

1 Joseph R. Saveri (State Bar No. 130064)
JOSEPH SAVERI LAW FIRM, LLP
 2 601 California Street, Suite 1000
 3 San Francisco, CA 94108
 Telephone: (415) 500-6800
 4 Facsimile: (415) 395-9940
 5 Email: jsaveri@saverilawfirm.com

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

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

15 *Counsel for Individual and Representative*
 16 *Plaintiffs and the Proposed Class*

17
 18 **UNITED STATES DISTRICT COURT**
 19 **NORTHERN DISTRICT OF CALIFORNIA**
 20 **SAN FRANCISCO DIVISION**

21 **Abdi Nazemian**, an individual;
 22 **Brian Keene**, an individual; and
Stewart O’Nan, an individual;

23 Individual and Representative Plaintiffs,

24 v.

25 **NVIDIA Corporation**, a Delaware corporation;
 26
 27 Defendant.

Case No.

COMPLAINT

CLASS ACTION

DEMAND FOR JURY TRIAL

1 Plaintiffs Abdi Nazemian, Brian Keene, and Stewart O’Nan (together “Plaintiffs”), on behalf of
2 themselves and all others similarly situated, bring this class-action complaint (“Complaint”) against
3 defendant NVIDIA Corporation (“NVIDIA” or “Defendant”).
4

5 OVERVIEW

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

8 2. A *large language model* is an AI software program designed to emit convincingly
9 naturalistic text outputs in response to user prompts. NeMo Megatron-GPT (“NeMo Megatron”) is a
10 series of large language models created by NVIDIA and released in September 2022.

11 3. Rather than being programmed in the traditional way—that is, by human programmers
12 writing code—a large language model is *trained* by copying an enormous quantity of textual works,
13 extracting protected expression from these works, and transforming that protected expression into a
14 large set of numbers called *weights* that are stored within the model. These weights are entirely and
15 uniquely derived from the protected expression in the training dataset. Whenever a large language
16 model generates text output in response to a user prompt, it is performing a computation that relies on
17 these stored weights, with the goal of imitating the protected expression ingested from the training
18 dataset.

19 4. Plaintiffs and Class members are authors. They own registered copyrights in certain
20 books that were included in the training dataset that NVIDIA has admitted copying to train its NeMo
21 Megatron models. Plaintiffs and Class members never authorized NVIDIA to use their copyrighted
22 works as training material.

23 5. NVIDIA copied these copyrighted works multiple times to train its NeMo Megatron
24 language models.
25

26 JURISDICTION AND VENUE

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

AGENTS AND CO-CONSPIRATORS

1
2 14. The unlawful acts alleged against the Defendant in this class action complaint were
3 authorized, ordered, or performed by the Defendant's respective officers, agents, employees,
4 representatives, or shareholders while actively engaged in the management, direction, or control of the
5 Defendant's businesses or affairs. The Defendant's agents operated under the explicit and apparent
6 authority of their principals. Defendant, and its subsidiaries, affiliates, and agents operated as a single
7 unified entity.

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

FACTUAL ALLEGATIONS

12
13
14 16. NVIDIA is a diversified technology company founded in 1993 that originally focused on
15 computer-graphics hardware and has since expanded to other computationally intensive fields,
16 including software and hardware for training and operating AI software programs.

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

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

24 19. During training, the LLM copies and ingests each textual work in the training dataset
25 and extracts protected expression from it. The LLM progressively adjusts its output to more closely
26 approximate the protected expression copied from the training dataset. The LLM records the results of
27 this process in a large set of numbers called *weights* that are stored within the model. These weights are
28 entirely and uniquely derived from the protected expression in the training dataset. For instance, the

1 NeMo Megatron-GPT 20B language model is so named because the model stores 20 billion (“20B”)
2 weights derived from protected expression in its training dataset.

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

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

11 22. In September 2022, NVIDIA first announced the availability of the NeMo Megatron
12 language models in a video on its website: “For the first time, NVIDIA is making its checkpoints
13 available publicly, where the checkpoints are trained with NeMo Megatron ... this is just to begin with.
14 And this is not the end. We will continue to add more checkpoints in the future.”¹ In this context
15 “checkpoints” is an alternate term for language models within the NeMo Megatron series. The
16 language models released in September 2022 include NeMo Megatron-GPT 1.3B, NeMo Megatron-
17 GPT 5B, NeMo Megatron-GPT 20B, and NeMo Megatron-T5 3B.

18 23. Each of the NeMo Megatron models is hosted on a website called Hugging Face, where
19 it has a *model card* that provides information about the model, including its training dataset. The model
20 card for each of the NeMo Megatron models states that, “The model was trained on ‘The Pile’ dataset
21 prepared by EleutherAI.”²

22
23
24
25 ¹ See <https://www.nvidia.com/en-us/on-demand/session/gtcfall22-a41200/?nvid=nv-int-tblg-881125>,
26 starting at 37:25.

27 ² See, e.g., <https://huggingface.co/nvidia/nemo-megatron-gpt-1.3B#training-data>,
28 <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>

1 24. The Pile is a training dataset curated by a research organization called EleutherAI. In
2 December 2020, EleutherAI introduced this dataset in a paper called “The Pile: An 800GB Dataset of
3 Diverse Text for Language Modeling”³ (the “EleutherAI Paper”).

