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

Stewart O'Nan, an individual;
Abdi Nazemian, an individual; and
Brian Keene, an individual;

Individual and Representative Plaintiffs,

v.

Databricks, Inc., a Delaware corporation; and
Mosaic ML, Inc., a Delaware corporation;

Defendants.

Case No.

COMPLAINT

CLASS ACTION

DEMAND FOR JURY TRIAL

Plaintiffs Stewart O'Nan, Abdi Nazemian, and Brian Keene (together "Plaintiffs"), on behalf of themselves and all others similarly situated, bring this class-action complaint ("Complaint") against defendants Mosaic ML, Inc. ("MosaicML") and Databricks, Inc. ("Databricks") (together "Defendants").

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. MosaicML Pretrained Transformer ("MPT") is a series of large language models created by MosaicML and distributed by Databricks.

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 MosaicML has admitted copying to train its MPT models. Plaintiffs and Class members never authorized MosaicML to use their copyrighted works as training material.

5. MosaicML copied these copyrighted works multiple times to train its MPT models.

6. Databricks, as the corporate parent of MosaicML and distributor of the MPT models, has also commercially benefitted from these acts of massive copyright infringement.

JURISDICTION AND VENUE

7. 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).

8. Jurisdiction and venue are proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Defendants are headquartered in this district. Defendants created the MPT language models and distribute 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. Defendants have 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. Defendants' 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.

PLAINTIFFS

10. 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*.

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

DEFENDANTS

14. Defendant Databricks is a Delaware corporation with its principal place of business at 160 Spear Street, 13th Floor, San Francisco CA 94105. Databricks acquired MosaicML in July 2023.

15. Defendant MosaicML is a Delaware corporation with its principal place of business at 501 2nd Street, Suite 202, San Francisco CA 94107. MosaicML operates as a subsidiary of Databricks.

AGENTS AND CO-CONSPIRATORS

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

17. 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 Defendants with respect to the acts, violations, and common course of conduct alleged herein.

FACTUAL ALLEGATIONS

18. MosaicML was founded in 2020 to provide tools to facilitate the training of AI models.

19. In May 2023, MosaicML released the first in its MPT series of *large language models*, called MPT-7B. A large language model ("LLM") is AI software designed to emit convincingly naturalistic text outputs in response to user prompts.

20. 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*.

21. 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 MPT-7B language model is so named because the model stores seven billion ("7B") weights derived from protected expression in its training dataset.

22. 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.

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

24. In May 2023, MosaicML first announced MPT-7B in a blog post called "Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs."[1] In the blog post, MosaicML describes the MPT-7B training dataset as a "MosaicML-curated mix of sources … [that] emphasizes English natural language text … and includes elements of the recently-released RedPajama dataset."

25. In a table describing the composition of the MPT-7B training dataset, MosaicML notes that a large quantity of data comes from a component dataset called "RedPajama—Books". MosaicML's blog post does not further describe the contents of the "RedPajama—Books" dataset.

26. But information about RedPajama is available elsewhere. The RedPajama dataset is hosted on a website called Hugging Face.[2] According to the documentation for the RedPajama dataset

[1] Available at https://www.databricks.com/blog/mpt-7b

[2] Available at https://huggingface.co/datasets/togethercomputer/RedPajama-Data-1T

found there, its "Books" component is actually a copy of the "Books3 dataset" that is "downloaded from Huggingface [sic]" when a user runs the script that automatically assembles the RedPajama dataset. Therefore, anyone who uses the "RedPajama—Books" dataset for training an AI model is actually using a copy of the Books3 dataset. The documentation for the RedPajama dataset does not further describe the contents of Books3.

27. But information about Books3 is available elsewhere. Books3 is a component of a separate AI training dataset called The Pile, 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] This paper 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]

28. 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.

29. 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.

30. 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.

31. 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]

32. But before then, anyone who used the "RedPajama—Books" dataset for training necessarily made a copy of the Books3 dataset. This includes MosaicML, which completed the training of MPT-7B before May 2023.

