[Counsel on signature page]

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
OAKLAND DIVISION

ABDI NAZEMIAN, et al, Individual and Representative Plaintiffs, vs. NVIDIA CORPORATION Defendant.	Master File Case No. 4:24-cv-01454-JST (SK) Consolidated with Case No. 4:24-cv-02655-JST (SK) **SECOND CONSOLIDATED AMENDED COMPLAINT** **Class Action** **Demand for Jury Trial**

Plaintiffs Abdi Nazemian, Brian Keene, Stewart O'Nan, Andre Dubus III, and Susan Orlean (together "Plaintiffs"), on behalf of themselves and all others similarly situated, bring this class action complaint ("Complaint") against Defendant NVIDIA Corporation ("NVIDIA" or "Defendant").

OVERVIEW

1. *Artificial intelligence*—commonly abbreviated "AI"—denotes software that is designed to algorithmically simulate human reasoning or inference, often using statistical methods.

2. A *large language model* is an AI software program designed to emit convincingly naturalistic text outputs in response to user prompts.

3. Rather than being programmed in the traditional way—that is, by human programmers writing code—a large language model is *trained* by copying an enormous quantity of textual works, extracting protected expression from these works, and transforming that protected expression into a large set of numbers called *weights* that are stored within the model. These weights are entirely and uniquely derived from the protected expression in the training dataset. Whenever a large language model 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.

4. Plaintiffs and Class members are authors. They own registered copyrights in certain books that NVIDIA has admitted copying, storing, and using to develop its AI language models.

5. NVIDIA copied these copyrighted works multiple times to train its language models, including from known pirated libraries (also known as "shadow libraries"). Those notorious shadow libraries include The Pile, Bibliotik, and Anna's Archive.

> I am on the data strategy team at NVIDIA, we are exploring including Anna's Archive in pre-training data for our LLMs.
> We are figuring out internally whether we are willing to accept the risk of using this data, but would like to speak with your team to get a better understanding of LLM-related work you have done.

6. NVIDIA "got the green light" to use Anna's Archive. NVIDIA did not hesitate in using pirated books from these illicit sources of copyrighted material, regardless of the "risk" or the harm to authors like the Plaintiffs.

7. And NVIDIA also caused numerous third parties to download and store Plaintiffs' copyrighted works by encouraging, facilitating, and promoting its customers to download copies of The Pile dataset, which includes more than one hundred thousand copyrighted books.

JURISDICTION AND VENUE

8. This Court has subject-matter jurisdiction under 28 U.S.C. § 1331 because this case arises under the Copyright Act (17 U.S.C. § 501).

9. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. § 1391(c)(2) because NVIDIA is headquartered in this district. NVIDIA created various large language models, including the NeMo Megatron models, and distributes them commercially. 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.

10. 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.

PLAINTIFFS

11. Plaintiff Abdi Nazemian is an author who lives in California. Mr. Nazemian owns registered copyrights in multiple books, including *Like a Love Story*.

12. Plaintiff Brian Keene is an author who lives in Pennsylvania. Mr. Keene owns registered copyrights in multiple books, including *Ghost Walk*.

13. Plaintiff Stewart O'Nan is an author who lives in Pennsylvania. Mr. O'Nan owns registered copyrights in multiple books, including *Last Night at the Lobster*.

14. Plaintiff Andre Dubus III is an author who lives in Massachusetts. Plaintiff Dubus owns registered copyrights in multiple books, including, *The Garden of Last Days*, *The Cage Keeper*, and *Townie: A Memoir*.

15. Plaintiff Susan Orlean is an author who lives in California. Plaintiff Orlean owns registered copyrights in multiple works, including, *The Orchid Thief* and *The Library Book*.

16. A non-exhaustive list of registered copyrights owned by Plaintiffs is included as Exhibit A.

DEFENDANT

17. Defendant NVIDIA is a Delaware corporation with its principal place of business at 2788 San Tomas Expressway, Santa Clara CA 95051.

AGENTS AND CO-CONSPIRATORS

18. The unlawful acts alleged against the Defendant in this class action 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.

19. 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, or for Defendant with respect to the acts, violations, and common course of conduct alleged herein.

FACTUAL ALLEGATIONS

20. NVIDIA is a diversified technology company founded in 1993 that originally focused on computer-graphics hardware, e.g., Graphics Processing Units ("GPUs"), and has since expanded to other computationally intensive fields, including software such as NVIDIA's "Compute Unified Device Architecture" and hardware, e.g. NVLink/NVLink Switch, for training and operating AI software programs. NVIDIA's hardware and software is used by all Frontier AI companies—companies that develop the most advanced AI systems— which has resulted in NVIDIA becoming the world's most valuable company.