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

7 26. The EleutherAI Paper further describes the contents of Books3:

8
9 Books3 is a dataset of books derived from a copy of the contents of the
10 Bibliotik private tracker ... Bibliotik consists of a mix of fiction and
11 nonfiction books and is almost an order of magnitude larger than our next
12 largest book dataset (BookCorpus2). We included Bibliotik because
13 books are invaluable for long-range context modeling research and
14 coherent storytelling.⁴

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

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

22 29. Plaintiffs’ copyrighted books listed in Exhibit A are among the works in the Books3
23 dataset. Below, these books are referred to as the **Infringed Works**.

24
25
26
27 ³ Available at <https://arxiv.org/pdf/2101.00027.pdf>

28 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

1 37. Plaintiffs have been injured by NVIDIA’s acts of direct copyright infringement.
2 Plaintiffs are entitled to statutory damages, actual damages, restitution of profits, and other remedies
3 provided by law.
4

5 **CLASS ALLEGATIONS**

6 38. The “**Class Period**” as defined in this Complaint begins on at least March 8, 2021 and
7 runs through the present. Because Plaintiffs do not yet know when the unlawful conduct alleged herein
8 began, but believe, on information and belief, that the conduct likely began earlier than March 8, 2021,
9 Plaintiffs reserve the right to amend the Class Period to comport with the facts and evidence uncovered
10 during further investigation or through discovery.

11 39. **Class definition.** Plaintiffs bring this action for damages and injunctive relief as a class
12 action under Federal Rules of Civil Procedure 23(a), 23(b)(2), and 23(b)(3), on behalf of the following
13 Class:

14 **All persons or entities domiciled in the United States that own a**
15 **United States copyright in any work that was used as training data for**
16 **the NeMo Megatron large language models during the Class Period.**

- 17 40. This Class definition excludes:
- 18 a. the Defendant named herein;
 - 19 b. any of the Defendant’s co-conspirators;
 - 20 c. any of Defendant’s parent companies, subsidiaries, and affiliates;
 - 21 d. any of Defendant’s officers, directors, management, employees, subsidiaries,
22 affiliates, or agents;
 - 23 e. all governmental entities; and
 - 24 f. the judges and chambers staff in this case, as well as any members of their
25 immediate families.

26 41. **Numerosity.** Plaintiffs do not know the exact number of members in the Class. This
27 information is in the exclusive control of Defendant. On information and belief, there are at least
28

1 thousands of members in the Class geographically dispersed throughout the United States. Therefore,
2 joinder of all members of the Class in the prosecution of this action is impracticable.

3 42. **Typicality.** Plaintiffs' claims are typical of the claims of other members of the Class
4 because Plaintiffs and all members of the Class were damaged by the same wrongful conduct of
5 Defendant as alleged herein, and the relief sought herein is common to all members of the Class.

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

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

- 14 a. Whether Defendant violated the copyrights of Plaintiffs and the Class when they
15 obtained copies of Plaintiffs' Infringed Works and used them to train the NeMo
16 Megatron language models.
- 17 b. Whether Defendant intended to cause further infringement of the Infringed Works with
18 the NeMo Megatron models because they have distributed these models under an open
19 license and advertised those models as a base from which to build further models.
- 20 c. Whether any affirmative defense excuses Defendant's conduct.
- 21 d. Whether any statutes of limitation constrain the potential for recovery for Plaintiffs and
22 the Class.

23 45. **Other class considerations.** Defendant has acted on grounds generally applicable to the
24 Class. This class action is superior to alternatives, if any, for the fair and efficient adjudication of this
25 controversy. Prosecuting the claims pleaded herein as a class action will eliminate the possibility of
26 repetitive litigation. There will be no material difficulty in the management of this action as a class
27 action. The prosecution of separate actions by individual Class members would create the risk of
28 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.

1 Dated: March 8, 2024

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

3 Joseph R. Saveri (State Bar No. 130064)
4 Christopher K. L. Young (State Bar No. 318371)
5 Elissa Buchanan (State Bar No. 249996)
6 **JOSEPH SAVERI LAW FIRM, LLP**
7 601 California Street, Suite 1000
8 San Francisco, CA 94108
9 Telephone: (415) 500-6800
10 Facsimile: (415) 395-9940
11 Email: jsaveri@saverilawfirm.com
12 cyoung@saverilawfirm.com
13 eabuchanan@saverilawfirm.com

14 Matthew Butterick (State Bar No. 250953)
15 1920 Hillhurst Avenue, #406
16 Los Angeles, CA 90027
17 Telephone: (323) 968-2632
18 Facsimile: (415) 395-9940
19 Email: mb@buttericklaw.com

20 Brian D. Clark (*pro hac vice* pending)
21 Laura M. Matson (*pro hac vice* pending)
22 Arielle S. Wagner (*pro hac vice* pending)
23 Eura Chang (*pro hac vice* pending)
24 **LOCKRIDGE GRINDAL NAUEN PLLP**
25 100 Washington Avenue South, Suite 2200
26 Minneapolis, MN 55401
27 Telephone: (612) 339-6900
28 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*