BURSOR & FISHER, P.A.
L. Timothy Fisher (State Bar No. 191626)
Joshua B. Glatt (State Bar No. 354064)
1990 North California Blvd., 9th Floor
Walnut Creek, CA 94596
Telephone: (925) 300-4455
Facsimile: (925) 407-2700
E-mail: ltfisher@bursor.com
jglatt@bursor.com

BURSOR & FISHER, P.A.
Joseph I. Marchese (*pro hac vice* forthcoming)
Julian C. Diamond (*pro hac vice* forthcoming)
1330 Avenue of the Americas, 32nd Floor
New York, NY 10019
Telephone: (646) 837-7150
Facsimile: (212) 989-9163
E-Mail: jmarchese@bursor.com
jdiamond@bursor.com

Attorneys for Plaintiff

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

DAVID MILLETTE, individually and on behalf of all others similarly situated,

Plaintiff,

v.

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

Defendants.

Case No.

CLASS ACTION COMPLAINT

JURY TRIAL DEMANDED

Plaintiff David Millette, (hereinafter "Plaintiff"), brings this action on behalf of himself and all others similarly situated 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"). Plaintiff seeks to recover injunctive relief and damages as a result of Defendants' unlawful conduct. Plaintiff makes the following allegations pursuant to the investigation of his counsel and are based upon information and belief, except as to the allegations specifically pertaining to himself, which are based on personal knowledge.

NATURE OF THE CASE

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

2. ChatGPT is currently powered by artificial-intelligence (hereinafter "AI") software programs called GPT-3.5, GPT-4, and GPT-4o also known as *large language models*. 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. Large language models' output is therefore entirely and uniquely reliant on the material in their training dataset. Every time they assemble text, video or image outputs, the models rely on the information they extracted from their training dataset.

4. This case addresses the surreptitious, non-consensual transcription of millions of YouTube users' videos by Defendants to train Defendants' AI software products. For years, YouTube has been a popular video sharing platform that allows content creators and users to upload and share videos with audiences worldwide. However, unbeknownst to those who upload videos to YouTube, Defendants have been covertly transcribing YouTube videos to create training datasets that they then use to train their AI products.

5. Plaintiff and Class members are YouTube users and video creators. Plaintiff and Class members have retained ownership rights in their uploaded videos, per YouTube's Terms of

Service. Plaintiff and Class members did not consent to the use of their videos as training material for ChatGPT. Nonetheless, their materials were transcribed and used to train ChatGPT.

6. By transcribing and using these videos in this way, Defendants profit from Plaintiff's and class members' data time and time again. As Defendants' AI products become more sophisticated through the use of training datasets, they become more valuable to prospective and current users, who purchase subscriptions to access Defendants' AI products.

7. By collecting and using this data without consent, Defendants have profited significantly from the use of Plaintiff's and Class members' materials, violated California's Unfair Competition Law ("UCL"), and been unjustly enriched at Plaintiff and Class members' expense.

JURISDICTION AND VENUE

8. This Court has subject matter Jurisdiction over this action pursuant to the Class Action Fairness Act ("CAFA"), 28 U.S.C. § 1332(d)(2) because this is a class action in which at least one member of the class is a citizen of a state different from any Defendants, the amount in controversy exceeds $5 million, exclusive of interest and costs, and the proposed class contains more than 100 members.

9. This Court has personal jurisdiction over the Defendants because Defendants maintain their principal places of business in this District and because a substantial part of the events or omissions giving rise to the claims asserted herein occurred in this District.

10. Venue is proper in this district pursuant to 28 U.S.C. § 1391 because a substantial part of the events or omissions giving rise to the claims asserted herein occurred in this District and because Defendants maintain their principal places of business in this District.

PARTIES

11. Plaintiff David Millette is a resident of Douglas, Massachusetts. Plaintiff created a YouTube account in or around June 2009. During that entire time, Plaintiff has retained ownership rights to the video content he has uploaded to YouTube, per YouTube's Terms of Service.

12. Defendants transcribed Plaintiff's videos to train their AI software products.

13. Defendant OpenAI, Inc. is a Delaware nonprofit corporation with its principal place of business located at 3180 18th Street, San Francisco, CA 94110.

