Joseph R. Saveri (SBN 130064)
Diane S. Rice (SBN 118303)
Christopher K.L. Young (SBN 318371)
William W. Castillo Guardado (SBN 294159)
SAVERI LAW FIRM, LLP
550 California Street, Suite 910
San Francisco, CA 94104
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
drice@saverilawfirm.com
cyoung@saverilawfirm.com
wcastillo@saverilawfirm.com

Attorneys for Individual and Representative Plaintiff E. Molly Tanzer

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

SAN FRANCISCO DIVISION

E. MOLLY TANZER, an individual, Individual and Representative Plaintiff, v. ADOBE INC. Defendant.	Case No. **COMPLAINT** **JURY TRIAL DEMANDED** **CLASS ACTION**

TABLE OF CONTENTS

Plaintiff E. Molly Tanzer ("Plaintiff"), on behalf of herself and all others similarly situated, brings this Complaint against Defendant Adobe Inc. ("Adobe").

OVERVIEW

1. Adobe engaged in a massive piracy scheme to develop the Nemotron and SlimLM family of large language models ("LLMs"). Specifically, Adobe partnered with NVIDIA to develop "next-generation NVIDIA AI foundation models," including the Nemotron family of models trained using millions of books. No public information exists showing that Adobe or NVIDIA licensed millions of books to train their AI models. Upon information and belief, Adobe and NVIDIA were only able to obtain such a large number of books by sourcing them from pirated sources, such as Anna's Archive, a shadow library that purposefully flouts copyright law and illegally distributes millions of copyrighted books.

2. Adobe also acquired the notorious Books3 dataset of nearly 200,000 pirated books to develop its SlimLM LLMs. The Books3 dataset is a subset of the SlimPajama dataset that Adobe acquired. Books3 contains the entirety of copyrighted books from the Bibliotik shadow library that were obtained without the authorization or consent of the authors.

3. Plaintiff and Class members are authors. They own registered copyrights in certain books that were included in datasets originating from Books3 and Anna's Archive that were used by Defendant and which were copied, stored, and used without their permission or compensation.

4. Adobe admits that it must compensate rightsholders, obtain their authorization, and respect their copyrights before including their works in training data for AI models. For example, Adobe's Firefly AI models are trained on *licensed* images whose creators are paid on a yearly basis for the inclusion of those images in Firefly's training data. In contrast, book authors have *never* received compensation for Adobe's use, storage, or copying of their works, nor have these authors ever provided Adobe with authorization to use, copy, or store their works. Adobe's claims of "respecting" the creative community ring hollow.

5. Adobe benefitted commercially from its acts of massive copyright infringement, including by securing contracts with enterprise customers for use of its LLMs, and by incorporating

the infringing Nemotron and SlimLM models into Adobe programs and tools, including Adobe Acrobat.

6. Through the above acts, Defendant has infringed Plaintiff's copyrighted works, and it continues to do so by continuing to store, copy, use, and process datasets containing copies of Plaintiff's and the putative Class's copyrighted books.

JURISDICTION AND VENUE

7. This Court has subject-matter jurisdiction under 28 U.S.C. § 1331 and §1338(a) because this is a civil action arising under the copyright laws of the United States, including 17 U.S.C. § 501.

8. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Adobe is headquartered in this district. Adobe acquired the Books3 dataset to create the SlimLM series of large language models, and, upon information and belief, developed the Nemotron models using datasets sourced from the shadow library Anna's Archive. Defendant commercially benefited from the creation of these models. Therefore, a substantial part of the events giving rise to the claim occurred in this District. A substantial portion of the affected interstate trade and commerce was carried out in this District. Defendant has transacted business, maintained substantial contacts, and/or committed overt acts in furtherance of the illegal scheme and conspiracy throughout the United States, including in this District. Defendant's conduct has had the intended and foreseeable effect of causing injury to persons residing in, located in, or doing business throughout the United States, including in this District.

9. Under Civil Local Rule 3-2(c), assignment of this case to the San Francisco Division is proper because this case pertains to intellectual-property rights, which is a district-wide case category under General Order No. 44, and therefore venue is proper in any courthouse in this District.

PARTIES

A. PLAINTIFF

10. Plaintiff E. Molly Tanzer is an author residing in Colorado who owns registered copyrights in multiple books, including *Creatures of Will and Temper*, *Creatures of Charm and Hunger*, and *Vermilion*.

11. A non-exhaustive list of registered copyrights owned by Plaintiff is included as Exhibit A.

B. DEFENDANT

12. Defendant Adobe Inc. is a Delaware corporation headquartered at 345 Park Avenue, San Jose, CA 95110.

C. AGENTS AND CO-CONSPIRATORS

13. The unlawful acts alleged against Defendant in this first amended consolidated complaint were authorized, ordered, or performed by the Defendant's respective officers, agents, employees, representatives, or shareholders while actively engaged in the management, direction, or control of the Defendant's businesses or affairs. The Defendant's agents operated under the explicit and apparent authority of their principals. Defendant and its subsidiaries, affiliates, and agents operated as a single unified entity.

14. Various persons or firms not named as defendants may have participated as co-conspirators in the violations alleged herein and may have performed acts and made statements in furtherance thereof. Each acted as the principal, agent, or joint venture of Defendant with respect to the acts, violations, and common course of conduct alleged herein.