21. In addition to the hardware and software products it sells to AI companies, NVIDIA itself has developed numerous AI models known as "large language models" ("LLMs"). An LLM is AI software designed to emit convincingly naturalistic text outputs in response to user prompts. NVIDIA sells products to its customers that rely on NVIDIA's LLMs.

22. Though LLMs are software programs, they are not created the way most software programs are—that is, by human software programmers writing code. Rather, LLMs are *trained* by copying an enormous quantity of textual works and then feeding these copies in pieces into the model. This corpus of input material is called the *training dataset*.

23. As set forth below, NVIDIA unlawfully copied copyrighted material from illegal pirate "shadow libraries." NVIDIA collated and stored this material in centralized servers which its engineers (and other employees) could access for any purpose. NVIDIA and its employees subsequently made additional unlawful copies of this illegally-obtained copyrighted material during the LLM development process.

24. During the training process, LLMs copy and ingest each textual work in the training dataset and extract protected expression from it. In a process somewhat resembling a guess-and-check quiz, the LLM is progressively adjusted to more closely approximate the protected expression copied from the training dataset. The LLM records the results of this process in a large set of numbers called *weights* or *parameters* that are stored within the model, and, in some sense, "are" the model. These weights are entirely and uniquely derived from the protected expression in the training dataset.

For instance, the NeMo Megatron–GPT 20B model—an LLM released in September 2022 as part of NVIDIA's NeMo Megatron series of LLMs—is so named because the model stores 20 billion ("20B") weights derived from protected expression in its training dataset.

25. Importantly, datasets may have multiple uses during the development process of an LLM even if the dataset does not become part of a model's final training dataset. For example, during the development of an LLM, the developer may initiate a *run* or *checkpoint* using certain datasets to see the effect of that dataset on the model. Once the checkpoint is finished, a full model is completed and its performance analyzed. The developer may then alter the datasets and conduct another checkpoint. This process may occur multiple times before a developer arrives at the final checkpoint for that model. All of the models created as part of the checkpoint process may never receive official names nor be publicly released.

26. Once the LLM has copied and ingested the textual works in the training dataset and transformed the protected expression into stored weights, the LLM 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.

27. Much of the material in NVIDIA's training dataset, however, comes from copyrighted works—including books written by Plaintiffs and Class members—that were acquired, copied and stored by NVIDIA without consent, without credit, and without compensation.

28. In November 2021, NVIDIA announced the "NeMo Megatron framework for training language models."[1] NVIDIA touted this framework as "provid[ing] a production-ready, enterprise-grade solution to simplify the development and deployment of large language models."[2]

29. In September 2022, NVIDIA announced the availability of the NeMo Megatron language models in a video on its website: "For the first time, NVIDIA is making its checkpoints available publicly, where the checkpoints are trained with NeMo Megatron … this is just to begin

[1] *See* https://nvidianews.nvidia.com/news/nvidia-brings-large-language-ai-models-to-enterprises-worldwide.
[2] *Id.*

with. And this is not the end. We will continue to add more checkpoints in the future."[3] In this context "checkpoints" is an alternate term for language models. The language models released in September 2022 include NeMo Megatron-GPT 1.3B, NeMo Megatron-GPT 5B, NeMo Megatron-GPT 20B, and NeMo Megatron-T5 3B models.

30. Each of these NeMo Megatron models was hosted on a website called Hugging Face, where a model card provides information about each model, including its training dataset. The model card for each of the NeMo Megatron models states, "The model was trained on 'The Pile' dataset prepared by EleutherAI."[4]

31. The Pile is a training dataset curated by a research organization called EleutherAI. In December 2020, EleutherAI introduced this dataset in a paper called "The Pile: An 800GB Dataset of Diverse Text for Language Modeling"[5] (the "EleutherAI Paper").

32. According to the EleutherAI Paper, one of the components of The Pile is a collection of books called Books3. The EleutherAI Paper reveals that the Books3 dataset comprises 108 gigabytes of data, or approximately 12% of the dataset, making it the third largest component of The Pile by size.

33. The EleutherAI Paper further describes the contents of Books3:

> 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.**[6]

---

[3] *See* https://www.nvidia.com/en-us/on-demand/session/gtcfall22-a41200/?nvid=nv-int-tblg-881125, starting at 37:25.

