1	Joseph R. Saveri (State Bar No. 130064) JOSEPH SAVERI LAW FIRM, LLP		Bryan L. Clobes (pro hac vice) CAFFERTY CLOBES MERIWETHER	
2	601 California Street, Suite 1000		& SPRENGEL LLP	
3	San Francisco	,	205 N. Moni	
3	Telephone:	(415) 500-6800	Media, PA 19	9063
4	Facsimile:	(415) 395-9940	Telephone:	(215) 864-2800
5	Email:	jsaveri@saverilawfirm.com	Email:	bclobes@caffertyclobes.com
6		terick (State Bar No. 250953)		
7	1920 Hillhurst Avenue, #406 Los Angeles, CA 90027			
8	Facsimile:	(323) 968-2632 (415) 395-9940		
9	Email:	mb@buttericklaw.com		
	Lillall.	mowoutteriexiaw.com		
10	Counsel for In	edividual and Representative Plaintiffs		
11		sed Class (continued on signature page)		
	unu me i rope.	sea Suuss (communu on signaiane page)		
12				
13		T COURT		
10		NORTHERN DIST	RICT OF CA	LIFORNIA
14		SAN FRANC	ISCO DIVIS	SION
15	IN RE OPE	NAI CHATGPT LITIGATION	Master File	No. 23-cv-3223-AMO
16	This docume	ent relates to:	FIDST COM	NSOLIDATED AMENDED
17	23-cv-03223-		COMPLAI	
	23-cv-03416-		COMILAI	NI
18	23-cv-04625-		Class Action	n
19			Demand for	· Iurv Trial
20				J y
21				
22				
23				
24				
25				
26				
27				

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

OVERVIEW

- 1. ChatGPT is a software product created, maintained, and sold by OpenAI.
- 2. ChatGPT is powered by two AI software programs called GPT-3.5 and GPT-4, also known as *large language models*. 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*. Once a large language model has copied and ingested the text in its training dataset, it is able to emit convincingly naturalistic text outputs in response to user prompts.
- 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.
- 4. Plaintiffs and Class members are authors of books. Plaintiffs and Class members have registered copyrights in the books they published. Plaintiffs and Class members did not consent to the use of their copyrighted books as training material for ChatGPT. Nonetheless, their copyrighted materials were ingested and used to train ChatGPT.
- 5. Indeed, when ChatGPT is prompted, ChatGPT generates summaries of Plaintiffs' copyrighted works—something only possible if ChatGPT was trained on Plaintiffs' copyrighted works.
- 6. Defendants, by and through the use of ChatGPT, benefit commercially and profit significantly from the use of Plaintiffs' and Class members' copyrighted materials.

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. § 101, et seq.).
- 8. Jurisdiction and venue is proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Defendant OpenAI, Inc. is headquartered in this District, and thus a substantial part of the events giving rise to the claims 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. Each 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. 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 Paul Tremblay is a writer who lives in Massachusetts and owns registered copyrights in multiple works, including *The Cabin at the End of the World*.
- 11. Plaintiff Sarah Silverman is a writer and performer who lives in California and owns a registered copyright in one work, *The Bedwetter*.
- 12. Plaintiff Christopher Golden is a writer who lives in Massachusetts and owns registered copyrights in multiple works, including *Ararat*.
- 13. Plaintiff Richard Kadrey is a writer who lives in Pennsylvania and owns registered copyrights in multiple works, including *Sandman Slim*.
- 14. Plaintiff Ta-Nehisi Coates is an author who lives in New York and owns registered copyrights in multiple works, including *The Water Dancer*.

L.L.C. is a wholly owned subsidiary of OpenAI, Inc. that is operated for profit. OpenAI, Inc. controls

OpenAI OpCo, L.L.C. directly and through the other OpenAI entities.

27

28

26.

with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI GP is the general partner of OpenAI, L.P. OpenAI GP manages and operates the day-to-day business and affairs of OpenAI, L.P. OpenAI GP was aware of the unlawful conduct alleged herein and exercised control over OpenAI, L.P. throughout the Class Period. OpenAI, Inc. directly controls OpenAI GP.