14. Defendant OpenAI, L.P. is a Delaware limited partnership with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI, L.P. is a wholly owned subsidiary of OpenAI Inc. that is operated for profit. OpenAI, Inc. controls OpenAI, L.P. directly and through the other OpenAI entities.

15. Defendant OpenAI OpCo, L.L.C. is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI OpCo, 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.

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

17. 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.

18. 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.

19. 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.

20. Each of the Defendants acted jointly to perpetrate the acts described herein. At all times relevant to the allegations in this matter, each of these Defendants acted in concert with, with the knowledge and approval of, and/or as the agent of the other Defendants within the course and scope of the agency, regarding the acts and omissions alleged.

GENERAL BACKGROUND

21. OpenAI creates and sells artificial-intelligence (AI) software products. AI software is designed to algorithmically simulate human reasoning or inference, often using statistical methods.

22. Certain AI products created and sold by OpenAI are known as large language models. Large language models ("LLMs") are types of AI software designed to parse and emit natural language. Though LLMs are software programs, they are not created the way most software programs are—that is, by human software engineers writing code. Rather, LLMs are "trained" by copying massive amounts of text from various sources and feeding these copies into the model. During training, these models copy each piece of information in the training dataset and extract expressive information from it. The LLMs progressively adjust their output to more closely resemble the images, videos, and sequences of words copied from the training dataset. Once these models have copied and ingested all these inputs, they are able to emit convincing simulations of natural written language, as well as videos and images as they appear in the training dataset.

23. Much of the material in OpenAI's training datasets, however, comes from works—including videos created and uploaded by Plaintiff—that were copied by OpenAI without consent, without credit, and without compensation.

24. OpenAI made a series of large language models, including without limitation GPT-1 (released June 2018), GPT-2 (February 2019), GPT-3 (May 2020), GPT-3.5 (March 2022), GPT-4 (March 2023) and GPT-4o (May 2024). "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]

25. Many kinds of material have been used to train large language models. Video transcriptions, however, are a key ingredient in training datasets for large language models because they offer copious examples of natural language.

26. In 2022, OpenAI released an automatic speech recognition (ASR) system called Whisper. The Whisper model, which transcribes audio into text, was trained on 680,000 hours of data collected from across the web. Tellingly, the exact names of the speech recognition corpora on which Whisper was trained are unavailable. But one of the world's largest open multilingual speech corpora, VoxPopuli, contains only 400,000 hours of unlabeled speech data. Libriheavy, an ASR corpus considered one of the largest freely available corpora of speech with supervisions, only consists of 50,000 hours of English speech derived from LibriVox.

27. There are only a handful of publicly available, internet-based speech corpora that can be utilized as training data for LLMs. As demonstrated, the two biggest corpora combined (VoxPopuli and Libriheavy) still fall more than 200,000 hours short of the duration of speech that comprises Whisper's training dataset.

28. A New York Times report claims that Whisper is capable of transcribing the audio from YouTube videos, and that an OpenAI team that included OpenAI's president, Greg Brockman, transcribed more than one million hours of video from YouTube.

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

29. On information and belief, OpenAI's Language Models' datasets include transcriptions of videos taken directly from YouTube, because these video transcriptions are one of the largest corpora of natural language data available for training and fine-tuning the OpenAI Language Models.

CLASS ALLEGATIONS

30. Plaintiff seeks to represent a class defined as all persons or entities domiciled in the United States that uploaded any YouTube video that was transcribed and then used as training data for the OpenAI Language Models without their consent (the "Nationwide Class").

31. Plaintiff also seeks to represent a class defined as all persons or entities domiciled in California that uploaded any YouTube video that was transcribed and then used as training data for the OpenAI Language Models without their consent (the "California Subclass") (collectively with the Nationwide Class, the "Class").

32. Specifically excluded from the Class are Defendants, Defendants' officers, directors, agents, trustees, parents, children, corporations, trusts, representatives, employees, principals, servants, partners, joint ventures, or entities controlled by Defendants, and their heirs, successors, assigns, or other persons or entities related to or affiliated with Defendants and/or Defendants officers and/or directors, the judge assigned to this action, and any member of the judge's immediate family.

