Joseph R. Saveri (State Bar No. 130064)
Cadio Zirpoli (State Bar No. 179108)
Christopher K.L. Young (State Bar No. 318371)
Kathleen J. McMahon (State Bar No. 340007)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1000
San Francisco, California 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
cyoung@saverilawfirm.com
kmcmahon@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

Counsel for Individual and Representative Plaintiffs and the Proposed Class

UNITED STATES DISTRICT COURT
NORTHERN DISTRICT OF CALIFORNIA
SAN FRANCISCO DIVISION

RICHARD KADREY, an individual; SARAH SILVERMAN, an individual; CHRISTOPHER GOLDEN, an individual;

Individual and Representative Plaintiffs,

v.

META PLATFORMS, INC., a Delaware corporation;

Defendant.

Case No.

COMPLAINT

CLASS ACTION

DEMAND FOR JURY TRIAL

Plaintiffs Richard Kadrey, Sarah Silverman, and Christopher Golden ("Plaintiffs"), on behalf of themselves and all others similarly situated, bring this Class Action Complaint (the "Complaint") against Defendant Meta Platforms, Inc.

I. OVERVIEW

1. LLaMA is a set of large language models created and maintained by Defendant Meta Platforms, Inc. A *large language model* is an AI software program designed to emit convincingly naturalistic text outputs in response to user prompts.

2. Rather than being programmed in the traditional way, a large language model is "trained" by copying massive amounts of text and extracting expressive information from it. This body of text is called the *training dataset*.

3. A large language model's output is therefore entirely and uniquely reliant on the material in its training dataset. Every time it assembles a text output, the model relies on the information it extracted from its training dataset. Thus, the decisions about what textual information to include in the training dataset are deliberate and important choices.

4. Plaintiffs and Class members are authors of books. Plaintiffs and Class members have copyrights in the books they published. Plaintiffs and Class members did not consent to the use of their copyrighted books as training material for LLaMA.

5. Nonetheless, their copyrighted materials were copied and ingested as part of training LLaMA. Many of Plaintiffs' copyrighted books appear in the dataset that Meta has admitted to using to train LLaMA.

II. 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) and the Digital Millennium Copyright Act (17 U.S.C. § 1202).

7. Jurisdiction and venue is proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Defendant Meta Platforms, Inc. ("Meta") is headquartered in this district, and thus a substantial part of the events giving rise to the claim occurred in this district; and because a substantial part of the events giving rise to Plaintiffs' claims occurred in this District, and 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(d), assignment of this case to the San Francisco or Oakland Division is proper because Meta is headquartered in San Mateo County, where a substantial part of the events giving rise to the claim occurred, a substantial amount part of the events giving rise to Plaintiffs' claims and the interstate trade and commerce involved and affected by Defendant's conduct giving rise to the claims herein occurred in this Division.

III. PARTIES

A. Plaintiffs

9. Plaintiff Richard Kadrey is a writer who lives in Pennsylvania. Plaintiff Kadrey owns registered copyrights in several books, including *Sandman Slim*. These books contain the copyright-management information customarily included in published books, including the name of the author and the year of publication.

10. Plaintiff Sarah Silverman is a writer and performer who lives in California. Plaintiff Silverman owns a registered copyright in one book, called *The Bedwetter.* This book contains copyright-management information customarily included in published books, including the name of the author and the year of publication.

11. Plaintiff Christopher Golden is a writer who lives in Massachusetts. Mr. Golden owns registered copyrights in several books, including *Ararat*. These books contain the copyright-management information customarily included in published books, including the name of the author and the year of publication.

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

B. Defendant

13. Defendant Meta is a Delaware corporation with its principal place of business at 1601 Willow Road, Menlo Park, California 94025.

IV. 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. Each Defendant, and its subsidiaries, affiliates, and agents operated as a single unified entity.

15. Various persons and/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 other Defendants with respect to the acts, violations, and common course of conduct alleged herein.

V. FACTUAL ALLEGATIONS

16. Meta is a diversified internet company that creates, markets, and sells software and hardware technology products, including Facebook, Instagram, and Horizon Worlds. Meta also has a large artificial-intelligence group called Meta AI that creates and distributes artificial-intelligence software products.

