

1 Joseph R. Saveri (State Bar No. 130064)
 2 Cadio Zirpoli (State Bar No. 179108)
 3 Christopher K.L. Young (State Bar No. 318371)
 4 Kathleen J. McMahon (State Bar No. 340007)
JOSEPH SAVERI LAW FIRM, LLP
 5 601 California Street, Suite 1000
 6 San Francisco, California 94108
 Telephone: (415) 500-6800
 7 Facsimile: (415) 395-9940
 Email: jsaveri@saverilawfirm.com
 8 czirpoli@saverilawfirm.com
 cyoung@saverilawfirm.com
 9 kmcmahon@saverilawfirm.com

10 Matthew Butterick (State Bar No. 250953)
 11 1920 Hillhurst Avenue, #406
 Los Angeles, CA 90027
 Telephone: (323) 968-2632
 12 Facsimile: (415) 395-9940
 13 Email: mb@buttericklaw.com

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

16 **UNITED STATES DISTRICT COURT**
 17 **NORTHERN DISTRICT OF CALIFORNIA**
SAN FRANCISCO DIVISION

18 RICHARD KADREY, an individual;
 19 SARAH SILVERMAN, an individual;
 20 CHRISTOPHER GOLDEN, an individual;
 21 Individual and Representative Plaintiffs,
 22 v.
 23 META PLATFORMS, INC., a Delaware
 24 corporation;
 25 Defendant.

Case No.
COMPLAINT
CLASS ACTION
DEMAND FOR JURY TRIAL

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

4 I. OVERVIEW

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

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

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

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

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

21 II. JURISDICTION AND VENUE

22 6. This Court has subject matter jurisdiction under 28 U.S.C. § 1331 because this case
23 arises under the Copyright Act (17 U.S.C. § 501) and the Digital Millennium Copyright Act (17 U.S.C.
24 § 1202).

25 7. Jurisdiction and venue is proper in this judicial district under 28 U.S.C. § 1391(c)(2)
26 because Defendant Meta Platforms, Inc. (“Meta”) is headquartered in this district, and thus a
27 substantial part of the events giving rise to the claim occurred in this district; and because a substantial
28 part of the events giving rise to Plaintiffs’ claims occurred in this District, and a substantial portion of

1 the affected interstate trade and commerce was carried out in this District. Defendant has transacted
2 business, maintained substantial contacts, and/or committed overt acts in furtherance of the illegal
3 scheme and conspiracy throughout the United States, including in this District. Defendant's conduct
4 has had the intended and foreseeable effect of causing injury to persons residing in, located in, or doing
5 business throughout the United States, including in this District.

6 8. Under Civil Local Rule 3.2(d), assignment of this case to the San Francisco or Oakland
7 Division is proper because Meta is headquartered in San Mateo County, where a substantial part of the
8 events giving rise to the claim occurred, a substantial amount part of the events giving rise to Plaintiffs'
9 claims and the interstate trade and commerce involved and affected by Defendant's conduct giving rise
10 to the claims herein occurred in this Division.

11 III. PARTIES

12 A. Plaintiffs

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

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

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

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

1 **B. Defendant**

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

4 **IV. AGENTS AND CO-CONSPIRATORS**

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

11 15. Various persons and/or firms not named as Defendants may have participated as co-
12 conspirators in the violations alleged herein and may have performed acts and made statements in
13 furtherance thereof. Each acted as the principal, agent, or joint venture of, or for other Defendants with
14 respect to the acts, violations, and common course of conduct alleged herein.

15 **V. FACTUAL ALLEGATIONS**

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

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

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

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

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

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

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

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

20 23. In a table describing the composition of the LLaMA training dataset, Meta notes that
21 85 gigabytes of the training data comes from a category called “Books.” Meta further elaborates that
22 “Books” comprises the text of books from two internet sources: (1) Project Gutenberg, an online
23 archive of approximately 70,000 books that are out of copyright, and (2) “the Books3 section of
24 ThePile . . . a publicly available dataset for training large language models.” Meta’s paper on LLaMA
25 does not further describe the contents of Books3 or ThePile.

26 24. But that information is available elsewhere. ThePile is a dataset assembled by a research
27 organization called EleutherAI. In December 2020, EleutherAI introduced this dataset in a paper
28 called “The Pile: An 800GB Dataset of Diverse Text for Language Modeling”.

1 25. The EleutherAI paper reveals that the Books3 dataset comprises 108 gigabytes of data,
2 or approximately 12% of the dataset, making it the third largest component of The Pile by size.

3 26. The EleutherAI paper describes the contents of Books3:

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

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

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

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

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

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

25 To maintain integrity and prevent misuse, we are releasing our model
26 under a noncommercial license focused on research use cases. Access to
27 the model will be granted on a case-by-case basis to academic
28 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.

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

3
4 **COUNT 5**
5 **UNJUST ENRICHMENT**
6 **CALIFORNIA COMMON LAW**

7 1. Plaintiffs incorporate by reference the preceding factual allegations.

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

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

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

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

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

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

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

21 **COUNT 6**
22 **NEGLIGENCE**
23 **CALIFORNIA COMMON LAW**

24 9. Plaintiffs incorporate by reference the preceding factual allegations.