33. Plaintiff reserves the right to expand, limit, modify, or amend the class definition, including the addition of one or more subclasses, in connection with his motion for class certification, or at any other time, based on, inter alia, changing circumstances and/or new facts obtained.

34. **Numerosity**. On information and belief, hundreds of thousands of video creators fall into the definitions of the Class. Members of the Class can be identified through Defendants' records, discovery, and other third-party sources.

35. **Commonality and Predominance**. Common questions of law and fact exist as to all members of the Class and predominate over any questions affecting only individual members of the Class. These common legal and factual questions include, but are not limited to, the following:

a. Whether Defendants violated the rights of Plaintiff and the Class when they transcribed Plaintiff's videos and used those transcriptions as part of their AI software's training datasets;

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

c. Whether this Court should enjoin Defendants from engaging in the unlawful conduct alleged herein, and what the scope of that injunction would be.

d. Whether any affirmative defense excuses Defendants' conduct.

e. Whether any statutes of limitation constrain the potential recovery for Plaintiff and the Class.

f. Whether Plaintiff and the other Class members are entitled to restitution or other relief.

36. **Typicality**. Plaintiff's claims are typical of the claims of the other members of the Class in that, among other things, all Class members were similarly situated and were comparably injured through Defendants' wrongful conduct as set forth herein. Further, there are no defenses available to Defendants that are unique to Plaintiff.

37. **Adequacy of Representation**. Plaintiff will fairly and adequately protect the interests of the Class. Plaintiff has retained counsel that is highly experienced in complex consumer class action litigation, and Plaintiff intends to vigorously prosecute this action on behalf of the Class. Furthermore, Plaintiff has no interests that are antagonistic to those of the Class.

38. **Superiority**. A class action is superior to all other available means for the fair and efficient adjudication of this controversy. The damages or other financial detriment suffered by individual Class members are relatively small compared to the burden and expense of individual litigation of their claims against Defendants. It would thus be virtually impossible for the Class to obtain effective redress for the wrongs committed against the members on an individual basis. Furthermore, even if Class members could afford such individualized litigation, the court system could not. Individualized litigation would create the danger of inconsistent or contradictory judgments arising from the same set of facts. Individualized litigation would also increase the

delay and expense to all parties and the court system from the issues raised by this action. By contrast, the class action device provides the benefits of adjudication of these issues in a single proceeding, economies of scale, and comprehensive supervision by a single court, and presents no unusual management difficulties under the circumstances.

39. Further, Defendants have acted and refused to act on grounds generally applicable to the proposed Class, thereby making appropriate final injunctive and declaratory relief with respect to the Class as a whole.

CAUSES OF ACTION

COUNT I
Unjust Enrichment or Restitution
(On behalf of Plaintiff and the Class)

40. Plaintiff incorporates by reference and re-alleges each and every allegation set forth above as though fully set forth herein.

41. Plaintiff brings this claim individually and on behalf of members of the Class against the Defendants.

42. To the extent required by law, Plaintiff brings this claim in the alternative to any legal claims that may be alleged.

43. Plaintiff also alternatively alleges this claim as a Quasi-Contract or Non-Quasi-Contract Claim for Restitution and Disgorgement.

44. Plaintiff and Class members unwittingly conferred a benefit upon Defendants. OpenAI acquired valuable information from Plaintiff and Class members' videos to expand their AI software's training datasets and used that information to develop and improve their products. In using Plaintiff's information to refine their Language Models, OpenAI made their products more valuable to prospective and current users, who purchase subscriptions to access them. Plaintiff and Class members received nothing from this transaction. Plaintiff lacks an adequate remedy at law, and pleads this cause of action in the alternative to the extent Plaintiff is required to do so.

45. Defendants have knowledge of such benefits.

46. Defendants have been unjustly enriched in retaining the revenues derived from the sales of their products trained on Plaintiff's and Class members' videos. Retention of those moneys under these circumstances is unjust and inequitable because Defendants did not obtain the meaningful consent of Plaintiff and Class members before using their videos as described above.