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

18. In February 2023, Meta released an AI product called LLaMA. LLaMA is a set of *large language models*. A large language model (or "LLM" for short) is AI software designed to parse and emit natural language. Though a large language model is a software program, it is not created the way most software programs are—that is, by human software engineers writing code. Rather, a large language model is "trained" by copying massive amounts of text from various sources and feeding these copies into the model. This corpus of input material is called the *training dataset*. During training, the large language model copies each piece of text in the training dataset and extracts expressive

information from it. The large language model progressively adjusts its output to more closely resemble the sequences of words copied from the training dataset. Once the large language model has copied and ingested all this text, it is able to emit convincing simulations of natural written language as it appears in the training dataset.

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

20. Authors, including Plaintiffs, publish books with certain copyright management information. This information includes the book's title, the ISBN number or copyright number, the author's name, the copyright holder's name, and terms and conditions of use. Most commonly, this information is bound on the back of the book's title page and is standard in any book, regardless of genre.

21. Meta introduced LLaMA in a paper called "LLaMA: Open and Efficient Foundation Language Models". In the paper, Meta describes the LLaMA training dataset as "a large quantity of textual data" that was chosen because it was "publicly available, and compatible with open sourcing."

22. *Open sourcing* refers to putting data under a permissive style of copyright license called an *open-source license*. Copyrighted materials, however, are not ordinarily "compatible with open sourcing" unless and until the copyright owner first places the material under an open-source license, thereby enabling others to do so later.

23. In a table describing the composition of the LLaMA training dataset, Meta notes that 85 gigabytes of the training data comes from a category called "Books." Meta further elaborates that "Books" comprises the text of books from two internet sources: (1) Project Gutenberg, an online archive of approximately 70,000 books that are out of copyright, and (2) "the Books3 section of ThePile . . . a publicly available dataset for training large language models." Meta's paper on LLaMA does not further describe the contents of Books3 or ThePile.

24. But that information is available elsewhere. ThePile is a dataset assembled 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".

25. 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 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.

27. Bibliotik is one of a number of notorious "shadow library" websites that also includes Library Genesis (aka LibGen), Z-Library (aka B-ok), and Sci-Hub. The books and other materials aggregated by these websites have also been available in bulk via torrent systems. These shadow libraries have long been of interest to the AI-training community because of the large quantity of copyrighted material they host. For that reason, these shadow libraries are also flagrantly illegal.

28. The person who assembled the Books3 dataset has confirmed in public statements that it represents "all of Bibliotik" and contains 196,640 books. EleutherAI currently distributes copies of Books3 through its website (https://pile.eleuther.ai/).

29. The Books3 dataset is also available from a popular AI project hosting service called Hugging Face (https://huggingface.co/datasets/the_pile_books3).

30. Many of Plaintiffs' books appear in the Books3 dataset. A list of Plaintiffs' books currently known to exist in the Books3 dataset is attached as Exhibit B. Together, these books are referred to as the **Infringed Works**.

31. Since the launch of the LLaMA language models in February 2023, Meta has made these models selectively available to organizations that request access, saying:

> To maintain integrity and prevent misuse, we are releasing our model under a noncommercial license focused on research use cases. Access to the model will be granted on a case-by-case basis to academic researchers; those affiliated with organizations in government, civil society, and academia; and industry research laboratories around the world. People interested in applying for access can find the link to the application in our research paper.

32. Meta has not disclosed what criteria it uses to decide who is eligible to receive the LLaMA language models, nor who has actually received them, or whether Meta has in fact adhered to its stated criteria. On information and belief, Meta has in fact distributed the LLaMA models to certain people and entities, continues to do so, and has benefited financially from these acts.

33. In March 2023, the LLaMA language models were leaked to a public internet site and have continued to circulate. Meta has not disclosed what role it had, if any, in the leak.

34. Later in March 2023, Meta issued a DMCA takedown notice to a programmer on GitHub who had released a tool that helped users download the leaked LLaMA language models. In the notice, Meta asserted copyright over the LLaMA language models.

35. According to reporting in June 2023, Meta plans to make the next version of LLaMA commercially available.

VI. CLAIMS FOR RELIEF

COUNT 1
DIRECT COPYRIGHT INFRINGEMENT
17 U.S.C. § 106

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 LLaMA language models, Meta copied the Books3 dataset, which includes the Infringed Works.