27. Defendant OpenAI Startup Fund I, L.P. ("OpenAI Startup Fund I") is a Delaware

Defendant OpenAI GP, L.L.C. ("OpenAI GP") is a Delaware limited liability company

- 27. Defendant OpenAI Startup Fund I, L.P. ("OpenAI Startup Fund I") is a Delaware limited partnership with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI Startup Fund I was instrumental in the foundation of OpenAI, L.P., including the creation of its business strategy and providing initial funding. OpenAI Startup Fund I was aware of the unlawful conduct alleged herein and exercised control over OpenAI, L.P. throughout the Class Period.
- 28. Defendant OpenAI Startup Fund GP I, L.L.C. ("OpenAI Startup Fund GP I") is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI Startup Fund GP I is the general partner of OpenAI Startup Fund I. OpenAI Startup Fund GP I is a party to the unlawful conduct alleged herein. OpenAI Startup Fund GP I manages and operates the day-to-day business and affairs of OpenAI Startup Fund I.
- 29. Defendant OpenAI Startup Fund Management, LLC ("OpenAI Startup Fund Management") is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI Startup Fund Management is a party to the unlawful conduct alleged herein. OpenAI Startup Fund Management was aware of the unlawful conduct alleged herein and exercised control over OpenAI, L.P. throughout the Class Period.

AGENTS AND CO-CONSPIRATORS

30. The unlawful acts alleged against the 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.

31. Various persons and/or firms not named as Defendants may have participated as coconspirators in the violations alleged herein and may have performed acts and made statements in furtherance thereof. Each acted as the principal, agent, or joint venturer of, or for, other Defendants with respect to the acts, violations, and common course of conduct alleged herein.

FACTUAL ALLEGATIONS

- 32. OpenAI creates and sells artificial-intelligence software products. *Artificial intelligence* is commonly abbreviated as "AI." AI software is designed to algorithmically simulate human reasoning or inference, often using statistical methods.
- 33. Certain AI products created and sold by OpenAI are known as *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.
- 34. Much of the material in OpenAI's training datasets, however, comes from copyrighted works—including books written by Plaintiffs—that were copied by OpenAI without consent, without credit, and without compensation.
- 35. 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. This information is most commonly found on the back of the book's title page and is customarily included in all books, regardless of genre.

- (released June 2018), GPT-2 (February 2019), GPT-3 (May 2020), GPT-3.5 (March 2022), and most recently GPT-4 (March 2023). "GPT" is an abbreviation for "generative pre-trained transformer," where *pre-trained* refers to the use of textual material for training, *generative* refers to the model's ability to emit text, and *transformer* refers to the underlying training algorithm. OpenAI offers certain language models in variant forms: for instance, the GPT-4 family of models includes publicly accessible variants called 'gpt-4-0125-preview,' 'gpt-4-turbo-preview,' and 'gpt-4-32k;' the GPT-3.5 Turbo family of models includes publicly accessible variants called 'gpt-3.5-turbo-0125,' 'gpt-3.5-turbo-1106,' and 'gpt-3.5-turbo-instruct.' On information and belief, OpenAI has made other language-model variants that are in commercial use but are not publicly accessible. In an interview with the Financial Times in November 2023, OpenAI CEO Sam Altman confirmed that GPT-5 is under development. Together, OpenAI's large language models, including any in development, will be referred to as the "OpenAI Language Models." ¹
- 37. Many kinds of material have been used to train large language models. Books, however, have always been a key ingredient in training datasets for large language models because books offer the best examples of high-quality longform writing.
- 38. For instance, in its June 2018 paper introducing GPT-1 (called "Improving Language Understanding by Generative Pre-Training"), OpenAI revealed that it trained GPT-1 on BookCorpus, a collection of "over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance." OpenAI confirmed why a dataset of books was so valuable: "Crucially, it contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information." Hundreds of large language models have been trained on BookCorpus, including those made by OpenAI, Google, Amazon, and others.
- 39. BookCorpus, however, is a controversial dataset. It was assembled in 2015 by a team of AI researchers for the purpose of training language models. They copied the books from a website called Smashwords.com that hosts unpublished novels that are available to readers at no cost. Those

