

DANIEL J. MULLER, SBN 193396
dmuller@venturahersey.com
VENTURA HERSEY & MULLER, LLP
1506 Hamilton Avenue
San Jose, California 95125
Telephone: (408) 512-3022
Facsimile: (408) 512-3023

Attorneys for Plaintiff and the Class

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA – SAN FRANCISCO DIVISION

MICHAEL CHABON, DAVID HENRY
HWANG, MATTHEW KLAM, RACHEL
LOUISE SNYDER, AND AYELET
WALDMAN,

individually and on behalf of all others
similarly situated,

Plaintiffs,

v.

OPENAI, INC., OPENAI, L.P., OPENAI
OPCO, LLC, OPENAI GP LLC, OPENAI
STARTUP FUND GP I, LLC, OPENAI
STARTUP FUND I, LP, and OPENAI
STARTUP FUND MANAGEMENT, LLC,

Defendants.

Case No.

CLASS ACTION COMPLAINT

CLASS ACTION

JURY TRIAL DEMANDED

1 Plaintiffs Michael Chabon, David Henry Hwang, Matthew Klam, Rachel Louise Snyder,
2 and Ayelet Waldman (“Plaintiffs”), individually and on behalf of all others similarly situated,
3 bring this action against Defendants OpenAI, Inc., OpenAI, LP, OpenAI OpCo, LLC, OpenAI
4 GP LLC, OpenAI Startup Fund I, LP, OpenAI Startup Fund GP I, LLC, and OpenAI Startup
5 Fund Management, LLC (collectively, “Defendants” or “OpenAI”). Plaintiffs allege as follows
6 based upon personal knowledge as to themselves and their own acts, and upon information and
7 belief as to all other matters:

8 **NATURE OF ACTION**

9 1. This is a class action lawsuit brought by Plaintiffs on behalf of themselves and a
10 Class of authors holding copyrights in their published works arising from OpenAI’s clear
11 infringement of their intellectual property.

12 2. OpenAI is a research company specializing in the development of artificial
13 intelligence (“AI”) products, such as ChatGPT.

14 3. ChatGPT is an AI chatbot, which produces responses to users’ text queries or
15 prompts in a way that mimics human conversation.

16 4. ChatGPT relies on other OpenAI products to function, namely Generative Pre-
17 trained Transformer (“GPT”) models. “Generative,” in GPT, represents the model’s ability to
18 respond to text inquiries, while “Pre-trained” refers to the model’s use of training datasets to
19 program its responses, and “Transformer” concerns the model’s underlying algorithm allowing
20 it to function.

21 5. OpenAI has released five versions of GPT models, and the current version of
22 ChatGPT runs on GPT-3.5 and GPT-4, depending on whether the user has subscribed to the
23 premium version of ChatGPT. Only the version of ChatGPT that runs on GPT-3.5 is available
24 at no cost to the public.

25 6. OpenAI’s GPT models are types of “large language model,” which is a form of
26 deep-learning algorithm programmed through “training datasets,” consisting of massive
27 amounts of text data copied from the internet by OpenAI. The GPT models extract information
28 from their training datasets in order to learn the statistical relationships between words, phrases,

1 and sentences, which allow them to generate coherent and contextually relevant responses to
2 user prompts or queries.

3 7. A large language model's responses to user prompts or queries are entirely and
4 uniquely dependent on the text contained in its training dataset, necessarily processing and
5 analyzing the information contained in its training dataset to generate responses.

6 8. OpenAI incorporated Plaintiffs' and Class members' copyrighted works in
7 datasets used to train its GPT models powering its ChatGPT product. Indeed, when ChatGPT is
8 prompted, it generates not only summaries, but in-depth analyses of the themes present in
9 Plaintiffs' copyrighted works, which is only possible if the underlying GPT model was trained
10 using Plaintiffs' works.

11 9. Plaintiffs and Class members did not consent to the use of their copyrighted
12 works as training material for GPT models or for use with ChatGPT.

13 10. Defendants, by and through their operation of ChatGPT, benefit commercially
14 and profit handsomely from their unauthorized and illegal use of Plaintiffs' and Class members'
15 copyrighted works.