FACTUAL ALLEGATIONS

15. Generative artificial intelligence ("AI") models are designed to generate output—such as text, images, video, or audio—in response to user prompts. An AI model is not created the way most software programs are—that is, by human software programmers writing code. Rather, an AI model is *trained* by copying an enormous quantity of data and then feeding these copies into the model. This corpus of input material is called the *training dataset*.

16. These models are trained with vast quantities of data to allow the output to closely recreate the content of its training data. For example, an image model instructed to generate a picture of a cat will emit a picture based on pictures of cats used to train the model.

17. A common type of generative AI model is known as a large language model ("LLM") designed to emit text outputs in response to user prompts. Often, the models are created to generate text mimicking human language.

18. Training consists of a multi-stage process that includes the acquisition and curation of the dataset, processing of the dataset, feeding the dataset into the model so that the model can extract the patterns and relationships from the protected expression contained therein, and further fine-tuning the model for more specialized uses with even more data.

19. The first step in training the model is acquiring and curating the data that goes into the model. Training an AI model is not only a function of quantity of data, but also of quality. As Adobe's own General Counsel Dana Rao acknowledged, generative AI "is only as good as the data on which it's trained."[1] The selection and curation of training data is therefore an important first step in training.

20. Copyrighted books tend to be high-quality data for training LLMs. Studies have found models to perform best when their training data comprises books,[2] such that when books are removed from the training mixture of AI models, model performance degrades more than with the removal of other types of training data.[3] Put differently, the inclusion of books in the training data mixture has a disproportionately large impact on a model's capabilities relative to their share of training tokens.[4] For these reasons, the inclusion of books in the training mixture of LLMs is ***essential.***

[1] Rachel Metz & Brody Ford, *Adobe's 'Ethical' Firefly AI Was Trained on Midjourney Images*, YAHOO FINANCE (Apr. 12, 2024), https://finance.yahoo.com/news/adobe-ethical-firefly-ai-trained-123004288.html (on file with counsel).

[2] Jack W. Rae, et al., *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*, (Jan. 21, 2022) [https://arxiv.org/abs/2112.11446].

[3] Shayne Longpre, et al., *A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity*, in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) 3245–3276, 3246 (2024).

[4] Yang Zhao et al., *Deciphering the Impact of Pretraining Data on Large Language Models through Machine Unlearning,* in *Findings of the Association for Computational Linguistics: ACL 2024* 9386–9406 (2024).

21. During training, the AI model copies and ingests each work in the training dataset and extracts protected expression from it. During what is known as *pretraining*, the AI model progressively adjusts its output to more closely approximate the protected expression copied from the training dataset. The AI model records the results of this process in a large set of numbers called weights (also known as *parameters*) that are stored within the model. These weights are entirely and uniquely derived from the protected expression in the training dataset. Once a model is pretrained, it results in a trained model known as a *base* or *foundational* model.

22. After a model is pretrained, it may then be *fine-tuned* with additional data. According to Adobe:

> During the build and train phase of AI development, training is not a "one and done" process. The model typically must be "tuned," that is, trained on either a larger, more diverse dataset or on a completely different dataset than originally expected. For example in the initial training of Adobe Firefly, the model produced hands with more than five fingers. This resulted in the product team realizing that they needed more images of hands in the dataset to improve the model.[5]

23. Engineers may also conduct experiments known as "ablation studies" that test the effect of certain data on the model. This can include, for example, determining whether there is a difference in the quality of a model's output if it is trained with books versus without. A dataset may be used to run such experiments but be excluded from the final training dataset of the model. Importantly, these datasets too may consist of copyrighted works, including books.

24. An LLM, once it has copied and ingested the textual works in the training dataset and transformed the protected expression into stored weights, is able to emit convincing simulations of natural written language in response to user prompts. Whenever an LLM generates text output in response to a user prompt, it is performing a computation that relies on these stored weights, with the goal of imitating the protected expression ingested from the training dataset.

25. Recent studies confirm that LLMs are capable of *memorizing* their training data. LLMs are able to emit verbatim or reproduce near-verbatim portions of their training data. One study found

[5] *Adobe Generative AI Built for Business Solution Brief*, ADOBE (2024), https://www.adobe.com/cc-shared/assets/pdf/trust-center/ungated/whitepapers/corporate/adobe-gen-ai-built-for-business-solution-brief.pdf (on file with counsel).

that the LLMs of several leading AI developers—Anthropic, Google, OpenAI, and xAI—can regurgitate substantial portions of their training data, even with the implementation of safeguards designed to prevent the regurgitation of training data.[6] Another recent study found that fine-tuning LLMs from OpenAI, Google, and DeepSeek with a small number of books unlocks the ability of those models to reproduce up to 90% of text from unrelated copyrighted books, "using only semantic descriptions as prompts and no actual book text."[7] The models identified in these studies rely on the same or similar architecture as the Nemotron and SlimLM models developed by Adobe. Upon information and belief, models developed by Adobe are capable of memorizing and therefore regurgitating substantial portions of their training data.

26. Throughout each step of the training process, the same dataset may be used multiple times. Indeed, given the cost of developing an AI model, it is a ubiquitous practice to retain datasets for future use, whether that use is to pretrain other models, to perform ablations on a model, or to fine-tune an already trained base model. The implication is that if a dataset contains unlawfully-obtained copyrighted material, each step of the training process may result in an unauthorized use (i.e., infringement) of that copyrighted work.

27. Adobe, founded in 1982, provides services and tools for image, audio, video, and text editing and viewing. It is well known as the creator of Photoshop (an image manipulation program), and Adobe Acrobat (a program to view and manipulate files stored in Adobe's proprietary Portable Document Format ("PDF") format).