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

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

- 40. OpenAI also copied many books while training GPT-3. In the July 2020 paper introducing GPT-3 (called "Language Models are Few-Shot Learners"), OpenAI disclosed that 15% of the enormous GPT-3 training dataset came from "two internet-based books corpora" that OpenAI simply called "Books1" and "Books2."
- 41. Tellingly, OpenAI has never revealed what books are part of the Books1 and Books2 datasets—though there are some clues. First, OpenAI admitted these are "internet-based books corpora." Second, both Books1 and Books2 are apparently much larger than BookCorpus. Based on numbers included in OpenAI's paper about GPT-3, Books1 is apparently about nine times larger; Books2 is about 42 times larger. Since BookCorpus contained about 7,000 titles, this suggests Books1 would contain about 63,000 titles; Books2 would contain about 294,000 titles.
- 42. But there are only a handful of "internet-based books corpora" that would be able to deliver this much material.
- 43. As noted in ¶ 41, the OpenAI Books1 dataset can be estimated to contain about 63,000 titles. Project Gutenberg is an online archive of e-books whose copyright has expired. In September 2020, Project Gutenberg claimed to have "over 60,000" titles. Project Gutenberg has long been popular for training AI systems due to the lack of copyright. In 2018, a team of AI researchers created the "Standardized Project Gutenberg Corpus," which contained "more than 50,000 books." On information and belief, the OpenAI Books1 dataset is based on either the Standardized Project Gutenberg Corpus or Project Gutenberg itself, because of the roughly similar sizes of the two datasets.
- 44. As noted in ¶ 41, the OpenAI Books2 dataset can be estimated to contain about 294,000 titles. The only "internet-based books corpora" that have ever offered that much material are notorious "shadow library" websites like Library Genesis (aka LibGen), Z-Library (aka B-ok), Sci-Hub, and Bibliotik. The books aggregated by these websites have also been available in bulk via torrent systems. These flagrantly illegal shadow libraries have long been of interest to the AI-training community: for instance, an AI training dataset published in December 2020 by EleutherAI called "Books3" includes a recreation of the Bibliotik collection and contains nearly 200,000 books. On information and belief, the

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

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

INTERROGATING THE OPENAI LANGUAGE MODELS USING CHATGPT

- 46. ChatGPT is a language model created and sold by OpenAI. As its name suggests, ChatGPT is designed to offer a conversational style of interaction with a user. OpenAI offers ChatGPT through a web interface to individual users for \$20 per month. Through the web interface, users can choose to use two versions of ChatGPT: one based on the GPT-3.5 model, and one based on the newer GPT-4 model.
- 47. OpenAI also offers ChatGPT to software developers through an application-programming interface (or "API"). The API allows developers to write programs that exchange data with ChatGPT. Access to ChatGPT through the API is billed on the basis of usage.
- A8. Regardless of how it is accessed—either through the web interface or through the API—ChatGPT allows users to enter text prompts, which ChatGPT then attempts to respond to in a natural way, i.e., ChatGPT can generate answers in a coherent and fluent way that closely mimics human language. If a user prompts ChatGPT with a question, ChatGPT will answer. If a user prompts ChatGPT with a command, ChatGPT will obey. If a user prompts ChatGPT to summarize a copyrighted book, it will do so.
- 49. ChatGPT's output, like other LLMs, relies on the data upon which it is trained to generate new content. LLMs generate output based on patterns and connections drawn from the training data. For example, if an LLM is prompted to generate a writing in the style of a certain author, the LLM would generate content based on patterns and connections it learned from analysis of that author's work within its training data.

- 50. On information and belief, the reason ChatGPT can accurately summarize a certain copyrighted book is because that book was copied by OpenAI and ingested by the underlying OpenAI Language Model (either GPT-3.5 or GPT-4) as part of its training data.
- 51. When ChatGPT was prompted to summarize books written by each of the Plaintiffs, it generated very accurate summaries. These summaries are attached as **Exhibit B**. The summaries get some details wrong, which is expected, since a large language model mixes together expressive material derived from many sources. Still, the rest of the summaries are accurate, which means that ChatGPT retains knowledge of particular works in the training dataset and is able to output similar textual content.

CLASS ALLEGATIONS

- 52. The "Class Period" as defined in this Complaint begins on at least June 28, 2020 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 June 28, 2020, Plaintiffs reserve the right to amend the Class Period to comport with the facts and evidence uncovered during further investigation or through discovery.
- 53. 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 OpenAI Language Models during the Class Period.

- 54. This Class definition excludes:
 - a. any of the 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;

Works in Exhibit A.