16 **JURISDICTION AND VENUE**

17 11. This Court has subject matter jurisdiction of this action pursuant to 28 U.S.C. §
18 1331 because this case arises under the Copyright Act (17 U.S.C. § 501) and the Digital
19 Millennium Copyright Act (17 U.S.C. § 1202).

20 12. This Court has personal jurisdiction over Defendants pursuant to 18 U.S.C.
21 §§ 1965(b) & (d), because they maintain their principal places of business in, and are thus
22 residents of, this judicial district, maintain minimum contacts with the United States, this judicial
23 district, and this State, and they intentionally avail themselves of the laws of the United States
24 and this state by conducting a substantial amount of business in California. For these same
25 reasons, venue properly lies in this District pursuant to 28 U.S.C. §§ 1391(a), (b) and (c).

PARTIES

A. Plaintiffs

13. Plaintiff Michael Chabon (“Plaintiff Chabon”) is a resident of California. Plaintiff Chabon is an author who owns registered copyrights in many works, including but not limited to, *The Mysteries of Pittsburgh*, *Wonder Boys*, *The Amazing Adventures of Kavalier & Clay*, *the Yiddish Policemen’s Union*, *Gentlemen of the Road*, *Telegraph Avenue*, *Fight of the Century*, *Kingdom of Olive and Ash*, and *Moonglow*. Plaintiff Chabon is the recipient of the Pulitzer Prize for Fiction, Hugo, Nebula, Los Angeles Times Book Prize, and the National Jewish Book Award, among many others achieved over the span of a writing career spanning more than 30 years. Plaintiff Chabon’s works include copyright-management information that provides information about the copyrighted work, including the title of the work, its ISBN or copyright registration number, the name of the author, and the year of publication.

14. Plaintiff David Henry Hwang (“Plaintiff Hwang”) is a resident of New York. Plaintiff Hwang is a playwright and screenwriter who owns registered copyrights in many works, including but not limited to, *M. Butterfly*, *Chinglish*, *Yellow Face*, *the Dance and the Railroad*, and *FOB*, as well as the Broadway musical, *Flower Drum Song* (2002 revival). Plaintiff Hwang is a Tony Award winner and three-time nominee, a Grammy Award winner who has been twice nominated, a three-time OBIE Award winner, and a three-time finalist for the Pulitzer Prize in Drama. Plaintiff Hwang’s works include copyright-management information that provides information about the copyrighted work, including the title of the work, its ISBN or copyright registration number, the name of the author, and the year of publication.

15. Plaintiff Matthew Klam (“Plaintiff Klam”) is a resident of Washington D.C. Plaintiff Klam is an author who owns registered copyrights in several works, including but not limited to, *Who is Rich?*, and *Sam the Cat and Other Stories*. Plaintiff Klam is a recipient of a Guggenheim Fellowship, a Robert Bingham/PEN Award, a Whiting Writer’s Award, and a National Endowment of the Arts. Plaintiff Klam’s works have been selected as Notable Books of the year by *The New York Times*, *The Los Angeles Times*, *the Kansas City Star*, and the

1 *Washington Post*. Plaintiff Klam’s works include copyright-management information that
2 provides information about the copyrighted work, including the title of the work, its ISBN or
3 copyright registration number, the name of the author, and the year of publication.

4 16. Plaintiff Rachel Louise Snyder (“Plaintiff Snyder”) is a resident of Washington,
5 D.C. Plaintiff Snyder is an author who owns registered copyrights in many works, including but
6 not limited to, *Women We Buried*, *Women We Burned*, *No Visible Bruises – What We Don’t*
7 *Know About Domestic Violence Can Kill Us*, *What We’ve Lost is Nothing*, and *Fugitive Denim:*
8 *A Moving Story of People and Pants in the Borderless World of Global Trade*. Plaintiff Snyder
9 is a Guggenheim fellow and the recipient of the J. Anthony Lukas Work-in-Progress Award, the
10 Hillman Prize, and the Helen Bernstein Book Award, and was a finalist for the National Book
11 Critics Circle Award, *Los Angeles Times* Book Prize, and Kirkus Award. Her work has appeared
12 in *The New Yorker*, *The New York Times*, *Slate*, and in many other publications. Plaintiff
13 Snyder’s works include copyright-management information that provides information about the
14 copyrighted work, including the title of the work, its ISBN or copyright registration number, the
15 name of the author, and the year of publication.