[4] *See, e.g.*, https://huggingface.co/nvidia/nemo-megatron-gpt-1.3B#training-data, https://huggingface.co/nvidia/nemo-megatron-gpt-5B#training-data, https://huggingface.co/nvidia/nemo-megatron-gpt-20B#training-data, https://huggingface.co/nvidia/nemo-megatron-t5-3B#training-data

[5] Available at https://arxiv.org/pdf/2101.00027.pdf

[6] *Id.* at 3–4 (emphasis added).

34. Bibliotik is one of a number of notorious "shadow library" websites which make, store, and distribute huge quantities of pirated copyrighted works via the BitTorrent Protocol.

35. The person who assembled the Books3 dataset, Shawn Presser, has confirmed in public statements that it represents "all of Bibliotik" and contains approximately 196,640 books.

36. Plaintiffs' copyrighted books listed in Exhibit A are among the works in the Books3 dataset. Below, these books are referred to as the **Infringed Works**.

37. Until October 2023, the Books3 dataset was available from Hugging Face. At that time, the Books3 dataset was removed with a message that it "is defunct and no longer accessible due to reported copyright infringement."[7]

38. NVIDIA has publicly admitted training its NeMo Megatron models on a copy of The Pile dataset. Therefore, NVIDIA necessarily also (1) acquired a copy of Books3 (because it is part of The Pile) and (2) made additional copies of Books3 during the course of developing LLMs, including (but not limited to) its NeMo Megatron models. Certain books written by Plaintiffs are part of Books3—including the Infringed Works—and thus NVIDIA necessarily (1) made unlawful copies of Plaintiffs' works when downloading Books3, and (2) made additional unlawful copies of Plaintiffs' works when developing its LLMs, including (but not limited to) its NeMo Megatron models. NVIDIA thus directly infringed Plaintiffs' copyrights.

39. But NVIDIA's use of Plaintiffs' Infringed Works was not limited to the models it *publicly* disclosed were trained on The Pile. NVIDIA and its engineers maintained The Pile in centralized servers and repeatedly (and extensively) used The Pile following its acquisition, including to develop multiple LLMs known internally as NeMo Megatron GPT 126M, NeMo Megatron GPT 40B, NeMo Megatron GPT 175B, NeMo Megatron T5 220M, NeMo Megatron T5 11B, and NeMo Megatron T5 23B.

40. NVIDIA's use of The Pile to develop language models was not limited to a single line or class of models either. Instead, language models across NVIDIA used The Pile.

[7] *See* https://huggingface.co/datasets/the_pile_books3

41. NVIDIA used The Pile to train and develop models that do not bear the NeMo Megatron name as well. For instance, NVIDIA included the Pile dataset as training data for an LLM known as Megatron 345M, which was publicly released as the Megatron GPT2 345m model. NVIDIA also used The Pile to train an LLM known as "NeMo GPT-3 10B." NVIDIA additionally developed the InstructRetro-48B and Retro-48B LLMs using the Books3 dataset from The Pile.

42. The Pile was not NVIDIA's only dataset that included Books3. NVIDIA also downloaded the SlimPajama dataset.[8] "SlimPajama was created by cleaning and deduplicating the 1.2T token RedPajama dataset from [the company] Together [AI]." And the RedPajama dataset itself originally included the Books3 dataset. The SlimPajama dataset included the Books3 dataset. NVIDIA used the SlimPajama dataset to test "both sentencepiece and BPE [tokenizers]." Tokenizers are software which is used to process training data for use in LLM training and development. In short, NVIDIA used the SlimPajama dataset to develop and test the software used in the development of its LLMs. As one NVIDIA employee remarked, "SlimPajama . . . is available in our org." NVIDIA, therefore, again infringed Plaintiffs' copyrights by downloading unauthorized copies of their works by downloading, storing, and using the SlimPajama dataset.

43. Upon information and belief, NVIDIA also developed a large number of internal models, including checkpoints, many of which were never given proper names or publications but which also unlawfully included datasets containing Plaintiffs' and Class members' works, such as The Pile.

44. Upon information and belief, NVIDIA also made unlawful copies of The Pile during the course of internal research which did not result in a fully trained LLM.

45. Not content to acquire, store, and use The Pile in its internal and external LLM research, development, and commercialization efforts, NVIDIA sought vastly more copyrighted works than The Pile could provide. Because the quality of an LLM depends on both the quality *and* quantity of its training data, NVIDIA found itself desperate for additional books. Books have the

[8] *See* https://huggingface.co/datasets/cerebras/SlimPajama-627B.

unique designation of being widely understood as high-quality LLM training data and being available illegally in large quantities from illicit shadow libraries.