39. Plaintiffs never authorized Meta 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 copyright law.

40. Meta made copies of the Infringed Works during the training process of the LLaMA language models without Plaintiffs' permission.

41. Because the LLaMA language models cannot function without the expressive information extracted from Plaintiffs' Infringed Works and retained inside the LLaMA language models, these LLaMA language models are themselves infringing derivative works, made without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act.

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

COUNT 2
VICARIOUS COPYRIGHT INFRINGEMENT
17 U.S.C. § 106

43. Plaintiffs incorporate by reference the preceding factual allegations.

44. Because the output of the LLaMA language models is based on expressive information extracted from Plaintiffs' Infringed Works, every output of the LLaMA language models is an infringing derivative work, made without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act.

45. Meta has the right and ability to control the output of the LLaMA language models. Meta has benefited financially from the infringing output of the LLaMA language models. Therefore, every output from the LLaMA language models constitutes an act of vicarious copyright infringement.

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

COUNT 3
REMOVAL OF COPYRIGHT-MANAGEMENT INFORMATION AND FALSE ASSERTION OF COPYRIGHT
17 U.S.C. § 1202(B)

47. Plaintiffs incorporate by reference the preceding factual allegations.

48. Plaintiffs included one or more forms of copyright-management information ("CMI") in each of the Plaintiffs' Infringed Works, including: copyright notice, title and other identifying information, or the name or other identifying information about the owners of each book, terms and conditions of use, and identifying numbers or symbols referring to CMI.

49. Without the authority of Plaintiffs and the Class, Meta copied the Plaintiffs' Infringed Works and used them as training data for the LLaMA language models. By design, the training process does not preserve any CMI. Therefore, Meta intentionally removed CMI from the Plaintiffs' Infringed Works in violation of 17 U.S.C. § 1202(b)(1).

50. Without the authority of Plaintiffs and the Class, Defendant created derivative works based on Plaintiffs' Infringed Works. By distributing these works without their CMI, Meta violated 17 U.S.C. § 1202(b)(3).

51. By falsely claiming that it has sole copyright in the LLaMA language models—which it cannot, because the LLaMA language models are infringing derivative works—Meta violated 17 U.S.C. § 1202(a)(1).

52. Meta knew or had reasonable grounds to know that this removal of CMI would facilitate copyright infringement by concealing the fact that every output from the LLaMA language models is an infringing derivative work, synthesized entirely from expressive information found in the training data.

53. Plaintiffs have been injured by Meta's removal of CMI. Plaintiffs are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law.

COUNT 4
UNFAIR COMPETITION
CAL. BUS. & PROF. CODE §§ 17200, *ET SEQ.*

54. Plaintiffs incorporate by reference the preceding factual allegations.

55. Defendant has engaged in unlawful business practices, including violating Plaintiffs' and the Class's rights under the DMCA, and using Plaintiffs' Infringed Works to train LLaMA without Plaintiffs' or the Class's authorization.

56. The unlawful business practices described herein violate California Business and Professions Code section 17200 *et seq*. because that conduct is otherwise unlawful by violating the DMCA.

57. The unlawful business practices described herein violate California Business and Professions Code section 17200 *et seq*. because they are unfair, immoral, unethical, oppressive, unscrupulous or injurious to consumers, because, among other reasons, Defendant used Plaintiffs' protected works to train LLaMA for Defendant's own gain without Plaintiffs' and the Class's authorization.

58. The unlawful business practices described herein violate California Business and Professions Code section 17200 *et seq*. as fraudulent because consumers are likely to be deceived

because, among other reasons, Meta caused LLaMA's output to be emitted without any credit to Plaintiffs' or the Class whose Infringed Works comprise LLaMA's training dataset.

COUNT 5
UNJUST ENRICHMENT
CALIFORNIA COMMON LAW

1. Plaintiffs incorporate by reference the preceding factual allegations.

2. Plaintiffs and the Class have invested substantial time and energy in creating the Infringed Works.

3. Defendants have unjustly utilized access to the Infringed Materials to train LLaMA.

4. Plaintiffs did not consent to the unauthorized use of the Infringed Materials to train LLaMA.

5. By using Plaintiffs' Infringed Works to train LLaMA, Plaintiffs and the Class were deprived of the benefits of their work, including monetary damages.