16 17. Plaintiff Ayelet Waldman (“Plaintiff Waldman”) is a resident of California.
17 Plaintiff Waldman is an author and screen and television writer who owns registered copyrights
18 in several works, including but not limited to, *Love and other Impossible Pursuits*, *Red Hook*
19 *Road*, *Love and Treasure*, *Bad Mother*, *Daughter’s Keeper*, *A Really Good Day*, *Fight of the*
20 *Century*, and *Kingdom of Olives and Ash*. Plaintiff Waldman has been nominated for an Emmy
21 and a Golden Globe and is the recipient of numerous awards including a Peabody, AFI award,
22 and a Pen Award, among others. Plaintiff Waldman’s works include copyright-management
23 information that provides information about the copyrighted work, including the title of the
24 work, its ISBN or copyright registration number, the name of the author, and the year of
25 publication.

26 18. At all times relevant hereto, Plaintiffs have been and remain the holders of the
27 exclusive rights under the Copyright Act of 1976 (17 U.S.C. §§ 101, *et seq.* and all amendments
28

1 thereto) to reproduce, distribute, display, or license the reproduction, distribution, and/or display
2 the works identified in paragraphs 13-17, *supra*.

3 **B. Defendants**

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

6 20. Defendant OpenAI, LP is a Delaware limited partnership with its principal place
7 of business located at 3180 18th St., San Francisco, CA 94110. OpenAI, LP is a wholly owned
8 subsidiary of OpenAI, Inc. that is operated for profit. OpenAI, Inc. controls OpenAI, LP directly
9 and through the other OpenAI entities.

10 21. Defendant OpenAI OpCo, LLC is a Delaware limited liability company with its
11 principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI
12 OpCo, LLC is a wholly owned subsidiary of OpenAI, Inc. that is operated for profit. OpenAI,
13 Inc. controls OpenAI OpCo, LLC directly and through the other OpenAI entities.

14 22. Defendant OpenAI GP, LLC is a Delaware limited liability company with its
15 principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI GP,
16 LLC is a general partner of OpenAI, LP. OpenAI GP manages and operates the day-to-day
17 business and affairs of OpenAI, LP. OpenAI GP was aware of the unlawful conduct alleged
18 herein and exercised control over OpenAI, LP throughout the Class Period. OpenAI, Inc. directly
19 controls OpenAI GP.

20 23. Defendant OpenAI Startup Fund I, LP is a Delaware limited partnership with its
21 principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI
22 Startup Fund I, LP was instrumental in the foundation of OpenAI, LP, including the creation of
23 its business strategy and providing initial funding. OpenAI Startup Fund I was aware of the
24 unlawful conduct alleged herein and exercised control over OpenAI, LP throughout the Class
25 Period.

26 24. Defendant OpenAI Startup Fund GP I, LLC is a Delaware limited liability
27 company with its principal place of business located at 3180 18th Street, San Francisco, CA
28 94110. OpenAI Startup Fund GP I, LLC is the general partner of OpenAI Startup Fund I.

1 OpenAI Startup Fund GP I is a party to the unlawful conduct alleged herein. OpenAI Startup
2 Fund GP I manages and operates the day-to-day business and affairs of OpenAI Startup Fund I.

3 25. Defendant OpenAI Startup Fund Management, LLC is a Delaware limited
4 liability company with its principal place of business located at 3180 18th Street, San Francisco,
5 CA 94110. OpenAI Startup Fund Management, LLC is a party to the unlawful conduct herein.
6 OpenAI Startup Fund Management was aware of the unlawful conduct alleged herein and
7 exercised control over OpenAI, LP throughout the Class Period.

8 **FACTUAL ALLEGATIONS**

9 **A. OpenAI’s Artificial Intelligence Products**

10 26. OpenAI researches, develops, releases, and maintains AI products with the
11 intention that its products “benefit all of humanity.”¹

12 27. ChatGPT is among the products OpenAI has developed, engineered, released,
13 and maintained, which utilizes another OpenAI product, GPT models, to respond to text prompts
14 and queries in a natural, coherent, and fluent way through a web interface.

