

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

Bryan L. Clobes (*pro hac vice*)
**CAFFERTY CLOBES MERIWETHER
& SPRENGEL LLP**
205 N. Monroe Street
Media, PA 19063
Telephone: (215) 864-2800
Email: bclobes@caffertyclobes.com

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

12 *Counsel for Individual and Representative Plaintiffs
13 and the Proposed Class (continued on signature page)*

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

17 **IN RE OPENAI CHATGPT LITIGATION**

18 This document relates to:
19 23-cv-03223-AMO
20 23-cv-03416-AMO
21 23-cv-04625-AMO

Master File No. 23-cv-3223-AMO

**FIRST CONSOLIDATED AMENDED
COMPLAINT**

Class Action

Demand for Jury Trial

1 Plaintiffs Paul Tremblay, Sarah Silverman, Christopher Golden, Richard Kadrey, Ta-Nehisi
2 Coates, Junot Díaz, Andrew Sean Greer, David Henry Hwang, Matthew Klam, Laura Lippman, Rachel
3 Louise Snyder, and Jacqueline Woodson (collectively “Plaintiffs”), on behalf of themselves and all
4 others similarly situated, bring this class-action complaint (“Complaint”) against defendants OpenAI,
5 Inc.; OpenAI, L.P.; OpenAI OpCo, L.L.C.; OpenAI GP, L.L.C.; OpenAI Startup Fund I, L.P.; OpenAI
6 Startup Fund GP I, L.L.C.; and OpenAI Startup Fund Management, LLC (collectively, “OpenAI” or
7 “Defendants”). Plaintiffs seek to recover injunctive relief and damages as a result of OpenAI’s
8 unlawful conduct.

10 OVERVIEW

11 1. ChatGPT is a software product created, maintained, and sold by OpenAI.

12 2. ChatGPT is powered by two AI software programs called GPT-3.5 and GPT-4, also
13 known as *large language models*. Rather than being programmed in the traditional way, a large language
14 model is “trained” by copying massive amounts of text and extracting expressive information from it.
15 This body of text is called the *training dataset*. Once a large language model has copied and ingested the
16 text in its training dataset, it is able to emit convincingly naturalistic text outputs in response to user
17 prompts.

18 3. A large language model’s output is therefore entirely and uniquely reliant on the
19 material in its training dataset. Every time it assembles a text output, the model relies on the
20 information it extracted from its training dataset.

21 4. Plaintiffs and Class members are authors of books. Plaintiffs and Class members have
22 registered copyrights in the books they published. Plaintiffs and Class members did not consent to the
23 use of their copyrighted books as training material for ChatGPT. Nonetheless, their copyrighted
24 materials were ingested and used to train ChatGPT.

25 5. Indeed, when ChatGPT is prompted, ChatGPT generates summaries of Plaintiffs’
26 copyrighted works—something only possible if ChatGPT was trained on Plaintiffs’ copyrighted works.

27 6. Defendants, by and through the use of ChatGPT, benefit commercially and profit
28 significantly from the use of Plaintiffs’ and Class members’ copyrighted materials.

1 36. OpenAI made a series of large language models, including without limitation GPT-1
2 (released June 2018), GPT-2 (February 2019), GPT-3 (May 2020), GPT-3.5 (March 2022), and most
3 recently GPT-4 (March 2023). “GPT” is an abbreviation for “generative pre-trained transformer,”
4 where *pre-trained* refers to the use of textual material for training, *generative* refers to the model’s ability
5 to emit text, and *transformer* refers to the underlying training algorithm. OpenAI offers certain language
6 models in variant forms: for instance, the GPT-4 family of models includes publicly accessible variants
7 called ‘gpt-4-0125-preview,’ ‘gpt-4-turbo-preview,’ and ‘gpt-4-32k;’ the GPT-3.5 Turbo family of
8 models includes publicly accessible variants called ‘gpt-3.5-turbo-0125,’ ‘gpt-3.5-turbo-1106,’ and ‘gpt-
9 3.5-turbo-instruct.’ On information and belief, OpenAI has made other language-model variants that
10 are in commercial use but are not publicly accessible. In an interview with the Financial Times in
11 November 2023, OpenAI CEO Sam Altman confirmed that GPT-5 is under development. Together,
12 OpenAI’s large language models, including any in development, will be referred to as the “OpenAI
13 Language Models.”¹

14 37. Many kinds of material have been used to train large language models. Books, however,
15 have always been a key ingredient in training datasets for large language models because books offer the
16 best examples of high-quality longform writing.