6. Defendants derived or intend to derive profit and other benefits from the use of the Infringed Materials to train LLaMA.

7. It would be unjust for Defendant to retain those benefits.

8. The conduct of Defendant is causing and, unless enjoined and restrained by this Court, will continue to cause Plaintiffs and the Class great and irreparable injury that cannot fully be compensated or measured in money.

COUNT 6
NEGLIGENCE
CALIFORNIA COMMON LAW

9. Plaintiffs incorporate by reference the preceding factual allegations.

10. Defendant owed a duty of care toward Plaintiffs and the Class based upon Defendant's relationship to them. This duty is based upon Defendant's obligations, custom and practice, right to control information in its possession, exercise of control over the information in its possession, authority to control the information in its possession, and the commission of affirmative acts that result in said harms and losses. Additionally, this duty is based on the requirements of California Civil Code

section 1714, requiring all "persons," including Defendant, to act in a reasonable manner toward others.

11. Defendant breached its duties by negligently, carelessly, and recklessly collecting, maintaining and controlling Plaintiffs' and Class members' Infringed Works and engineering, designing, maintaining and controlling systems—including LLaMA—which are trained on Plaintiffs' and Class members' Infringed Works without their authorization.

12. Defendant owed Plaintiffs and Class members a duty of care to maintain Plaintiffs' Infringed Works once collected and ingested for training LLaMA.

13. Defendant also owed Plaintiffs and Class members a duty of care to not use the Infringed Works in a way that would foreseeably cause Plaintiffs and Class members injury, for instance, by using the Infringed Works to train LLaMA.

14. Defendant breached their duties by, *inter alia*, the Infringed Works to train LLaMA.

VII. CLASS ALLEGATIONS

A. Class Definition

15. 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 LLaMA language models during the Class Period.**

16. This Class definition excludes:

a. Defendant named herein;

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

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

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

e. all governmental entities; and

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

B. Numerosity

17. 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.

C. Typicality

18. 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.

D. Adequacy

19. 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.

E. Commonality and Predominance

20. Numerous questions of law or fact common to each Class arise from Defendant's conduct:

a. whether Defendant violated the copyrights of Plaintiffs and the Class when they downloaded copies of Plaintiff's Infringed Works and used them to train the LLaMA language models;

b. whether the LLaMA language models are themselves infringing derivative works based on Plaintiffs' Infringed Works;

c. whether the text outputs of the LLaMA language models are infringing derivative works based on Plaintiffs' Infringed Works;

d. whether Defendant violated the DMCA by removing copyright-management information (CMI) from Plaintiffs' Infringed Works.

e. Whether Defendant was unjustly enriched by the unlawful conduct alleged herein.

f. Whether Defendant's conduct alleged herein constitutes Unfair Competition under California Business and Professions Code section 17200 *et seq.*

g. Whether Defendant's conduct alleged herein constitutes common unfair competition

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

i. Whether any statutes of limitation limits Plaintiffs' and the Class's potential for recovery.

21. These and other questions of law and fact are common to the Class predominate over any questions affecting the members of the Class individually.

F. Other Class Considerations

22. 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.

23. The prosecution of separate actions by individual Class members would create the risk of inconsistent or varying adjudications, establishing incompatible standards of conduct for Defendants.

VIII. 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) Permanent injunctive relief, including but not limited to changes to the LLaMA language models to ensure that all applicable information set forth in 17 U.S.C. § 1203(b)(1) is included when appropriate.

e) An order of costs and allowable attorney's fees under 17 U.S.C. § 1203(b)(4)–(5).

f) An award of statutory damages under 17 U.S.C. § 1203(b)(3) and 17 U.S.C. § 1203(c)(3), or in the alternative, an award of actual damages and any additional profits under 17 U.S.C. § 1203(c)(2) (including tripling damages under 17 U.S.C. § 1203(c)(4) if applicable).

g) 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.

h) 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.

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

IX. 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: July 7, 2023

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)
Kathleen J. McMahon (State Bar No. 340007)
JOSEPH SAVERI LAW FIRM, LLP
601 California Street, Suite 1000
San Francisco, California 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
Email: jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
cyoung@saverilawfirm.com
kmcmahon@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

Counsel for Individual and Representative Plaintiffs and the Proposed Class