15 28. OpenAI has released a series of upgrades to its GPT model, including GPT-1
16 (released June 2018), GPT-2 (February 2019), GPT-3 (May 2020), GPT-3.5 (March 2022), and
17 most recently, GPT-4 (March 2023)².

18 29. The current version of ChatGPT utilizes both GPT-3.5 and GPT-4; however, the
19 version of ChatGPT that allows users to choose between using GPT-3.5 and GPT-4 is only
20 available to subscribers at a cost of \$20 per month. Otherwise, users are only able to access the
21 version of ChatGPT that relies on the GPT-3.5 model.³

22 30. OpenAI makes ChatGPT available to software developers through an
23 application-programming interface (“API”), which allows developers to write software
24

25
26 ¹ *About*, OpenAI, <https://openai.com/about>

27 ² Fawad Ali, *GPT-1 to GPT-4: Each of OpenAI’s GPT Models Explained and Compared*,
28 *Make Use Of* (Apr. 11, 2023) <https://www.makeuseof.com/gpt-models-explained-and-compared/>

³ *Introducing ChatGPT Plus*, OpenAI (Feb. 1, 2023) <https://openai.com/blog/chatgpt-plus>

1 programs that exchange data with ChatGPT.⁴ OpenAI charges developers for access to ChatGPT
2 by the API on the basis of usage.

3 **B. OpenAI Uses Copyrighted Works in its Training Datasets**

4 31. As mentioned in paragraph 6, *supra*, OpenAI pre-trains its GPT models using a
5 dataset consisting of various sources and content types, including books, plays, articles, and
6 webpage and other written works, to respond accurately to users' prompts and queries.

7 32. OpenAI has admitted that, of all sources and content types that can be used to
8 train the GPT models, written works, plays and articles are valuable training material because
9 they offer the best examples of high-quality, long form writing and "contain[] long stretches of
10 contiguous text, which allows the generative model to learn to condition on long-range
11 information."⁵

12 33. Upon information and belief, OpenAI builds the dataset it uses to train its GPT
13 models by scraping the internet for text data.

14 34. While casting a wide net across the internet to capture the most comprehensive
15 set of content available allows OpenAI to better train its GPT models, this practice necessarily
16 leads OpenAI to capture, download, and copy copyrighted written works, plays and articles.

17 35. Among the content OpenAI has scraped from the internet to construct its training
18 datasets are Plaintiffs' copyrighted works.

19 36. In its June 2018 paper introducing the GPT-1 model, *Improving Language*
20 *Understanding by Generative Pre-Training*, OpenAI revealed that it trained the GPT-1 model
21 using two datasets: "Common Crawl," which is a massive dataset of web pages containing
22 billions of words, and "BookCorpus," which is a collection of "over 7,000 unique unpublished
23 books from a variety of genres including Adventure, Fantasy, and Romance."⁶

24
25
26 ⁴ *OpenAI API*, OpenAI (June 11, 2020) <https://openai.com/blog/openai-api>

27 ⁵ Alec Radford, *Improving Language Understanding by Generative-Pre-Training*, OpenAI
28 (June 11, 2018).

⁶ *Id.*; see also Fawad Ali, *GPT-1 to GPT-4: Each of OpenAI's GPT Models Explained and Compared*, Make Use Of (Apr. 11, 2023) <https://www.makeuseof.com/gpt-models-explained-and-compared/>

1 37. BookCorpus is a controversial dataset, assembled in 2015 by a team of AI
2 researchers funded by Google and Samsung for the sole purpose of training language models
3 like GPT by copying written works from a website called Smashwords, which hosts self-
4 published novels, making them available to readers at no cost.⁷ Despite those novels being
5 largely under copyright, they were copied into the BookCorpus dataset without consent, credit,
6 or compensation to the authors.⁸