17 38. For instance, in its June 2018 paper introducing GPT-1 (called “Improving Language
18 Understanding by Generative Pre-Training”), OpenAI revealed that it trained GPT-1 on BookCorpus,
19 a collection of “over 7,000 unique unpublished books from a variety of genres including Adventure,
20 Fantasy, and Romance.” OpenAI confirmed why a dataset of books was so valuable: “Crucially, it
21 contains long stretches of contiguous text, which allows the generative model to learn to condition on
22 long-range information.” Hundreds of large language models have been trained on BookCorpus,
23 including those made by OpenAI, Google, Amazon, and others.

24 39. BookCorpus, however, is a controversial dataset. It was assembled in 2015 by a team of
25 AI researchers for the purpose of training language models. They copied the books from a website
26 called Smashwords.com that hosts unpublished novels that are available to readers at no cost. Those

27
28 ¹ The definition of “OpenAI Language Models” encompasses any language models developed (or in development) by OpenAI, irrespective of whether those models underly ChatGPT.

1 novels, however, are largely under copyright. They were copied into the BookCorpus dataset without
2 consent, credit, or compensation to the authors.

3 40. OpenAI also copied many books while training GPT-3. In the July 2020 paper
4 introducing GPT-3 (called “Language Models are Few-Shot Learners”), OpenAI disclosed that 15% of
5 the enormous GPT-3 training dataset came from “two internet-based books corpora” that OpenAI
6 simply called “Books1” and “Books2.”

7 41. Tellingly, OpenAI has never revealed what books are part of the Books1 and Books2
8 datasets—though there are some clues. First, OpenAI admitted these are “internet-based books
9 corpora.” Second, both Books1 and Books2 are apparently much larger than BookCorpus. Based on
10 numbers included in OpenAI’s paper about GPT-3, Books1 is apparently about nine times larger;
11 Books2 is about 42 times larger. Since BookCorpus contained about 7,000 titles, this suggests Books1
12 would contain about 63,000 titles; Books2 would contain about 294,000 titles.

13 42. But there are only a handful of “internet-based books corpora” that would be able to
14 deliver this much material.

15 43. As noted in ¶ 41, the OpenAI Books1 dataset can be estimated to contain about 63,000
16 titles. Project Gutenberg is an online archive of e-books whose copyright has expired. In September
17 2020, Project Gutenberg claimed to have “over 60,000” titles. Project Gutenberg has long been
18 popular for training AI systems due to the lack of copyright. In 2018, a team of AI researchers created
19 the “Standardized Project Gutenberg Corpus,” which contained “more than 50,000 books.” On
20 information and belief, the OpenAI Books1 dataset is based on either the Standardized Project
21 Gutenberg Corpus or Project Gutenberg itself, because of the roughly similar sizes of the two datasets.

22 44. As noted in ¶ 41, the OpenAI Books2 dataset can be estimated to contain about 294,000
23 titles. The only “internet-based books corpora” that have ever offered that much material are notorious
24 “shadow library” websites like Library Genesis (aka LibGen), Z-Library (aka B-ok), Sci-Hub, and
25 Bibliotik. The books aggregated by these websites have also been available in bulk via torrent systems.
26 These flagrantly illegal shadow libraries have long been of interest to the AI-training community: for
27 instance, an AI training dataset published in December 2020 by EleutherAI called “Books3” includes a
28 recreation of the Bibliotik collection and contains nearly 200,000 books. On information and belief, the

1 OpenAI Books2 dataset includes books copied from these “shadow libraries,” because those are the
2 largest sources of trainable books most similar in nature and size to OpenAI’s description of Books2.

3 45. In March 2023, OpenAI’s paper introducing GPT-4 contained no information about its
4 dataset at all: OpenAI claimed that “[g]iven both the competitive landscape and the safety implications
5 of large-scale models like GPT-4, this report contains no further details about ... dataset construction.”
6 Later in the paper, OpenAI concedes it did “filter[] our dataset ... to specifically reduce the quantity of
7 inappropriate erotic text content.”

8 9 **INTERROGATING THE OPENAI LANGUAGE MODELS USING CHATGPT**

10 46. ChatGPT is a language model created and sold by OpenAI. As its name suggests,
11 ChatGPT is designed to offer a conversational style of interaction with a user. OpenAI offers ChatGPT
12 through a web interface to individual users for \$20 per month. Through the web interface, users can
13 choose to use two versions of ChatGPT: one based on the GPT-3.5 model, and one based on the newer
14 GPT-4 model.