28. Adobe has in recent years added artificial intelligence capabilities into its products. Photoshop, for example, now incorporates a family of generative AI models known as Adobe Firefly that are capable of generating and manipulating images, videos, and audio. Adobe claims that the output of its Adobe Firefly models are "commercially safe" because Adobe used only licensed or public domain images and videos to develop these models. According to Adobe, these images are "vetted and

[6] Ahmed Ahmed et. al., *Extracting Books from Production Language Models* (Jan. 6, 2026) [https://arxiv.org/pdf/2601.02671].

[7] Xinyue Liu et al., *Alignment Whack-a-Mole: Finetuning Activates Verbatim Recall of Copyrighted Books in Large Language Models* (2026) [https://arxiv.org/abs/2603.20957].

humanly verified to be copyright compliant,"[8] and "customers can use those commercially safe solutions to confidently publish outputs knowing Adobe has ***responsibly addressed licensing and copyright issues***."[9]

29. Adobe Chief Strategy Officer Scott Belsky described other models in the industry as built on data that is "openly scraped,"[10] unlike Adobe's AI models that "show respect for the creative community" by training only on licensed content or public domain content where the copyright has expired.[11] Adobe CEO Shantanu Narayen further noted:

> I think other companies . . . are not completely transparent yet about what data they use and [if] they scrape the internet, and that will play out in the industry. But I like the approach that we've taken, and I like the way in which we've engaged with our community on this.[12]

30. Adobe claims to purchase or acquire images for which "Adobe has monetarily compensated the creator, secured the consent of person photographed, or confirmed that the copyright license has expired."[13]

31. Adobe even pays the creators of licensed works a yearly "Firefly Contributor Bonus" based on the number of images used to train the Firefly models.[14]

---

[8] *Adobe Generative AI Built for Business Solution Brief*, ADOBE (2024), https://www.adobe.com/cc-shared/assets/pdf/trust-center/ungated/whitepapers/corporate/adobe-gen-ai-built-for-business-solution-brief.pdf (on file with counsel).

[9] *Id*. (emphasis added).

[10] *Adobe's 'Ethical' Firefly AI Was Trained on Midjourney Images*, CNBC TV18 (Apr. 12, 2024), https://www.cnbctv18.com/technology/adobes-ethical-firefly-ai-was-trained-on-midjourney-images-19396150.htm (on file with counsel).

[11] *Adobe Firefly vs. DALL-E 3*, ADOBE, https://www.adobe.com/products/firefly/discover/firefly-vs-dalle.html (on file with counsel) (last visited May 17, 2026).

[12] Nilay Patel, *Why Adobe CEO Shantanu Narayen is confident we'll all adapt to AI,* THE VERGE (May 13, 2024, 7:00 AM PDT), https://www.theverge.com/24153956/adobe-shantanu-narayen-ai-firefly-premiere-photoshop-pdf-creativity-commerce (on file with counsel).

[13] *Adobe Generative AI Built for Business Solution Brief*, ADOBE (2024), https://www.adobe.com/cc-shared/assets/pdf/trust-center/ungated/whitepapers/corporate/adobe-gen-ai-built-for-business-solution-brief.pdf (on file with counsel).

[14] Matthew Smith, *Adobe Firefly + Adobe Stock*, ADOBE BLOG (Sept. 13, 2023) https://blog.adobe.com/en/publish/2023/09/13/adobe-firefly-adobe-stock (on file with counsel).

32. Adobe has gone as far as creating infrastructure to allow image creators to exclude their works from the training data used for Adobe's Firefly models. This is done through a "do not train" tag affixed to images so that a creator's content will not be used to train AI models.[15] According to Adobe Chief Technology Officer Ely Greenfield, Adobe's AI models already prevent the output of copyrighted content, stating that if Adobe users ask an AI model to generate an image in the style of a particular artist, "it won't generate an image that is aping that person's style . . . If someone wants to use your style, you can actually sell a customer the right to use your style."[16]

33. Adobe acknowledges the legal risk associated with training an AI model with copyrighted content. Adobe's white paper, "Generative AI Built for Business, Adobe's Commitment to Building AI Responsibly," notes that "if an individual or a company uses an image that has been generated by an AI model trained on copyrighted or otherwise legally protected content, that individual or company can be subject to reputational damage and expensive lawsuits for the use of that protected content."[17] Adobe General Counsel Dana Rao further noted:

> When we looked at Firefly or generative models and said, "How are we going to train it? We could train it off the web. We could scrape the web and build that model, or we could try to be more thoughtful about how we train that model given the potential copyright issues and the fact that we have creative customers who are concerned about people training on their content without their permission." . . . Now, the good news about the choice we made on copyright is that it respects our creative customers who are concerned about this. And enterprise customers were very excited to know that they're going to be able to use an image generative model that doesn't have IP issues, doesn't have brand issues, isn't trained on unsafe content, because all of that has been either not in the database at all in the beginning because of the way we trained it or we can use content moderation to address it before it even gets into Firefly.[18]

---

[15] Dawn Chmielewski & Stephen Nellis, *Adobe, Nvidia AI imagery systems aim to resolve copyright questions*, REUTERS (Mar. 21, 2023, at 12:52 PM PDT), https://www.reuters.com/technology/adobe-nvidia-ai-imagery-systems-aim-resolve-copyright-questions-2023-03-21/ (on file with counsel).