7 38. OpenAI also copied many books while training GPT-3. In the July 2020 paper
8 introducing GPT-3, *Language Models are Few-Shot Learners*, OpenAI disclosed, in addition to
9 using the “Common Crawl” and “WebText” datasets that capture web pages, 16% of the GPT-
10 3 training dataset came from “two internet-based book corpora,” which OpenAI simply refers
11 to as “Books1” and “Books2.”⁹

12 39. OpenAI has never revealed what books are part of the Books1 and Books2
13 datasets or how they were obtained. OpenAI has offered a few clues, admitting that these are
14 internet-based datasets that are much larger than BookCorpus.¹⁰ Based on the figures provided
15 in its GPT-3 introductory paper, Books1 is nine times larger than BookCorpus, meaning it
16 contains roughly 63,000 titles, and Books2 is 42 times larger, meaning it contains about 294,000
17 titles.¹¹

18 40. A limited number of internet-based book corpora exist that contain this much
19 material, meaning there are only a handful of possible sources OpenAI could have used to train
20 the GPT-3 model.

21 41. Project Gutenberg is an online archive of e-books whose copyrights have expired.
22 Project Gutenberg has long been popular for training AI systems due to the lack of copyright. In
23 2018, a team of AI researchers created the “Standardized Project Gutenberg Corpus,” which
24

25 ⁷ Jack Bandy, *Dirty Secrets of BookCorpus, a Key Dataset in Machine Learning*, Medium
26 (May 12, 2021) <https://towardsdatascience.com/dirty-secrets-of-bookcorpus-a-key-dataset-in-machine-learning-6ee2927e8650>

27 ⁸ *Id.*

⁹ Tom B. Brown, *Language Models are Few-Shot Learners*, OpenAI (July 22, 2020).

28 ¹⁰ *Id.* at 9.

¹¹ *Id.*

1 contained “more than 50,000 books.”¹² On that information and belief, the OpenAI Books1
2 dataset is based on either the Standardized Project Gutenberg Corpus or Project Gutenberg itself,
3 because of the roughly similar sizes of the two datasets.

4 42. As for the Books2 dataset, the only “internet-based books corpora” that have ever
5 offered that much material are infamous “shadow library” websites, like Library Genesis
6 (“LibGen”), Z-Library, Sci-Hub, and Bibliotik, which host massive collections of pirated books,
7 research papers, and other text-based materials.¹³ The materials aggregated by these websites
8 have also been available in bulk through torrent systems.¹⁴

9 43. These illegal shadow libraries have long been of interest to the AI-training
10 community. For instance, an AI training dataset published in December 2020 by EleutherAI
11 called “Books3” includes a recreation of the Bibliotik collection and contains nearly 200,000
12 books.¹⁵ On information and belief, the OpenAI Books2 dataset includes books copied from
13 these “shadow libraries,” because those are the sources of trainable books most similar in nature
14 and size to OpenAI’s description of Books2.

15 44. When OpenAI introduced GPT-4 in March 2023, the introductory paper
16 contained no information about the dataset used to train it.¹⁶ Instead, OpenAI claims that,
17 “[g]iven both the competitive landscape and the safety implications of large-scale models like
18 GPT-4, this report contains no further details about . . . dataset construction.”¹⁷

19 45. Regarding GPT-4, OpenAI has conceded that it did filter its dataset “to
20 specifically reduce the quantity of inappropriate erotic text content,” implying that it again used
21 a large dataset containing text works.¹⁸

22 **C. OpenAI Unlawfully Infringed Plaintiffs’ Copyrights**

23 ¹² Martin Gerlach, et al., *A standardized Project Gutenberg corpus for statistical analysis of*
24 *natural language and quantitative linguistics*, Cornell University (Dec. 19, 2018),
<https://arxiv.org/pdf/1812.08092.pdf>

25 ¹³ See Claire Woodcock, *‘Shadow Libraries’ Are Moving Their Pirated Books to The Dark*
Web After Fed Crackdowns, Vice (Nov. 30, 2022).

26 ¹⁴ *Id.*

27 ¹⁵ See Alex Perry, *A giant online book collection Meta used to train its AI is gone over*
copyright issues, Mashable (Aug. 18, 2023).

28 ¹⁶ *GPT-4 Technical Report*, OpenAI (Mar. 27, 2023).