15 47. OpenAI also offers ChatGPT to software developers through an application-
16 programming interface (or “API”). The API allows developers to write programs that exchange data
17 with ChatGPT. Access to ChatGPT through the API is billed on the basis of usage.

18 48. Regardless of how it is accessed—either through the web interface or through the API—
19 ChatGPT allows users to enter text prompts, which ChatGPT then attempts to respond to in a natural
20 way, i.e., ChatGPT can generate answers in a coherent and fluent way that closely mimics human
21 language. If a user prompts ChatGPT with a question, ChatGPT will answer. If a user prompts
22 ChatGPT with a command, ChatGPT will obey. If a user prompts ChatGPT to summarize a
23 copyrighted book, it will do so.

24 49. ChatGPT’s output, like other LLMs, relies on the data upon which it is trained to
25 generate new content. LLMs generate output based on patterns and connections drawn from the
26 training data. For example, if an LLM is prompted to generate a writing in the style of a certain author,
27 the LLM would generate content based on patterns and connections it learned from analysis of that
28 author’s work within its training data.

- 1 e. all governmental entities; and
- 2 f. the judges and chambers staff in this case, as well as any members of their
- 3 immediate families.

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

8 56. **Typicality.** Plaintiffs' claims are typical of the claims of other members of the Class
9 because Plaintiffs and all members of the Class were damaged by the same wrongful conduct of
10 Defendants as alleged herein, and the relief sought herein is common to all members of the Class.

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

16 58. **Commonality and predominance.** Numerous questions of law or fact common to each
17 Class arise from Defendants' conduct:

- 18 a. whether Defendants violated the copyrights of Plaintiffs and the Class when they
- 19 downloaded copies of Plaintiffs' copyrighted books and used them to train ChatGPT;
- 20 b. whether ChatGPT itself is an infringing derivative work based on Plaintiffs' copyrighted
- 21 books;
- 22 c. Whether Defendants' conduct alleged herein constitutes Unfair Competition under
- 23 California Business and Professions Code § 17200 *et seq.*
- 24 d. Whether this Court should enjoin Defendants from engaging in the unlawful conduct
- 25 alleged herein. And what the scope of that injunction would be.
- 26 e. Whether any affirmative defense excuses Defendants' conduct.
- 27 f. Whether any statutes of limitation constrain the potential recovery for Plaintiffs and the
- 28 Class.

1 copies of Plaintiffs' copyrighted work. Defendants deceptively marketed their product in a manner that
2 fails to attribute the success of their product to the copyrighted work on which it is based.

3
4 **DEMAND FOR JUDGMENT**

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

- 7 a) This Action may proceed as a class action, with Plaintiffs serving as Class
8 Representatives, and with Plaintiffs' counsel as Class Counsel.
- 9 b) A declaration that Defendants have infringed Plaintiffs' and the Class' exclusive
10 copyrights in the Infringed Works under the Copyright Act.
- 11 c) A declaration that such infringement is willful.
- 12 d) Judgment in favor of Plaintiffs and the Class and against Defendants.
- 13 e) An award of statutory and other damages under 17 U.S.C. § 504 for Defendants' willful
14 infringement of Plaintiffs' and the Class' exclusive copyrights in the Infringed Works.
- 15 f) Reasonable attorneys' fees and costs as available under 17 U.S.C. § 505 or other
16 applicable statute.
- 17 g) Pre- and post-judgment interest on the damages awarded to Plaintiffs and the Class, and
18 that such interest be awarded at the highest legal rate from and after the date this class
19 action complaint is first served on Defendants.
- 20 h) Defendants are to be jointly and severally responsible financially for the costs and
21 expenses of a Court-approved notice program through post and media designed to give
22 immediate notification to the Class.
- 23 i) Further relief for Plaintiffs and the Class as may be just and proper.
- 24
25
26
27
28

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 13, 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)
Holden Benon (State Bar No. 325847)
Aaron Cera (State Bar No. 351163)
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
 hbenon@saverilawfirm.com
 acera@saverilawfirm.com

Bryan L. Clobes (*pro hac vice*)
**CAFFERTY CLOBES MERIWETHER
& SPRENGEL LLP**
205 N. Monroe Street
Media, PA 19063
Telephone: (215) 864-2800
Email: bclobes@caffertyclobes.com

Alexander J. Sweatman (*pro hac vice* forthcoming)
**CAFFERTY CLOBES MERIWETHER
& SPRENGEL LLP**
135 South LaSalle Street, Suite 3210
Chicago, IL 60603
Telephone: (312) 782-4880
Email: asweatman@caffertyclobes.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*