Joseph R. Saveri (State Bar No. 130064)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1000
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com

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

Laura M. Matson (*pro hac vice* pending)
LOCKRIDGE GRINDAL NAUEN PLLP
100 Washington Avenue South, Suite 2200
Minneapolis, MN 55401
Telephone: (612) 339-6900
Facsimile: (612) 339-0981
Email: lmmatson@locklaw.com

Counsel for Individual and Representative
Plaintiffs and the Proposed Class

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

Abdi Nazemian, an individual; **Brian Keene**, an individual; and **Stewart O'Nan**, an individual; Individual and Representative Plaintiffs, v. **NVIDIA Corporation**, a Delaware corporation; Defendant.	Case No. **COMPLAINT** **CLASS ACTION** **DEMAND FOR JURY TRIAL**

Plaintiffs Abdi Nazemian, Brian Keene, and Stewart O'Nan (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. NeMo Megatron–GPT ("NeMo Megatron") is a series of large language models created by NVIDIA and released in September 2022.

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 were included in the training dataset that NVIDIA has admitted copying to train its NeMo Megatron models. Plaintiffs and Class members never authorized NVIDIA to use their copyrighted works as training material.

5. NVIDIA copied these copyrighted works multiple times to train its NeMo Megatron language models.

JURISDICTION AND VENUE

6. 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).

7. 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 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.

8. 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

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

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

11. 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*.

12. A nonexhaustive list of registered copyrights owned by Plaintiffs is included as Exhibit A.

DEFENDANT

13. 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

14. 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.

15. 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

16. NVIDIA is a diversified technology company founded in 1993 that originally focused on computer-graphics hardware and has since expanded to other computationally intensive fields, including software and hardware for training and operating AI software programs.

17. In September 2022, NVIDIA released its NeMo Megatron series of *large language models*. A large language model ("LLM") is AI software designed to emit convincingly naturalistic text outputs in response to user prompts.

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

19. During training, the LLM copies and ingests each textual work in the training dataset and extracts protected expression from it. The LLM progressively adjusts its output 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* that are stored within the model. These weights are entirely and uniquely derived from the protected expression in the training dataset. For instance, the

NeMo Megatron–GPT 20B language model is so named because the model stores 20 billion ("20B") weights derived from protected expression in its training dataset.

20. 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.

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

22. In September 2022, NVIDIA first 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 with. And this is not the end. We will continue to add more checkpoints in the future."[1] In this context "checkpoints" is an alternate term for language models within the NeMo Megatron series. 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.

23. Each of the NeMo Megatron models is hosted on a website called Hugging Face, where it has a *model card* that provides information about the model, including its training dataset. The model card for each of the NeMo Megatron models states that, "The model was trained on 'The Pile' dataset prepared by EleutherAI."[2]

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

[2] 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

24. 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"[3] (the "EleutherAI Paper").

25. 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.

26. 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.[4]

27. Bibliotik is one of a number of notorious "shadow library" websites that also includes Library Genesis (aka LibGen), Z-Library (aka B-ok), Sci-Hub, and Anna's Archive. These shadow libraries have long been of interest to the AI-training community because they host and distribute vast quantities of unlicensed copyrighted material. For that reason, these shadow libraries also violate the U.S. Copyright Act.

28. 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.

29. 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**.

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

[4] *Id.* at 3–4.

30. 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."[5]

31. In sum, NVIDIA has admitted training its NeMo Megatron models on a copy of The Pile dataset. Therefore, NVIDIA necessarily also trained its NeMo Megatron models on a copy of Books3, because Books3 is part of The Pile. Certain books written by Plaintiffs are part of Books3—including the Infringed Works—and thus NVIDIA necessarily trained its NeMo Megatron models on one or more copies of the Infringed Works, thereby directly infringing the copyrights of the Plaintiffs.

COUNT 1
DIRECT COPYRIGHT INFRINGEMENT (17 U.S.C. § 501)
AGAINST NVIDIA

32. Plaintiffs incorporate by reference the preceding factual allegations.

33. 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.

34. To train the NeMo Megatron language models, NVIDIA copied The Pile dataset. The Pile dataset includes the Books3 dataset, which includes the Infringed Works. NVIDIA made multiple copies of the Books3 dataset while training the NeMo Megatron models.

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

36. NVIDIA made multiple copies of the Infringed Works during the training of the NeMo Megatron models without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act. On information and belief, NVIDIA has continued to make copies of the Infringed Works for training other models.

[5] See https://huggingface.co/datasets/the_pile_books3

37. 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.

CLASS ALLEGATIONS

38. The "**Class Period**" as defined in this Complaint begins on at least 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.

39. **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 domiciled in the United States that own a United States copyright in any work that was used as training data for the NeMo Megatron large language models during the Class Period.**

40. 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.

41. **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

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.

42. **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.

43. **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.

44. **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 and used them to train the NeMo Megatron language models.

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

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

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

45. **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 is to be financially responsible 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.

Dated: March 8, 2024

By: */s/ Joseph R. Saveri*
Joseph R. Saveri

Joseph R. Saveri (State Bar No. 130064)
Christopher K. L. Young (State Bar No. 318371)
Elissa Buchanan (State Bar No. 249996)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1000
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
cyoung@saverilawfirm.com
eabuchanan@saverilawfirm.com

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

Brian D. Clark *(pro hac vice* pending)
Laura M. Matson *(pro hac vice* pending)
Arielle S. Wagner *(pro hac vice* pending)
Eura Chang *(pro hac vice* pending)
LOCKRIDGE GRINDAL NAUEN PLLP
100 Washington Avenue South, Suite 2200
Minneapolis, MN 55401
Telephone: (612) 339-6900
Facsimile: (612) 339-0981
Email: bdclark@locklaw.com
lmmatson@locklaw.com
aswagner@locklaw.com
echang@locklaw.com

Counsel for Individual and Representative Plaintiffs and the Proposed Class