46. In addition to Bibliotik (the source of Books3, discussed above), those shadow libraries include: (1) Library Genesis ("LibGen") which has been 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; (2) Z-Library (aka B-ok) which began as a for-profit LibGen mirror which enabled expedited downloads for a fee until it was seized by law enforcement as part of an operation which resulted in its founders being arrested and indicted (they have since fled the country); and (3) Sci-hub which, like LibGen, has been repeatedly enjoined by federal courts for copyright infringement in default proceedings.

47. The most active current shadow library is known as "Anna's Archive." The successor to Z-library, Anna's Archive began existence as "Pirate Library Mirror," a name derived from the fact that it "mirrored" (that is to say, hosted all the same books as) Z-Library. Shortly after its launch in 2022, it rebranded to "Anna's Archive" and quickly expanded to host all of LibGen, Z-Library, Sci-Hub, and additional books sourced from pirated libraries. Anna's Archive hosts millions of pirated books.

48. Many of these shadow libraries enable increased download speeds or quantities for paying members. *See, e.g.,* https://annas-archive.org/donate.

49. These "shadow libraries" have long been of interest to the AI industry—and their insatiable quest for more data—because they illegally host and distribute vast quantities of high-quality copyrighted material and because they are willing to move LLM developers to the "front of the line" for download speeds—in exchange for a fee.

50. 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." https://annas-archive.org/llm. Shadow libraries provide "high-speed . . . enterprise-level access [to their collections] . . . [in exchange] for donations in the range of tens of thousands USD." In other words: paid piracy.

51. As revealed publicly over the last year,[9] it is an industry-wide practice to use shadow libraries such as Library Genesis, Z-Library, and Pirate Library Mirror. Virtually every one of the major LLM developers—including OpenAI, Meta, and Anthropic—pirated books from Library Genesis, Z-Library, Sci-Hub, and/or Pirate Library Mirror. NVIDIA followed this industry-wide practice and pirated troves of books from shadow libraries.

52. The shadow libraries themselves have noted that the explosion in piracy and patronage by LLM companies has saved shadow libraries from extinction. As a post by the admins of Anna's Archive put it:

> 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. . . We have given high-speed access to about 30 companies***. https://annas-archive.org/blog/ai-copyright.html (emphasis added_)

53. Internal documents show competitive pressures drove NVIDIA to piracy. In the fall of 2023, NVIDIA faced a rapidly approaching deadline in the form of its annual developer day. In the year since the launch of the NeMo Megatron series in September 2022, OpenAI had released ChatGPT to massive success, resulting in a substantial increase in investor attention on AI. In response, NVIDIA sought to develop and demonstrate cutting edge LLMs at its fall 2023 developer day. In seeking to acquire data for what it internally called "NextLargeLLM," "NextLLMLarge" and "Next Generation LLM" (collectively, "NextLargeLLM"). NVIDIA was"[h]yper [f]ocused on books corpuses." NVIDIA knew that "published books under copyright" are "the most valuable" for developing LLMs and NVIDIA knew that only books were available in sufficient quantities. And

[9] *See, e.g.*, Alex Reisner, *The Unbelievable Scale of AI's Pirated-Books Problem*, The Atlantic (March 20, 2025), https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/; *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 pf copyrighted books).

NVIDIA needed to achieve 8 trillion tokens for the "NextLargeLLM," and books provided this means.

54. In August 2023, NVIDIA contacted books publishers to obtain fast "access to large volumes of unique, high-quality datasets" or "ie. books." But on information and belief, NVIDIA could not secure this fast access to the huge quantity of books it needed through publishers. As one book publisher told NVIDIA, it was " not in a position to engage directly just yet but will be in touch." In 2023, NVIDIA had "chatted with multiple publishers . . . but none [] wanted to enter into data licensing deals."

55. Desperate for books, NVIDIA contacted Anna's Archive—the largest and most brazen of the remaining shadow libraries—about acquiring its millions of pirated materials and "including Anna's Archive in pre-training data for our LLMs." Because Anna's Archive charged tens of thousands of dollars for "high-speed access" to its pirated collections, *see* https://annas-archive.org/llm, NVIDIA sought to find out what "high-speed access" to the data would look like.

56. In correspondence with NVIDIA executives, Anna's Archive stated that, because its collections were illegally acquired and maintained, NVIDIA executives would need to "let [Anna's Archive] know when you have decided internally that this is something you can pursue. We have wasted too much time on people who could not get internal buy-in."