¹⁷ *Id.* at 2.

¹⁸ *Id.* at 61.

1 46. As explained, ChatGPT’s responses to user queries or prompts, like other large
2 language models, rely on the data upon which it is trained to generate responsive content. For
3 example, if ChatGPT is prompted to generate a writing in the style of a certain author, GPT
4 would generate content based on patterns and connections it learned from analysis of that
5 author’s work within its training dataset.

6 47. On information and belief, the reason ChatGPT can generate a writing in the style
7 of a certain author or accurately summarize a certain copyrighted book and provide in-depth
8 analysis of that book is because it was copied by OpenAI and copied and analyzed by the
9 underlying GPT model as part of its training data.

10 48. When ChatGPT is prompted to summarize copyrighted written works authored
11 by Plaintiffs, it generates accurate, in-depth summaries and analyses of their works.

12 49. For example, when prompted, ChatGPT accurately summarized Plaintiff
13 Chabon’s novel *The Amazing Adventures of Kavalier & Clay*. When prompted to identify
14 examples of trauma in the *Amazing Adventures of Kavalier & Clay*, ChatGPT identified six
15 specific examples, including how the main character’s “experiences in Europe, including
16 witnessing the persecution of Jews and the loss of his family, haunt him throughout the story.”
17 When asked to write a paragraph in the style of *The Amazing Adventures of Kavalier & Clay*,
18 ChatGPT generated a passage imitating Plaintiff Chabon’s writing style including references to
19 the characters dealing with “the weight of the world at war.” *Exhibit A*.

20 50. ChatGPT similarly provided in depth summaries and analyses of Plaintiff
21 Hwang’s play, *The Dance and the Railroad*. For example, when prompted, ChatGPT identified
22 five key themes from *The Dance and the Railroad*, including “art and creativity as a form of
23 resistance” and “using art as a form of escape from the harsh realities and dehumanization of
24 labor.” Additionally, when prompted to produce a screenplay in the style of *The Dance and the*
25 *Railroad*, ChatGPT produced a script written in Plaintiff Hwang’s style, which generated a
26 screenplay involving a Chinese laborer toiling on the Central Pacific Railroad that “believe[s]
27 in the power of art to keep [their] spirits alive.” *Exhibit B*.

28

1 51. Likewise, ChatGPT provided in depth summaries and analyses of Plaintiff
2 Klam’s works. For example, when prompted, Chat GPT accurately summarized Plaintiff Klam’s
3 novel *Who is Rich?* and correctly analyzed the key relationships between the novel’s central
4 character and the other characters in the novel. When asked to identify the main themes in *Who*
5 *is Rich?* Chat GPT accurately identified seven main themes of the novel including “mid-life
6 crisis and identify.” Further, when prompted to write a paragraph in the style of *Who is Rich?*,
7 ChatGPT generated random passages authentically written in Plaintiff Klam’s writing style,
8 including a reference to navigating the “treacherous waters of midlife.” *Exhibit C.*

9 52. In the same vein, after being prompted to summarize Plaintiff Snyder’s book,
10 *What We’ve Lost is Nothing*, ChatGPT accurately identified themes included within the novel,
11 such as “safety, perception, and the fragility of human relationships.” Similarly, once prompted,
12 ChatGPT accurately analyzed the theme of safety using a specific example from the text of
13 Plaintiff Snyder’s copyrighted work, explaining that “the theme of safety is examined through
14 the lens of a series of burglaries that occur in a suburban neighborhood . . . and how these
15 incidents affect the characters and their perceptions of the world around them.” ChatGPT was
16 also able to generate random passages authentically written in Plaintiff Snyder’s writing style
17 when prompted. *Exhibit D.*

18 53. Additionally, ChatGPT provided in depth summaries and analyses of Plaintiff
19 Waldman’s works. For instance, when prompted to summarize Plaintiff Waldman’s novel *Love*
20 *and Other Impossible Pursuits*, Chat GPT accurately provided a summary and analysis of the
21 novel. When prompted to identify specific instances of grief in *Love and other Impossible*
22 *Pursuits*, ChatGPT identified five specific instances of grief, including the protagonist Emelia’s
23 loss of her infant daughter, a “loss that occurred before the events of the novel and [that] continue
24 to haunt Emelia, affecting her emotional state and relationships.” When prompted to write a
25 paragraph in the style of *Love and Other Impossible Pursuits*, ChatGPT generated a paragraph
26 imitating Plaintiff Waldman’s writing style, including references to the “weight of her
27 daughter’s absence.” *Exhibit E.*