[16] *Id.*

[17] *Adobe Generative AI Built for Business Solution Brief*, ADOBE (2024), https://www.adobe.com/cc-shared/assets/pdf/trust-center/ungated/whitepapers/corporate/adobe-gen-ai-built-for-business-solution-brief.pdf (on file with counsel).

[18] Nilay Patel, *How Adobe is managing the AI copyright dilemma, with general counsel Dana Rao*, THE VERGE, (Jan 9, 2024, 7:00 AM PST), https://www.theverge.com/24027198/adobe-dana-rao-ai-copyright-fair-use-figma-acquisition-deal-decoder-interview (on file with counsel).

34. Yet while making these representations, Adobe was simultaneously training its LLMs on *hundreds of thousands* of pirated books without permission, without payment, and without any attempt to obtain licenses from rightsholders. This is far from "showing respect for the creative community." In fact, Adobe did not even attempt to seek *any* licenses for the use of copyrighted books. It simply took the works.

35. In March 2023, Adobe announced a partnership with NVIDIA to "create the next generation of generative AI models" with a focus on transparency.[19]

36. A year later, in March 2024, Adobe again announced it was working together with NVIDIA to "train new NVIDIA LLMs."[20] Specifically, using Adobe's PDF Extract tool, Adobe was "building datasets to train and tune next-generation NVIDIA AI foundation models, including NVIDIA Nemotron LLMs."[21]

37. NVIDIA's Nemotron models are a family of AI models trained on books. When NVIDIA announced the release of the Nemotron-4 15B model in February 2024, it did not identify the specific datasets used to train the model.[22] NVIDIA instead indicated generally that the model was trained with "books."[23] The entire training corpus consisted of 8 trillion tokens. 5.6 trillion (or seventy percent) of those tokens were made up of "English natural language" data, of which 240.8 billion (or 4.6%) were English-language books.[24] To put these figures in perspective, one token is about four characters, or 3/4

---

[19] *NVIDIA Brings Generative AI to World's Enterprises With Cloud Services for Creating Large Language and Visual Models*, NVIDIA (Mar. 21, 2023), https://investor.nvidia.com/news/press-release-details/2023/NVIDIA-Brings-Generative-AI-to-Worlds-Enterprises-With-Cloud-Services-for-Creating-Large-Language-and-Visual-Models/default.aspx (on file with counsel).

[20] Abhigyan Modi, *Adobe Partners with NVIDIA to Harness the Power of PDF Intelligence with Next-Gen LLMs*, ADOBE BLOG (Mar. 18, 2024), https://blog.adobe.com/en/publish/2024/03/18/adobe-partners-nvidia-harness-power-pdf-intelligence-next-gen-llms (on file with counsel).

[21] *Id.*

[22] Jupinder Parmar et al., *Nemotron-4 15B Technical Report*, NVIDIA, Working Paper (Feb. 27, 2024) [https://arxiv.org/pdf/2402.16819]. NVIDIA had not always concealed the datasets it used to train its models. When releasing previous models, NVIDIA disclosed the specific datasets it used to train those models, which revealed NVIDIA's use of the Books3 dataset containing illicit copies of copyrighted books.

[23] *Id.*

[24] *Id.*

of a word, and 100 tokens is about 75 words.[25] An average-length novel of 250 to 300 pages containing 250 to 300 words per page is about 128,000 tokens.[26] A dataset consisting of 240.8 billion tokens of books data, with an average book length of 128,000 tokens, would contain **1,881,250** books.

38. In June 2024, NVIDIA released the Nemotron-4 340B model that used the same 8 trillion tokens from the Nemotron-4 15B model, but was fine-tuned on an additional 1 trillion tokens.[27] Adobe and NVIDIA have never announced any agreement with book authors or publishers to license millions of books to train any AI model. In fact, a licensing deal of that magnitude does not exist *with any AI company.* Upon information and belief, obtaining such a large number of English-language books to train the Nemotron models could only be possible by pirating copyrighted books from sources such as Anna's Archive that already make millions of copyrighted books available for download for free.

39. Adobe and NVIDIA intentionally concealed the source of books data used to train the Nemotron models to prevent exposure of their conduct and the enforcement of copyrights by rightsholders.

40. Adobe itself admitted to training and fine-tuning Nemotron models, and building datasets to train the Nemotron models using its Adobe PDF Extract tool[28] (an AI-powered tool allowing the extraction of content from PDFs into a format suitable for LLM training).[29] This facilitates using the text within PDF files to train or fine-tune LLMs.

---

[25] *What are tokens and how to count them?*, OPENAI HELP CENTER, https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them (on file with counsel) (last visited May 17, 2026).

[26] *Everything You Need to Know About Tokens, Data Volumes and Processing in ChatGPT*, UINTENT (Nov. 26, 2024), https://www.uintent.com/case-studies-und-blog/everything-you-need-to-know-about-token-chatgpt (on file with counsel).

[27] *Nemotron-4 340B Technical Report*, NVIDIA, Working Paper (Aug. 5, 2024) [https://arxiv.org/pdf/2406.11704].

[28] Abhigyan Modi, *Adobe Partners with NVIDIA to Harness the Power of PDF Intelligence with Next-Gen LLMs*, ADOBE BLOG (Mar. 18, 2024), https://blog.adobe.com/en/publish/2024/03/18/adobe-partners-nvidia-harness-power-pdf-intelligence-next-gen-llms (on file with counsel).