57. Within a week of contacting Anna's Archive, and days after being warned by Anna's Archive of the illegal nature of their collections, NVIDIA management gave "the green light" to proceed with the piracy. Anna's Archive offered NVIDIA millions of pirated copyrighted books. Anna's Archive also offered access to several million books from Internet Archive, which were only normally available through Internet Archive's digital lending system (a system which was found to be copyright infringement by the Second Circuit, *see Hachette Book Grp., Inc. v. Internet Archive*, 115 F.4th 163 (2d Cir. 2024)). Anna's Archive promised NVIDIA access to "a lot of books," totaling roughly 500 terabytes of data. By downloading Anna's Archive, NVIDIA pirated additional copies of Plaintiff's Infringed Works.

58. On information and belief, in addition to Anna's Archive and The Pile, NVIDIA also downloaded books hosted or sourced from other shadow libraries, including LibGen, Sci-Hub, and Z-Library.

59. About four months after its exchange with Anna's Archive, in February 2024, NVIDIA released a model known as Nemotron-4 15B. The training data for this model was not publicly disclosed. Public documents, however, indicate that it was trained on 8 trillion tokens. The sources of the training data were never identified, and NVIDIA stated that it included "books." NVIDIA, however, has publicly stated that the training data for this model encompasses 70% from an "English natural language" dataset. This dataset itself is composed of 4.6% of books. Upon information and belief, to reach this percentage of tokens derived from books, the training data would need to include millions of books.

60. And a few months later, NVIDIA released the Nemotron-4 340B model. This model included the same 8 trillion tokens from the Nemotron-4 15B but added an additional 1 trillion tokens.

61. Upon information and belief, NVIDIA could not obtain the level of books needed for the Nemotron models without pirating copyrighted books, including Plaintiffs' Infringed Works.

62. In sum, NVIDIA has extensively and repeatedly violated the copyrights of Plaintiffs' Infringed Works including by acquiring these works from pirated sources, storing them, and enabling its employees to use them for any purpose, and copying them during the LLM training process.

63. Plaintiff Abdi Nazemian's book, *Like a Love Story*, was included in the Books3 dataset, based on public reporting about the dataset. This work is also available online through Anna's Archive, LibGen, and Z-Library.

64. Plaintiff Brian Keene's book, *Ghost Walk*, was included in the Books3 dataset, based on public reporting about the dataset. This work is also available online through Anna's Archive, LibGen, Z-Library, and Internet Archive.

65. Plaintiff Stewart O'Nan's book, *Last Night at the Lobster*, was included in the Books3 dataset, based on public reporting about the dataset. This work is also available online through Anna's Archive, LibGen, Z-Library, and Internet Archive.

66. Plaintiff Andre Dubus's books, *The Garden of Last Days*, *The Cage Keeper*, and *Townie: A Memoir* were included in the Books3 dataset, based on public reporting about the dataset. These works are also available online through Anna's Archive, LibGen, Z-Library, and Internet Archive.

67. Plaintiff Susan Orlean's books, *The Orchid Thief* and *The Library Book* were included in the Books3 dataset, based on public reporting about the dataset. These works are also available online through Anna's Archive, LibGen, and Z-Library.

68. NVIDIA's infringing activities, however, were not limited to downloading pirated copyrighted material to develop and train its own language models. NVIDIA also uploaded and distributed pirated copyrighted materials to third parties and provided the tools and means for numerous others to infringe Plaintiffs' copyrights.

69. NVIDIA downloaded pirated books from Anna's Archive, LibGen, and other shadow libraries using software called aria2c and a service called put.io, which use the BitTorrent peer-to-peer file-sharing protocol. Unlike a standard web browser download, with BitTorrent, one breaks a file into pieces and then each user's BitTorrent client in the same torrent "swarm" exchanges those pieces to download the file. Depending on the settings used, downloading content using the BitTorrent protocol can also result in uploading to other third parties in the swarm and making that content available for others in the swarm to download. In a BitTorrent swarm, uploading can occur while downloading, even when a user has only a partial copy of the file: as a user receives pieces of a file, by default the user also makes those pieces available to other users who have not yet received them. When NVIDIA ran the BitTorrent client, NVIDIA uploaded pieces of pirated book files to third parties during the downloading process (i.e., before the download was complete). NVIDIA also "seeded" pirated books (i.e., made available and uploaded parts of a completely downloaded file). While NVIDIA took some steps to limit seeding, it did not prevent seeding altogether. Plaintiffs'

books were among those that NVIDIA downloaded and, on information and belief, uploaded and distributed via torrent.