28

- 1 downloaded and copied Plaintiffs’ and the Class’s copyrighted books;
- 2 c. Whether ChatGPT itself is an infringing derivative work based on Plaintiffs’ and
- 3 the Class’s copyrighted books;
- 4 d. Whether the text responses of ChatGPT are infringing derivative works based on
- 5 Plaintiffs’ and the Class’s copyrighted books;
- 6 e. Whether Defendants violated the DMCA by removing copyright-management
- 7 information from Plaintiffs’ and the Class’s copyrighted books;
- 8 f. Whether Defendants were unjustly enriched by the unlawful conduct alleged
- 9 herein;
- 10 g. Whether Defendants’ conduct violates the California Unfair Competition Law;
- 11 h. Whether Plaintiffs and the other Class members are entitled to equitable relief,
- 12 including, but not limited to, restitution or injunctive relief; and
- 13 i. Whether Plaintiffs and the other Class members are entitled to damages and other
- 14 monetary relief and, if so, in what amount.

15 61. Typicality. Federal Rule of Civil Procedure 23(a)(3): Plaintiffs’ claims are
 16 typical of the other Class members’ claims because, among other things, all Class members were
 17 comparably injured through Defendants’ wrongful conduct as described above.

18 62. Adequacy. Federal Rule of Civil Procedure 23(a)(4): Plaintiffs are adequate
 19 Class representative because their interests do not conflict with the interests of the other
 20 members of the Class they seeks to represent; Plaintiff have retained counsel competent and
 21 experienced in complex class action litigation; and Plaintiffs intend to prosecute this action
 22 vigorously. The interests of the Class will be fairly and adequately protected by Plaintiffs and
 23 their counsel.

24 63. Declaratory and Injunctive Relief. Federal Rule of Civil Procedure 23(b)(2):
 25 Defendants have acted or refused to act on grounds generally applicable to Plaintiffs and the
 26 other members of the Class, thereby making appropriate final injunctive relief and declaratory
 27 relief with respect to the Class as a whole.

28 64. Superiority. Federal Rule of Civil Procedure 23(b)(3): A class action is superior

1 to any other available means for the fair and efficient adjudication of this controversy, and no
2 unusual difficulties are likely to be encountered in the management of this class action. The
3 damages or other financial detriment suffered by Plaintiffs and the other Class members are
4 relatively small compared to the burden and expense that would be required to individually
5 litigate their claims against Defendants, so it would be impracticable for the members of the
6 Class to individually seek redress for Defendants’ wrongful conduct. Even if Class members
7 could afford individual litigation, the court system could not. Individualized litigation creates a
8 potential for inconsistent or contradictory judgments, and increases the delay and expense to all
9 parties and the court system. By contrast, the class action device presents far fewer management
10 difficulties, and provides the benefits of single adjudication, economy of scale, and
11 comprehensive supervision by a single court.

12 **CAUSES OF ACTION**

13 **FIRST CAUSE OF ACTION**

14 **DIRECT COPYRIGHT INFRINGEMENT,**
15 **17 U.S.C. § 106, et seq.**

16 65. Plaintiffs hereby incorporate by reference the allegations contained in the
17 preceding paragraphs of this Complaint.

18 66. Plaintiffs bring this claim on behalf of themselves and on behalf of the Class
19 against Defendants.

20 67. As the owners of the registered copyrights in books used to train OpenAI’s GPT
21 models, Plaintiffs and the Class hold the exclusive rights to those works under 17 U.S.C. § 106.

22 68. Plaintiffs have obtained copyright registrations for each of the works identified
23 in Exhibit B.

24 69. On information and belief, to train OpenAI’s GPT models, OpenAI relied on
25 harvesting mass quantities of content from the public internet, including Plaintiffs’ and the
26 Class’s books, which are available in digital formats.