[29] *PDF Extract API Overview*, ADOBE DEVELOPER, https://developer.adobe.com/document-services/docs/overview/pdf-extract-api/ (on file with counsel) (last visited May 17, 2026).

41. Pirated books are commonly stored in PDF format. Anna's Archive, for example, makes available for download PDF versions of books, including Plaintiff's works.

42. NVIDIA acquired millions of unauthorized copies of copyrighted books from the shadow library Anna's Archive in order to develop its LLMs. Upon information and belief, Adobe used these millions of pirated books to train and fine-tune the Nemotron models, including through the use of its PDF Extract tool.

43. Shadow libraries are illicit online repositories of copyrighted content that openly defy copyright laws. Typically, they allow anyone to search for and download pirated books, including in PDF format. In spite of enforcement actions to curtail the ability of shadow libraries to distribute large collections of books, several shadow libraries have flourished and gained notoriety. These include (1) Library Genesis ("LibGen"), a website repeatedly enjoined by federal courts for copyright infringement in default proceedings and which has been designated a "notorious" repository of pirated works by the United States Trade Representative; its founders have faced federal criminal charges, (2) Z-Library (aka B-ok), a shadow library that originally began as a for-profit LibGen mirror charging a fee for expedited downloading, and which was seized by law enforcement; its founders have been arrested and indicted with criminal copyright infringement (and have since fled the country), and (3) Sci-hub, another shadow library operating in default of multiple federal court judgements for copyright infringement.

44. The most active and notorious shadow library is "Anna's Archive," which first began as "Pirate Library Mirror" (named this way because it "mirrored" all of the books from Z-Library). The shadow library rebranded to "Anna's Archive" in 2022 and soon made available *all* of the content of other shadow libraries, including LibGen, Z-Library, Sci-Hub, and Internet Archive, and additional books sourced from pirated libraries. Anna's Archive makes *millions* of pirated books available for download to the public. Books can be downloaded individually, or entire datasets with hundreds of thousands or millions of books can be downloaded wholesale through the peer-to-peer BitTorrent Protocol (also known as "torrenting"). When someone downloads a file through BitTorrent, they must also "seed" the file (that is, distribute it) to others who are also downloading the file. As a result, users of the BitTorrent protocol are engaged in both piracy *and* the distribution of illegal content.

45. The use of pirated books from Anna's Archive, LibGen, Z-Library, and other shadow libraries has been commonplace within the AI development community. [30] Major AI companies, including OpenAI, Meta, Anthropic, and NVIDIA pirated books from Anna's Archive, Library Genesis, Z-Library, Sci-Hub, and/or Pirate Library Mirror.

46. Shadow libraries provide an easily accessible source of high-quality copyrighted material—for free. As Anna's Archive explained, "[i]t is well understood that LLMs thrive on high-quality data. We have the largest collection of books, papers, magazines, etc. in the world, which are some of the highest quality text sources."[31] Shadow libraries themselves acknowledge that the explosion in piracy and patronage by LLM companies has saved shadow libraries from extinction. As the creators of Anna's Archive wrote:

> Not too long ago, "shadow-libraries" were dying. Sci-Hub, the massive illegal archive of academic papers, had stopped taking in new works, due to lawsuits. "Z-Library", the largest illegal library of books, saw its alleged creators arrested on criminal copyright charges . . . . ***Then came AI. Virtually 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. Most are Chinese, though we've also worked with companies from the US***.[32]

47. Adobe's disregard for the rights of copyright holders did not stop there. In August 2025, Adobe released its SlimLM series of "Small Language Models" ("SLMs"), a type of LLM.[33] These models were trained on the "SlimPajama" dataset that contains the Books3 dataset.

---

[30] *See, e.g.*, Alex Reisner, *The Unbelievable Scale of AI's Pirated-Books Problem*, THE ATLANTIC (Mar. 20, 2025), https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/ (on file with counsel); *Bartz v. Anthropic PBC*, 787 F. Supp. 3d 1007, 1015 (N.D. Cal. 2025) (noting Anthropic's use of LibGen and Pirate Library Mirror to download millions of copyrighted books).

[31] *LLM Data*, ANNA'S ARCHIVE, https://annas-archive.pk/llm (on file with counsel) (last visited May 17, 2026).

[32] *Copyright reform is necessary for national security*, ANNA'S ARCHIVE BLOG (Jan. 31, 2025) https://annas-archive.pk/blog/ai-copyright.html (on file with counsel) (emphasis added).

[33] Thang M. Pham et al., *SlimLM: An Efficient Small Language Model for On-Device Document Assistance*, Working Paper (Nov. 25, 2024) [https://arxiv.org/pdf/2411.09944].

48. The SlimPajama dataset was created by AI development company Cerebras AI in June 2023 as a smaller version of the popular dataset known as RedPajama.[34]

49. RedPajama (also known as RedPajama-V1) is a dataset created by Together AI in April 2023 for use in training AI models. The dataset aimed to "create a set of leading open-source models and to rigorously understand the ingredients that yield good performance."[35] In fact, the RedPajama dataset specifically replicated the dataset mixture used by Meta to train the then-state-of-the-art LLaMA series of models, noting that "for each data slice, [Together AI] conduct[ed] careful data pre-processing and filtering. . . to roughly match the number of tokens as reported by Meta AI in the LLaMA paper."[36]

50. One of the subsets of RedPajama is the "Books" dataset, described by Together AI as "[a] corpus of open books." Since RedPajama mirrored the dataset used to create the LLaMA models, more detail about the RedPajama-Books subset is found in the paper "LLaMA: Open and Efficient Foundation Language Models:"

> We include two book corpora in our training dataset: the Gutenberg Project, which contains books that are in the public domain, and the Books3 section of The Pile (Gao et al., 2020), a publicly available dataset for training large language models.[37]

[34] Daria Soboleva, *SlimPajama: A 627B token, cleaned and deduplicated version of RedPajama*, CEREBRAS BLOG (June 9, 2023), https://www.cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama (on file with counsel).