70. NVIDIA also distributed books it downloaded from shadow libraries to its business partners to further its commercial projects. These partnerships included a "PDF Extraction Project" with Adobe, Inc. involving 250 million pages acquired in part from shadow libraries such as Anna's Archive and LibGen (which, on information and belief, included Plaintiffs' books). NVIDIA gave its pirated books to Adobe because it wanted to "get tokens out of PDFs to create differentiated models" and "include PDF extraction into our NeMo pipeline so that customers can leverage their PDF content[.]" In return, Adobe committed to using NVIDIA's LLMs in "Document Cloud products (Acrobat)." This was a huge business opportunity for NVIDIA: Adobe had the promise to become a "big customer for NeMo and NVAIE – NeMo models powering every one of Adobe PDFs." When it came to sign the contract, NVIDIA went to great lengths to "expedit[e] internally to get it signed ASAP to move the project forward." On information and belief, NVIDIA proceeded to distribute Plaintiffs' works to Adobe.

71. NVIDIA also entered into partnerships with businesses like Cohere and Imbue that involved "sharing base data" and "offer[ing]" books that NVIDIA acquired from shadow libraries such as Anna's Archive and LibGen. On information and belief, NVIDIA proceeded to distribute Plaintiffs' works to Cohere, Imbue, and other customers interested in using NVIDIA's base training dataset (which included Plaintiffs' books).

72. NVIDIA also assisted and encouraged its customers to infringe Plaintiffs' copyrights. As CEO Jensen Huang explained in the keynote address at NVIDIA's 2023 GPU Technology Conference, as part of NVIDIA's "AI Foundations," customers can use the NeMo Framework (otherwise known as the NeMo Megatron Framework), to create and build their own AI models. As he stated, "[t]hroughout the entire process, NVIDIA AI experts will work with you, from creating your proprietary model to operations."[10]

[10] https://www.youtube.com/watch?v=DiGB5uAYKAg (40:00-:45).

73. Through the NeMo Megatron Framework and BigNLP platforms, NVIDIA provided customers with "scripts to automatically download and preprocess The Pile dataset which, until recently, was hosted externally by Eleuther Al." Meaning, NVIDIA provided tools and resources for its customers to use the NVIDIA platform to download The Pile, thereby infringing on Plaintiffs' copyrights. They scripts were developed to help their customers access these pirated datasets more quickly and easily. NVIDIA employees expressed concern about the "[t]ime needed for downloading pile files," so they developed and distributed code to "download and extract[] 30 pile files [in] ~70 minutes[,] which clearly shows the need for data prep parallelism."

74. For example, NVIDIA provided resources, guidance, and tools for its customer Writer Inc. to develop its line of Palmyra models using the NeMo Megatron Framework. On information and belief, NVIDIA provided the tools and scripts for Writer to download The Pile. NVIDIA provided similar assistance in downloading and processing The Pile to clients Persimmon AI Labs and Amazon. On information and belief, NVIDIA materially aided numerous other customers in downloading, using, and storing The Pile (and Books3) dataset.

75. NVIDIA provided the hardware too. Using the NeMo Framework, a customer could expect to quickly develop a language model trained on The Pile in only 9.8 days using NVIDIA's servers.

76. NVIDIA directly benefited from facilitating, supporting, and encouraging these infringing activities and attracted customers to use the NeMo Megatron Framework by providing quick access to The Pile (and Plaintiffs' books). In short, The Pile (and Books3) was key to NVIDIA attracting customers, and NVIDIA materially aided its customers to infringe Plaintiffs copyrights.

COUNT 1
Direct Copyright Infringement (17 U.S.C. § 501)
against NVIDIA

77. Plaintiffs incorporate by reference the preceding factual allegations.

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

79. To develop NVIDIA's LLMs, NVIDIA downloaded and copied The Pile and SlimPajama datasets. The Pile and SlimPajama datasets include the Books3 dataset, which includes the Infringed Works. NVIDIA made multiple copies of the Books3 dataset while developing its LLMs.

80. To develop NVIDIA's LLMs, NVIDIA downloaded and copied a dataset of books from Anna's Archive, which includes the Infringed Works. NVIDIA made multiple copies of this dataset while training its LLMs.

81. On information and belief, NVIDIA downloaded books hosted or sourced from other shadow libraries, including LibGen, Sci-Hub, and Z-Library.

82. NVIDIA used aria2c and put.io to download books from shadow libraries. The mechanism by which NVIDIA downloaded books using aria2c and put.io also made those books available to third parties and distributed them. NVIDIA therefore both reproduced and distributed downloaded books (on information and belief, including Plaintiffs' books) to third parties.

83. NVIDIA also shared books it downloaded from shadow libraries (on information and belief, including Plaintiffs' books) with other businesses in furtherance of joint commercial projects, an infringement of both the right of reproduction and distribution.

84. Plaintiffs and the Class members never authorized NVIDIA to make copies of their Infringed Works, make derivative works, publicly display copies (or derivative works), store copies, or distribute copies (or derivative works). All those rights belong exclusively to Plaintiffs under the U.S. Copyright Act.

