventy percent of the text of one copyrighted novel from Grok 3 without any jailbreak. Removing copyright management information from the training copies ensures that when Grok reproduces text from Plaintiffs' Works, that text reaches users without any title, author, owner, or copyright notice. Users who receive it cannot tell that it is copyrighted or whom it belongs to, and Plaintiffs cannot identify Grok's outputs as copies of their

works. Defendants thereby enabled and facilitated infringement by Grok and its users and concealed that infringement from Plaintiffs.

135. Plaintiffs were harmed by Defendants' removal of copyright management information from their works and are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law. Plaintiffs are entitled to recover attorneys' fees and costs under 17 U.S.C. § 1203(b)(5).

PRAYER FOR RELIEF

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

a. Judgment in favor of Plaintiffs against Defendants;

b. A declaration that Defendants have infringed Plaintiffs' exclusive copyrights under the Copyright Act;

c. A declaration that such infringement is willful;

d. A declaration that Defendants violated 17 U.S.C. § 1202(b) through their removal of copyright management information;

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

f. An injunction prohibiting Defendants from removing or altering copyright management information from Plaintiffs' Works, and requiring Defendants to preserve records of the provenance of, and the processing applied to, every copy of Plaintiffs' Works used in training;

g. That Defendants be directed to account to Plaintiffs for all gains, profits, and advantages derived from their unlawful acts;

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

i. An award of statutory or actual damages under 17 U.S.C. § 1203(c);

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

k. Pre- and post-judgment interest on the damages awarded to Plaintiffs; and

l. Further relief for Plaintiffs as the Court may deem just and proper.

JURY TRIAL DEMANDED

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

Dated: November 4, 2026

Respectfully submitted,

/s/ Elizabeth Brannen

Elizabeth Brannen (SBN 226234)
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
lmartin@stris.com

Kyle Roche (*pro hac vice* forthcoming)
Joseph Delich (*pro hac vice* forthcoming)
FREEDMAN NORMAND FRIEDLAND LLP
155 E. 44th Street, Ste 915
New York NY 10017
T: (646) 494-2900
kroche@fnf.law
jdelich@fnf.law

Counsel for Plaintiffs