[35] *Releasing 3B and 7B RedPajama-INCITE family of models including base, instruction-tuned & chat models*, TOGETHERAI BLOG (May 5, 2023), https://www.together.ai/blog/redpajama-models-v1 (on file with counsel).

[36] *Id.*; *RedPajama, a project to create leading open-source models, starts by reproducing LLaMA training dataset of over 1.2 trillion tokens*, TOGETHERAI BLOG (Apr. 17, 2023), https://www.together.ai/blog/redpajama (on file with counsel).

[37] Hugo Touvron et al, *LLaMA: Open and Efficient Foundation Language Models*, Meta AI Working Paper (Feb. 27, 2023) [https://arxiv.org/pdf/2302.13971] (on file with counsel).

51. The Books3 dataset was created by Shawn Presser in October 2020. Books3 is included within the "SlimPajama-Books" subset used to train Adobe models. The paper, "The Pile: An 800GB Dataset of Diverse Text for Language Modeling,"[38] details the contents of the Books3 dataset:

> Books3 is a dataset of books derived from a copy of the contents of the Bibliotik private tracker . . . Bibliotik consists of a mix of fiction and nonfiction books and is almost an order of magnitude larger than our next largest book dataset (BookCorpus2). We included Bibliotik because books are invaluable for long-range context modeling research and coherent storytelling.[39]

52. Bibliotik is a notorious shadow library similar to the other shadow libraries described above, such as Library Genesis (aka LibGen), Z-Library (aka B-ok), Sci-Hub, and Anna's Archive. Mr. Presser has confirmed in public statements that Books3 represents "all of Bibliotik" and contains approximately 196,640 books.

53. The RedPajama dataset is available for download from Hugging Face. Before October 2023, the Books subset containing Books3 was "downloaded from Huggingface [sic]" when a user ran the script that automatically assembled the RedPajama dataset.[40] But in October 2023, the Books3 dataset was removed with a message that it "is defunct and no longer accessible due to reported copyright infringement."[41] After the "Books3 dataset" was removed from Hugging Face, the RedPajama dataset documentation also added a message that Books3 is defunct "due to reported

---

[38] Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling*, EleutherAI Working Paper (Dec. 31, 2020) [https://arxiv.org/pdf/2101.00027.pdf].

[39] *Id.* at 3–4.

[40] *RedPajama-Data-1T*, Hugging Face Datasets, https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T [https://web.archive.org/web/20230420075601/https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T].

[41] *the_pile_books3*, Hugging Face Datasets, https://huggingface.co/datasets/the_pile_books3 [https://web.archive.org/web/20231127101818/https://huggingface.co/datasets/the_pile_books3].

copyright infringement."[42] Together AI has also acknowledged it took down Books3 "due to copyright issues."[43]

54. The SlimPajama dataset was previously available for download on HuggingFace,[44] but was removed entirely sometime in 2026. The SlimPajama dataset downloaded and used by Adobe contained the SlimPajama-Books dataset that included Books3.

55. Plaintiff's copyrighted books listed in Exhibit A are among the works available at Anna's Archive and included in the Books3 dataset. Below, these books are referred to as the Asserted Works.

56. Adobe benefitted commercially from its use of pirated books by incorporating Nemotron and SlimLM models into its programs and services, thereby attracting more customers and revenue. For example, in March 2026 Adobe and NVIDIA announced a collaboration in which "Adobe will bring NVIDIA Nemotron capabilities to Adobe Acrobat to further elevate the quality of AI output and increase productivity for business professionals, consumers and enterprises."[45] Adobe Acrobat is the widely used program to view and manipulate PDFs. The announcement also noted that Nemotron models would power Adobe's "agentic workflows."[46] And in April 2026 Adobe announced it was embedding "NVIDIA Nemotron open models into the CX Enterprise Coworker platform."[47]

---

[42] *RedPajama-Data-1T*, Hugging Face Datasets, https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T [https://web.archive.org/web/20240510231649/https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T].

[43] Maurice Weber et al., *RedPajama: an Open Dataset for Training Large Language Models*, Working Paper (Nov. 19, 2024) [https://arxiv.org/pdf/2411.12372].

[44] *SlimPajama-627B*, Hugging Face Datasets, https://huggingface.co/datasets/cerebras/SlimPajama-627B [https://web.archive.org/web/20230609191207/https://huggingface.co/datasets/cerebras/SlimPajama-627B].

[45] *Adobe and NVIDIA Announce Strategic Partnership to Deliver the Next Generation of Firefly Models and Creative, Marketing and Agentic Workflows*, ADOBE NEWSROOM (Mar. 16, 2026, at 1:30 PM), https://news.adobe.com/news/2026/03/adobe-and-nvidia-announce-strategic-partnership (on file with counsel).