25 10. Defendant owed a duty of care toward Plaintiffs and the Class based upon Defendant's
26 relationship to them. This duty is based upon Defendant's obligations, custom and practice, right to
27 control information in its possession, exercise of control over the information in its possession,
28 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

1 section 1714, requiring all “persons,” including Defendant, to act in a reasonable manner toward
2 others.

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

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

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

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

13 VII. CLASS ALLEGATIONS

14 A. Class Definition

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

17 **All persons or entities domiciled in the United States that own a**
18 **United States copyright in any work that was used as training data**
19 **for the LLaMA language models during the Class Period.**

20 16. This Class definition excludes:

- 21 a. Defendant named herein;
- 22 b. any of the Defendant’s co-conspirators;
- 23 c. any of Defendant’s parent companies, subsidiaries, and affiliates;
- 24 d. any of Defendant’s officers, directors, management, employees, subsidiaries,
25 affiliates, or agents;
- 26 e. all governmental entities; and
- 27 f. the judges and chambers staff in this case, as well as any members of their
28 immediate families.

1 **B. Numerosity**

2 17. Plaintiffs do not know the exact number of members in the Class. This information is in
3 the exclusive control of Defendant. On information and belief, there are at least thousands of members
4 in the Class geographically dispersed throughout the United States. Therefore, joinder of all members
5 of the Class in the prosecution of this action is impracticable.

6 **C. Typicality**

7 18. Plaintiffs' claims are typical of the claims of other members of the Class because
8 Plaintiffs and all members of the Class were damaged by the same wrongful conduct of Defendant as
9 alleged herein, and the relief sought herein is common to all members of the Class.

10 **D. Adequacy**

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

16 **E. Commonality and Predominance**

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

- 19 a. whether Defendant violated the copyrights of Plaintiffs and the Class when they
20 downloaded copies of Plaintiff's Infringed Works and used them to train the LLaMA
21 language models;
- 22 b. whether the LLaMA language models are themselves infringing derivative works based
23 on Plaintiffs' Infringed Works;
- 24 c. whether the text outputs of the LLaMA language models are infringing derivative works
25 based on Plaintiffs' Infringed Works;
- 26 d. whether Defendant violated the DMCA by removing copyright-management information
27 (CMI) from Plaintiffs' Infringed Works.
- 28 e. Whether Defendant was unjustly enriched by the unlawful conduct alleged herein.

- 1 f. Whether Defendant's conduct alleged herein constitutes Unfair Competition under
- 2 California Business and Professions Code section 17200 *et seq.*
- 3 g. Whether Defendant's conduct alleged herein constitutes common unfair competition
- 4 h. Whether any affirmative defense excuses Defendant's conduct.
- 5 i. Whether any statutes of limitation limits Plaintiffs' and the Class's potential for recovery.

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

8 **F. Other Class Considerations**

9 22. Defendants have acted on grounds generally applicable to the Class. This class action is
10 superior to alternatives, if any, for the fair and efficient adjudication of this controversy. Prosecuting the
11 claims pleaded herein as a class action will eliminate the possibility of repetitive litigation. There will be
12 no material difficulty in the management of this action as a class action.

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

16 **VIII. DEMAND FOR JUDGMENT**

17 WHEREFORE, Plaintiffs request that the Court enter judgment on their behalf and on behalf of
18 the Class defined herein, by ordering:

- 19 a) This action may proceed as a class action, with Plaintiffs serving as Class
20 Representatives, and with Plaintiffs' counsel as Class Counsel.
- 21 b) Judgment in favor of Plaintiffs and the Class and against Defendant.
- 22 c) An award of statutory and other damages under 17 U.S.C. § 504 for violations of the
23 copyrights of Plaintiffs and the Class by Defendant.
- 24 d) Permanent injunctive relief, including but not limited to changes to the LLaMA
25 language models to ensure that all applicable information set forth in 17 U.S.C. §
26 1203(b)(1) is included when appropriate.
- 27 e) An order of costs and allowable attorney's fees under 17 U.S.C. § 1203(b)(4)–(5).

- 1 f) An award of statutory damages under 17 U.S.C. § 1203(b)(3) and 17 U.S.C. § 1203(c)(3),
- 2 or in the alternative, an award of actual damages and any additional profits under 17
- 3 U.S.C. § 1203(c)(2) (including tripling damages under 17 U.S.C. § 1203(c)(4) if
- 4 applicable).
- 5 g) Pre- and post-judgment interest on the damages awarded to Plaintiffs and the Class, and
- 6 that such interest be awarded at the highest legal rate from and after the date this class
- 7 action complaint is first served on Defendant.
- 8 h) Defendants are to be jointly and severally responsible financially for the costs and
- 9 expenses of a Court approved notice program through post and media designed to give
- 10 immediate notification to the Class.
- 11 i) Further relief for Plaintiffs and the Class as may be just and proper.

IX. JURY TRIAL DEMANDED

12 Under Federal Rule of Civil Procedure 38(b), Plaintiffs demand a trial by jury of all the claims
13 asserted in this Complaint so triable.
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

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
kcmahon@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*