85. NVIDIA made multiple copies of the Infringed Works, including when it downloaded these works from shadow libraries, and when it made additional copies during the training and development of its language models without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act. On information and belief, NVIDIA has continued to store and make copies of the Infringed Works.

86. Plaintiffs have been injured by NVIDIA's acts of direct copyright infringement. Plaintiffs are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law.

87. NVIDIA's violation of Plaintiffs' and Class members' exclusive right was willful because NVIDIA knew the datasets it downloaded, copied, and stored, and on which it "trained" its LLMs contained copyrighted works.

COUNT II

Contributory Copyright Infringement

against NVIDIA

88. Plaintiffs incorporate by reference the preceding factual allegations.

89. NVIDIA materially contributed to and directly assisted in the direct infringement by multiple customers, including at least Amazon, Persimmon AI, and Writer, by providing the technology, personnel, access to datasets, and other resources, such as the NeMo Megatron Framework, and variations of similar platforms and scripts that performed the same function; controlling or managing the property or other assets with which the direct infringement was accomplished; or providing business, legal, strategic, or operational guidance that allowed its customers to download, copy, and store Plaintiffs' and Class members' copyrighted works.

90. NVIDIA materially contributed to and directly assisted in the direct infringement of third parties by using aria2c and put.io to download books from shadow libraries. By using those mechanisms, NVIDIA uploaded and distributed downloaded books (on information and belief, including Plaintiffs' books) to third parties.

91. NVIDIA materially contributed to and directly assisted in the direct infringement of third parties by sharing books it downloaded from shadow libraries (on information and belief, including Plaintiffs' books) with other businesses.

92. NVIDIA knew or had reason to know of the direct infringement by others using the NeMo Megatron framework, because NVIDIA is fully aware of the capabilities of its own product, platforms and tools upon which third parties downloaded and acquired at least The Pile dataset, and potentially other datasets including copyrighted books as well.

93. Defendant is contributorily liable for the direct infringement of others that used the NeMo Framework to download and acquire The Pile dataset (and potentially other datasets containing copyrighted books as well).

COUNT III
Vicarious Copyright Infringement
against NVIDIA

94. Plaintiffs incorporate by reference the preceding factual allegations.

95. NVIDIA had the right and ability to control the direct infringements of customers, including at least Amazon, Persimmon AI, and Writer, using the NeMo Megatron Framework, and variations of similar platforms and scripts provided by NVIDIA that performed the same function, to download The Pile dataset (and potentially other datasets containing copyrighted books as well). NVIDIA failed to exert is right and ability to control its customers infringing acts.

96. NVIDIA has directly benefitted financially from the direct infringement of its customers because NVIDIA generated revenue from customers using the NeMo Megatron Framework to download The Pile (and Books3) dataset (and potentially other datasets containing copyrighted books as well).

97. Plaintiffs have been injured by NVIDIA's acts of vicarious copyright infringement. Plaintiffs are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law.

CLASS ALLEGATIONS

98. The "**Class Period**" as defined in this Complaint begins no later than March 8, 2021 and runs through the present. Because Plaintiffs do not yet know when the unlawful conduct alleged herein began, but believe, on information and belief, that the conduct likely began earlier than March 8, 2021, Plaintiffs reserve the right to amend the Class Period to comport with the facts and evidence uncovered during further investigation or through discovery.

99. **Class definition**. Plaintiffs bring 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 Class:

All persons or entities that own a registered United States copyright in any literary work that was downloaded or otherwise copied by Defendant and / or used by Defendant in LLM training, research, or development during the Class Period.

100. This Class definition excludes:

a. the 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.

101. **Numerosity**. Plaintiffs do 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 tens or hundreds of 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.

102. **Typicality.** Plaintiffs' claims are typical of the claims of other members of the Class because Plaintiffs 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.

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

104. **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 Plaintiffs and the Class when they obtained copies of Plaintiffs' Infringed Works

b. Whether Defendant violated the copyrights of Plaintiffs and the Class when they used them to research, develop, and train language models.

c. Whether Defendant intended to cause further infringement of the Infringed Works with these language models because they have distributed these models under an open license and advertised those models as a base from which to build further models.

d. Whether Defendant's support, facilitation, and encouragement of the infringement by NVIDIA's customers of Plaintiffs' and Proposed Class Members' copyrighted works constitutes vicarious or contributory infringement under the Copyright Act

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

f. Whether any statutes of limitation constrain the potential for recovery for Plaintiffs and the Class.

105. **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.