[46] *Id.*

[47] Dom Nicastro, *Nvidia CEO Jensen Huang Tells the SaaS World: Agentic Is Here. Adobe Is Listening*, CMSWIRE (Apr. 20, 2026), https://www.cmswire.com/digital-experience/nvidia-ceo-jensen-huang-told-the-saas-world-agentic-is-here-adobe-was-listening/ (on file with counsel).

57. Adobe infringed on Plaintiff and Class members' copyrighted works on a massive scale. Adobe acquired books originating from the Books3 subset of SlimPajama without authorization from or compensation to their authors. Upon information and belief, Adobe also acquired books originating from Anna's Archive without authorization from or compensation to their authors. Adobe then continued copying and storing the datasets, and used them to develop and train the Nemotron and SlimLM series of LLMs.

58. Upon information and belief, Adobe used Plaintiff and Class members' copyrighted works to train other models, including non-public models, and to develop other tools, programs, or services.

59. In a speech at Adobe Summit 2023, Adobe CEO Shantanu Narayen stated that Adobe has "been very transparent about the training data" Adobe used, while "a lot of the other companies are actually using data that potentially—they're scraping the internet, they're accessing data that they may or may not have rights and license to."[48] Rather than focusing on the conduct of other companies, however, Mr. Narayen accurately described Adobe's own conduct: Adobe accessed datasets of hundreds of thousands of books for which it held no rights or license to copy.

60. Mr. Narayen also championed the right of creators to exclude their works from being used to train AI models, stating, "We think that that's important because people are going to say, and our creative community is going to say, 'I want do not track. So I don't want to be part of that training data.' So we created that entire infrastructure to allow people to do that." [49] Adobe, however, never implemented any infrastructure to allow *book authors* to exclude their works from the training data of Adobe's AI models.

[48] *ADBE.OQ – Adobe. Inc. Presents at Adobe Summit 2023*, Edited Transcript (Mar. 21, 2023), https://www.adobe.com/cc-shared/assets/investor-relations/pdfs/adbe-summit2023-qna.pdf (on file with counsel).

[49] *Id*.

61. Copyright law does not envision a two-tiered system whereby the copyrights of some creators (e.g., image artists) are afforded more protection than those of others (e.g., authors). But that is precisely what Adobe has done. Adobe goes to great lengths to protect the copyright of image creators: it licenses images for use in its Firefly AI models, compensates creators on a yearly basis, prevents Firefly models from emitting content in the style of those creators, and allows creators to exclude their works from AI training. In contrast, *none* of these protections are afforded to book authors: Adobe failed to secure licenses from copyright holders of books whose works originate from Books3 or Anna's Archive, it provides no royalties to authors for using their books to train the Nemotron and SlimLM models, has provided no ability for authors to opt-out of training, and it has created LLMs capable of memorizing and regurgitating substantial portions of books in the training data, all without compensating the rightsholders. Adobe has failed to follow its own AI Ethics Governance commitments to respect the rights of the creative community. It has failed to responsibly address licensing and copyright issues in its training data. In all, it has shown a blatant disregard for the rights of book authors.

CLASS ALLEGATIONS

62. The "**Class Period**" as defined in this complaint begins on at least May 18, 2023 and runs through the present. Because Plaintiff does not yet know when the unlawful conduct alleged herein began, but believe, on information and belief, that the conduct likely began earlier than May 18, 2023, Plaintiff reserves the right to amend the Class Period to comport with the facts and evidence uncovered during further investigation or through discovery.

63. **Class definition**. Plaintiff brings this action for damages and injunctive relief as a class action under Federal Rules of Civil Procedure 23(a), 23(b)(2), and 23(b)(3), on behalf of the following Classes:

> **All legal or beneficial owners of copyrighted works registered with the United States Copyright Office that Adobe acquired, copied, stored, or otherwise used to develop any large language model, program, or tool during the Class Period.**
>
> ***Anna's Archive subclass*. All legal or beneficial owners of copyrighted works registered with the United States Copyright Office originating from Anna's Archive that Adobe acquired, copied, stored, or**

otherwise used to develop any large language model, program, or tool during the Class Period.

***Books3 subclass*. All legal or beneficial owners of copyrighted works registered with the United States Copyright Office originating from Books3 that Adobe acquired, copied, stored, or otherwise used to develop any large language model, program, or tool during the Class Period**

64. This Class definition excludes:

a. Defendant named herein;

b. any of the Defendant's co-conspirators;

c. any of Defendant's parent companies, subsidiaries, and affiliates;

d. any of Defendant's officers, directors, management, employees, subsidiaries, affiliates, or agents;

e. all governmental entities; and

f. the judges and chambers staff in this case, as well as any members of their immediate families.

65. **Numerosity**. Plaintiff does not know the exact number of members in the Class. This information is in the exclusive control of Defendant. On information and belief, there are at least thousands of members in the Class geographically dispersed throughout the United States. Therefore, joinder of all members of the Class in the prosecution of this action is impracticable.

66. **Typicality**. Plaintiff's claims are typical of the claims of other members of the Class because Plaintiff and all members of the Class were damaged by the same wrongful conduct of Defendant as alleged herein, and the relief sought herein is common to all members of the Class.

67. **Adequacy**. Plaintiff will fairly and adequately represent the interests of the members of the Class because the Plaintiff has experienced the same harms as the members of the Class and has no conflicts with any other members of the Class. Furthermore, Plaintiff has retained sophisticated and competent counsel who are experienced in prosecuting federal and state class actions, as well as other complex litigation.