47. Because Defendants' retention of the non-gratuitous benefits conferred on it by Plaintiff and Class members is unjust and inequitable, Defendants must pay restitution to Plaintiff and the Class members for their unjust enrichment, as ordered by the Court.

48. Plaintiff and the members of the Class lack an adequate remedy at law to address the unfair conduct at issue here. Legal remedies available to Plaintiff and class members are inadequate because they are not equally prompt and certain and in other ways efficient as equitable relief. Damages are not equally certain as restitution because the standard that governs restitution is different than the standard that governs damages. Hence, the Court may award restitution even if it determines that Plaintiff fails to sufficiently adduce evidence to support an award of damages. Damages and restitution are not the same amount. Unlike damages, restitution is not limited to the amount of money a defendant wrongfully acquired plus the legal rate of interest. Equitable relief, including restitution, entitles the plaintiff to recover all profits from the wrongdoing, even where the original funds taken have grown far greater than the legal rate of interest would recognize. Legal claims for damages are not equally certain as restitution because claims for restitution entail few elements. In short, significant differences in proof and certainty establish that any potential legal claim cannot serve as an adequate remedy at law.

49. Plaintiff and members of the putative class seek non-restitutionary disgorgement of the financial profits that Defendants obtained as a result of their unjust conduct.

COUNT II
UCL – Unfair Competition
Cal. Bus. & Prof. Code §§ 17200 *et seq.*
(On behalf of Plaintiff and the Class)

50. Plaintiff incorporates by reference and re-alleges each and every allegation set forth above as though fully set forth herein.

51. Defendants engaged in unfair business practices by, among other things, using Plaintiff's videos to train their Language Models without permission from Plaintiff or Class members.

52. The unfair business practices described herein violate California Business and Professions Code § 17200 *et seq*. (the "UCL").

53. 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 Plaintiff's videos to train their Language Models for Defendants' own commercial profit without the authorization of Plaintiff or the Class. Defendants unfairly profit from and take credit for developing a commercial product based on unattributed reproductions of those stolen videos and ideas.

54. The unlawful business practices described herein violate the UCL because consumers are likely to be deceived. Defendants knowingly and secretively trained ChatGPT using unauthorized transcriptions and copies of Plaintiff's videos. Defendants deceptively marketed their product in a manner that fails to attribute the success of their product to the work on which it is based.

PRAYER FOR RELIEF

WHEREFORE, Plaintiff, individually and on behalf of all others similarly situated, seeks judgment against Defendants, as follows:

a. For an order certifying the Classes under Rule 23 of the Federal Rules of Civil Procedure and naming Plaintiff as the representative for the Class and Plaintiff's attorneys as Class Counsel;

b. For an order declaring that Defendants' conduct violates the laws referenced herein;

c. For an order finding in favor of Plaintiff and the Class on all counts asserted herein;

d. For compensatory, statutory, and punitive damages in amounts to be determined by the Court and/or jury;

e. For prejudgment interest on all amounts awarded;

f. For an order of restitution and all other forms of equitable monetary relief;

g. For injunctive relief as the Court may deem proper; and

h. For an order awarding Plaintiff and the Class their reasonable attorneys' fees and expenses and costs of suit.

DEMAND FOR TRIAL BY JURY

Pursuant to Federal Rule of Civil Procedure 38(b), Plaintiff demands a trial by jury of any and all issues in this action so triable of right.

Dated: August 2, 2024

BURSOR & FISHER, P.A.

By: /s/ *L. Timothy Fisher*
L. Timothy Fisher

L Timothy Fisher (State Bar No. 191626)
Joshua B. Glatt (State Bar No. 354064)
1990 North California Blvd., 9th Floor
Walnut Creek, CA 94596
Telephone: (925) 300-4455
Facsimile: (925) 407-2700
E-mail: ltfisher@bursor.com
jglatt@bursor.com

BURSOR & FISHER, P.A.
Joseph I. Marchese (*pro hac vice* forthcoming)
Julian C. Diamond (*pro hac vice* forthcoming)
1330 Avenue of the Americas, 32nd Floor
New York, NY 10019
Telephone: (646) 837-7150
Facsimile: (212) 989-9163
E-Mail: jmarchese@bursor.com
jdiamond@bursor.com

Attorneys for Plaintiff