DEMAND FOR JUDGMENT

Wherefore, Plaintiffs request 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 Plaintiffs serving as Class Representatives, and with Plaintiffs' counsel as Class Counsel.

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

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

d) Reasonable attorneys' fees as available under 17 U.S.C. § 505 or other applicable statute.

e) Destruction or other reasonable disposition of all copies Defendant made or used in violation of the exclusive rights of Plaintiffs and the Class, under 17 U.S.C. § 503(b).

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

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

h) Further relief for Plaintiffs and the Class as may be just and proper.

JURY TRIAL DEMANDED

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

November 24, 2026

By: /s/ *Rohit D. Nath*

Joseph R. Saveri (CSB No. 130064)
Diane S. Rice (CSB No. 118303)
Christopher K.L. Young (CSB No. 318371)
Evan Creutz (CSB No. 349728)
Elissa A. Buchanan (CSB No. 249996)
William Waldir Castillo Guardado (CSB No. 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
ecreutz@saverilawfirm.com
eabuchanan@saverilawfirm.com
wcastillo@saverilawfirm.com

Avery M. Wolff (*pro hac vice*)
SAVERI LAW FIRM, LLP
780 Third Avenue, Suite 1200
New York, NY 10017
Telephone: (212) 970-0300
Facsimile: (212) 970-0301
Email: awolff@saverilawfirm.com

Bryan L. Clobes (admitted pro hac vice)
Mohammed A. Rathur (admitted pro hac vice)
Nabihah S. Maqbool (admitted pro hac vice)
CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP
135 South LaSalle Street, Suite 3210
Chicago, IL 60603
Telephone: (312) 782-4880
bclobes@caffertyclobes.com
mrathur@caffertyclobes.com
nmaqbool@caffertyclobes.com

Justin A. Nelson (admitted *pro hac vice*)

Alejandra C. Salinas (admitted *pro hac vice*)
SUSMAN GODFREY L.L.P.
1000 Louisiana Street, Suite 5100
Houston, TX 77002
Telephone: (713) 651-9366
Facsimile: (713) 654-6666
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (SBN 316062)
SUSMAN GODFREY L.L.P
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067-2906
Telephone: (310) 789-3100
RNath@susmangodfrey.com

Elisha Barron (admitted *pro hac vice*)
Craig Smyser (admitted *pro hac vice*)
SUSMAN GODFREY L.L.P.
One Manhattan West, 51st Floor
New York, NY 10019
Telephone: (212) 336-8330
ebarron@susmangodfrey.com
csmyser@susmangodfrey.com

Jordan W. Connors (admitted *pro hac vice*)
Trevor D. Nystrom (admitted *pro hac vice*)
Dylan B. Salzman (admitted *pro hac vice*)
SUSMAN GODFREY L.L.P
401 Union Street, Suite 3000
Seattle, WA 98101
Telephone: (206) 516-3880
jconnors@susmangodfrey.com
tnystrom@susmangodfrey.com
dsalzman@susmangodfrey.com

Rachel J. Geman (*pro hac vice*)
Danna Z. Elmasry (pro *hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, NY 10013
Tel.: 212.355.9500
rgeman@lchb.com
delmasry@lchb.com

Anne B. Shaver

LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111
Tel.: 415.956.1000
ashaver@lchb.com

Betsy A. Sugar (pro hac vice)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
222 2nd Avenue S. Suite 1640
Nashville, TN 37201
Tel.: 615.313.9000
bsugar@lchb.com

David A. Straite (admitted *pro hac vice*)
DiCELLO LEVITT LLP
485 Lexington Avenue, Suite 1001
New York, NY 10017
Tel. (646) 933-1000
dstraite@dicellolevitt.com

Amy E. Keller (admitted *pro hac vice*)
Nada Djordjevic (admitted *pro hac vice*)
James A. Ulwick (admitted *pro hac vice*)
DiCELLO LEVITT LLP
Ten North Dearborn Street, Sixth Floor
Chicago, Illinois 60602
Tel. (312) 214-7900
akeller@dicellolevitt.com
ndjordjevic@dicellolevitt.com
julwick@dicellolevitt.com

Brian O'Mara (SBN 229737)
DiCELLO LEVITT LLP
4747 Executive Drive
San Diego, California 92121
Telephone: (619) 923-3939
Facsimile: (619) 923-4233
briano@dicellolevitt.com

Matthew Butterick (State Bar No. 250953)
1920 Hillhurst Avenue, #406
Los Angeles, CA 90027
Telephone: (323) 968-2632
Facsimile: (415) 395-9940
mb@buttericklaw.com

Counsel for Individual and Representative Plaintiffs and the Proposed Class