68. **Commonality** and **predominance**. Numerous questions of law or fact common to each Class member arise from Defendant's conduct and predominate over any questions affecting the members of the Class individually:

a. Whether Defendant violated the copyrights of Plaintiff and the Class when it obtained copies of Plaintiff's Asserted Works, stored the Asserted Works, copied the Asserted Works, and used the Asserted Works to develop LLMs and tools.

b. Whether any affirmative defense excuses Defendant's conduct.

c. Whether any statutes of limitation limits the potential for recovery for Plaintiff and the Class.

69. **Other class considerations**. Defendant has acted on grounds generally applicable to the Class. This class action is superior to alternatives, if any, for the fair and efficient adjudication of this controversy. Prosecuting the claims pleaded herein as a class action will eliminate the possibility of repetitive litigation. There will be no material difficulty in the management of this action as a class action. The prosecution of separate actions by individual Class members would create the risk of inconsistent or varying adjudications, establishing incompatible standards of conduct for Defendant.

CAUSES OF ACTION

COUNT I

Direct Copyright Infringement (17 U.S.C. § 501) against Defendant

70. Plaintiff incorporates by reference the preceding factual allegations.

71. As the owner of the registered copyrights in the Asserted Works, Plaintiff holds the exclusive rights to those books under 17 U.S.C. § 106.

72. Adobe downloaded, ingested, acquired, or stored copies of books originating from the Books3 dataset, which includes the Asserted Works.

73. Upon information and belief, Adobe acquired, ingested, or stored copies of books originating from Anna's Archive, a shadow library, which include the Asserted Works.

74. To develop its SlimLM language models, Adobe copied datasets containing Books3 and incorporated them into the training dataset for its LLMs. Adobe made multiple copies of these datasets during the development of the SlimLM series of LLMs.

75. Upon information and belief, to develop the Nemotron language models, Adobe copied datasets originating from Anna's Archive to develop these models and incorporate the datasets into the models' training data. Adobe made multiple copies of these datasets during the development of the Nemotron models.

76. Plaintiff and the Class members never authorized Defendant to make copies of their Asserted Works, make derivative works, publicly display copies (or derivative works), or distribute copies (or derivative works). All those rights belong exclusively to Plaintiff and the Class members under the U.S. Copyright Act.

77. By copying, storing, processing, reproducing, and using the datasets containing copies of Plaintiff's original works, Defendant has directly infringed Plaintiff's exclusive rights in her copyrighted works, in violation of the Copyright Act.

78. By copying, storing, processing, and reproducing the Nemotron and SlimLM models—each trained on Plaintiff's copyrighted works—Adobe has further directly infringed upon Plaintiff's exclusive rights in her copyrighted works, in violation of the Copyright Act.

79. Defendant repeatedly copied, stored, and used the Asserted Works without her authorization or consent. Such copies were made in direct violation of her exclusive rights under the Copyright Act.

80. Defendant's infringing conduct, as alleged herein, was and continues to be willful and carried out with full knowledge of Plaintiff's rights in and to the Asserted Works. As a direct and proximate result of its conduct, Defendant has wrongfully profited from copyrighted works to which it holds no ownership interest.

81. By and through the actions alleged above, Defendant has infringed and will continue to infringe Plaintiff's copyrights.

82. Plaintiff has been injured by Defendant's acts of direct copyright infringement. Plaintiff is entitled to statutory damages, actual damages, restitution of profits, and all appropriate legal and equitable relief.

DEMAND FOR JUDGMENT

Wherefore, Plaintiff requests that the Court enter judgment on their behalf and on behalf of the Class defined herein, by ordering:

a) This action may proceed as a class action, with Plaintiff serving as Class Representative, and with Plaintiff's counsel as Class Counsel.

b) Judgment in favor of Plaintiff and the Class and against Defendant.

c) An award of statutory and other damages under 17 U.S.C. § 504 for Defendant's violations of the copyrights of Plaintiff and the Class.

d) An award of reasonable attorneys' fees and reimbursement of all costs of this action pursuant to 17 U.S.C. § 505, or as otherwise permitted by law.

e) A declaration that such infringement is willful.

f) An order directing the destruction or other reasonable disposition of all copies Defendant made or used in violation of the exclusive rights of Plaintiff and the Class, pursuant to 17 U.S.C. § 503(b).

g) Injunctive relief sufficient to alleviate and stop Defendant's unlawful conduct alleged herein.

h) Pre- and post-judgment interest on the damages awarded to Plaintiff and the Class, and that such interest be awarded at the highest legal rate from and after the date this complaint is first served on Defendant.

i) Defendant is responsible financially for the costs and expenses of a Court-approved notice program through post and media designed to give immediate notification to the Class.

j) Further relief for Plaintiff and the Class as may be appropriate.

//

//

//

JURY TRIAL DEMANDED

Under Federal Rule of Civil Procedure 38(b), Plaintiff demands a trial by jury of all the claims asserted in this complaint so triable.

Dated: May 18, 2026

By: */s/ Joseph R. Saveri*

Joseph R. Saveri (SBN 130064)
Diane S. Rice (SBN 118303) Christopher K.L. Young (SBN 318371) William W. Castillo Guardado (SBN 294159) **SAVERI LAW FIRM, LLP**
550 California Street, Suite 910
San Francisco, CA 94104
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
drice@saverilawfirm.com
cyoung@saverilawfirm.com
wcastillo@saverilawfirm.com

Attorneys for Individual and Representative Plaintiff E. Molly Tanzer