33. In sum, MosaicML has admitted training its MPT-7B model on a copy of the "RedPajama—Books" dataset, which in turn is a copy of the Books3 dataset. Therefore, MosaicML necessarily trained its MPT-7B model on a copy of Books3. Certain books written by Plaintiffs are part of Books3—including the Infringed Works—and thus MosaicML necessarily trained MPT-7B on one or more copies of the Infringed Works, thereby directly infringing the copyrights of the Plaintiffs.

34. In June 2023, MosaicML released another member of the MPT series of large language models, called MPT-30B. As the name suggests, MPT-30B contained 30 billion weights—over quadruple the size of MPT-7B—derived from its training dataset. In a table describing the composition of the MPT-30B training dataset, MosaicML admitted that once again, a large quantity of training data came from "RedPajama—Books."[6] By copying the "RedPajama—Books" dataset, MosaicML necessarily copied Books3—including the Infringed Works—thereby again directly infringing the copyrights of the Plaintiffs.

35. On information and belief, since it was acquired by Databricks in July 2023, MosaicML has continued to make copies of the Infringed Works for LLM training and other commercial purposes.

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

[6] Available at https://www.mosaicml.com/blog/mpt-30b

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

36. Plaintiffs incorporate by reference the preceding factual allegations.

37. 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.

38. To train the MPT-7B and MPT-30B language models, MosaicML copied the Books3 dataset, which includes the Infringed Works. MosaicML made multiple copies of the Books3 dataset during the training of the MPT-7B and MPT-30B models.

39. On information and belief, MosaicML made further copies of the Books3 dataset or subsets thereof to train other models in the MPT family. For instance, MosaicML released a model called MPT-7B-StoryWriter-65k+ ("the StoryWriter model"), a variant of MPT-7B that MosaicML admits was further trained on "a filtered fiction subset of the [B]ooks3 dataset".[7] The stated purpose of the StoryWriter model is "to read and write stories"—or, put another way, to generate works that directly compete with works in the training dataset.

40. Plaintiffs and the Class members never authorized MosaicML 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 and the Class members under the U.S. Copyright Act.

41. MosaicML repeatedly copied the Infringed Works without Plaintiffs' permission. MosaicML made these copies without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act.

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

[7] See https://www.databricks.com/blog/mpt-7b

COUNT 2
VICARIOUS COPYRIGHT INFRINGEMENT AGAINST DATABRICKS

43. Plaintiffs incorporate by reference the preceding factual allegations.

44. Databricks acquired MosaicML in July 2023. As the corporate parent of MosaicML, Databricks had the right and ability to control the direct infringements alleged in Count 1 committed by MosaicML, at minimum those occurring after the acquisition. Databricks failed to exert its right and ability to control MosaicML's infringements.

45. Databricks has directly benefitted financially from the direct infringement alleged in Count 1 because MosaicML generates revenue from its infringing activities, and this revenue belongs to Databricks.

46. Plaintiffs have been injured by Databricks'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

47. 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.

48. **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 MPT series of language models during the Class Period.**

49. This Class definition excludes:

a. Defendants named herein;

b. any of the Defendants' co-conspirators;

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

d. any of Defendants' 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.

50. **Numerosity**. Plaintiffs do not know the exact number of members in the Class. This information is in the exclusive control of Defendants. 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.

51. **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 Defendants as alleged herein, and the relief sought herein is common to all members of the Class.

52. **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.

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

a. Whether Defendants violated the copyrights of Plaintiffs and the Class when they obtained copies of Plaintiffs' Infringed Works and copied the Infringed Works into the dataset used to train the MPT models.

b. Whether Defendants intended to cause further infringement of the Infringed Works with the MPT models because they have distributed these models as so-called "open source" models and advertised those models as a base from which to build further models.

c. Whether any affirmative defense excuses Defendants' conduct.

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

54. **Other class considerations**. Defendants have 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 Defendants.

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 Defendants.

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 Defendants 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 Defendants.

g) Defendants are to be jointly and severally 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.

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)
Cadio Zirpoli (State Bar No. 179108)
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
czirpoli@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