27

28

- 67. Licensing copyrighted material to train AI models is plainly feasible. It already happens. Indeed, OpenAI itself sought to license copyrighted materials to train its LLMs. For instance, OpenAI reached agreements with the Associated Press and Axel Springer to license textual materials for the purpose of LLM training. OpenAI has reportedly been in negotiations with other publishers as well.
- 68. Because the OpenAI Language Models cannot function without the expressive information extracted from Plaintiffs' works (and others) and retained inside them, the OpenAI Language Models are themselves infringing derivative works, made without Plaintiffs' permission and in violation of their exclusive rights under the Copyright Act.
- 69. Plaintiffs have been injured by OpenAI'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

UCL — Unfair Competition

CAL. BUS. & PROF. CODE §§ 17200 ET SEQ.

- 70. Plaintiffs incorporate by reference the preceding factual allegations.
- 71. Defendants engaged in unfair business practices by, among other things, using Plaintiffs' Infringed Works to train ChatGPT without permission from Plaintiffs or Class members.
- 72. The unfair business practices described herein violate California Business and Professions Code § 17200 *et seq.* (the "UCL").
- 73. The unfair business practices described herein violate the UCL because they are unfair, immoral, unethical, oppressive, unscrupulous, or injurious to consumers, and because Defendants used Plaintiffs' protected works to train ChatGPT for Defendants' own commercial profit without the authorization of Plaintiffs or the Class. Defendants unfairly profit from and take credit for developing a commercial product based on unattributed reproductions of those stolen writings and ideas.
- 74. The unlawful business practices described herein violate the UCL because consumers are likely to be deceived. Defendants knowingly and secretively trained ChatGPT using unauthorized

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	

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

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) A declaration that Defendants have infringed Plaintiffs' and the Class' exclusive copyrights in the Infringed Works under the Copyright Act.
- c) A declaration that such infringement is willful.
- d) Judgment in favor of Plaintiffs and the Class and against Defendants.
- e) An award of statutory and other damages under 17 U.S.C. § 504 for Defendants' willful infringement of Plaintiffs' and the Class' exclusive copyrights in the Infringed Works.
- f) Reasonable attorneys' fees and costs as available under 17 U.S.C. § 505 or other applicable statute.
- 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 Defendants.
- 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.

1	JURY TRIAL DEMANDED				
2	Under Federal Rule of Civil Procedure 38(b), Plaintiffs demand a trial by jury of all the claims				
3	asserted in this Complaint so triable.				
4					
5	Dated: March 13, 2024	By: /s/Joseph R. Saveri			
6		Joseph R. Saveri			
7	Joseph R. Saveri (State Bar No. 130064)	Bryan L. Clobes (pro hac vice) CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP 205 N. Monroe Street			
8	Cadio Zirpoli (State Bar No. 179108)				
9	Christopher K.L. Young (State Bar No. 318371) Holden Benon (State Bar No. 325847)				
9	Aaron Cera (State Bar No. 351163)	Media, PA 19063			
10	JOSEPH SAVERI LAW FIRM, LLP	Telephone: (215) 864-2800			
11	601 California Street, Suite 1000	Email: bclobes@caffertyclobes.com			
12	San Francisco, California 94108 Telephone: (415) 500-6800	Alexander J. Sweatman (pro hac vice forthcoming)			
12	Facsimile: (415) 395-9940	CAFFERTY CLOBES MERIWETHER			
13	Email: jsaveri@saverilawfirm.com	& SPRENGEL LLP			
14	czirpoli@saverilawfirm.com	135 South LaSalle Street, Suite 3210			
	cyoung@saverilawfirm.com	Chicago, IL 60603			
15	hbenon@saverilawfirm.com	Telephone: (312) 782-4880			
16	acera@saverilawfirm.com	Email: asweatman@caffertyclobes.com			
17	Matthew Butterick (State Bar No. 250953)	Counsel for Individual and Representative Plaintiffs and the Proposed Class			
18	1920 Hillhurst Avenue, #406 Los Angeles, CA 90027				
19	Telephone: (323) 968-2632				
20	Facsimile: (415) 395-9940 Email: mb@buttericklaw.com				
21	Linan. mowoutenesiaw.com				
22					
23					
24					
25					
26					
27					
28					