27 70. Because OpenAI’s GPT models cannot function without the expressive
28 information extracted from Plaintiffs’ and Class members’ works and retained by the GPT

1 models, GPT and ChatGPT are themselves infringing derivative works without Plaintiffs' and
2 Class members' permission and in violation of their exclusive rights under the Copyright Act.

3 71. Plaintiffs and the Class never authorized OpenAI to make copies of their written
4 works, make derivative works, publicly display copies (or derivative works), or distribute copies
5 (or derivative works). Each of those rights belong exclusively to Plaintiffs and Class members
6 under copyright law.

7 72. By and through the actions alleged above, OpenAI has infringed and will
8 continue to infringe Plaintiffs' and the Class's copyrights.

9 73. OpenAI's acts of copyright infringement have been intentional, willful, and in
10 callous disregard of Plaintiffs' and Class members' rights. OpenAI knew at all relevant times
11 that the datasets it used to train its GPT models contained copyrighted materials, and that its acts
12 were in violation of the terms of use of the materials.

13 74. OpenAI engaged in the infringing acts described herein for its own commercial
14 benefit.

15 75. As a direct and proximate result of OpenAI's wrongful conduct, Plaintiffs have
16 been substantially and irreparably injured by OpenAI's acts of direct copyright infringement in
17 an amount not readily capable of determination and, unless permanently enjoined from further
18 acts of infringement and continuing to use and distribute GPT models trained using Plaintiffs'
19 and Class members' copyrighted materials without permission, OpenAI will cause additional
20 irreparable harm for which there is no adequate remedy at law. Plaintiff and the Class are thus
21 entitled to permanent injunctive relief preventing OpenAI from engaging in any further
22 infringement of Plaintiffs' and the Class's copyrighted works.

23 76. Plaintiffs are further entitled to recover statutory damages, actual damages,
24 restitution of profits, and other remedies provided by law.

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

SECOND CAUSE OF ACTION

**VICARIOUS COPYRIGHT INFRINGEMENT
17 U.S.C. § 106**

77. Plaintiffs incorporate by reference all allegations of the preceding paragraphs as though fully set forth herein.

78. Plaintiffs bring this claim on behalf of herself and on behalf of the Class against Defendants.

79. Defendant OpenAI, LP is the for-profit subsidiary of Defendant OpenAI, Inc. and is principally responsible for and dedicated to the development of the GPT models and ChatGPT products at issue in this action. Defendant OpenAI Startup Fund Management, LLC exercised control over Defendant OpenAI, LP, along with Defendant OpenAI GP, LLC, which is the general partner of Defendant OpenAI, LP, responsible for managing and operating the day-to-day business affairs of Defendant OpenAI, LP, and is wholly owned and controlled by Defendant OpenAI, Inc., along with Defendant OpenAI OpCo, LLC. Upon information and belief, Defendant OpenAI Startup Fund I, LP played a vital role in the foundation of Defendant OpenAI, LP, including providing initial funding and creating its business strategy, while Defendant OpenAI Startup Fund GP I, LLC is the general partner of Defendant OpenAI Startup Fund I, LP, responsible for managing and operating the day-to-day business affairs of Defendant OpenAI Startup Fund I, LP.

80. Defendant OpenAI, LP directly infringed upon Plaintiffs’ and Class members’ copyrighted works through the unauthorized use and reproduction of the works, and preparation of derivative works by ChatGPT. As discussed above, Plaintiffs’ and Class members’ protected works were used to train GPT models. Because the GPT models are based on expressive information extracted from Plaintiffs’ and Class members’ works, Defendant OpenAI, LP is directly liable for unauthorized use, reproduction, display of copyrighted works, as well as creation of derivative works through ChatGPT’s responses. Therefore, Defendant OpenAI, LP directly infringed upon Plaintiffs’ and Class members’ exclusive rights under 17 U.S.C. § 106.

1 81. Defendants OpenAI, Inc., OpenAI OpCo, LLC, OpenAI GP, LLC, OpenAI
2 Startup Fund GP I, LLC, OpenAI Startup Fund I, LP, and OpenAI Startup Management LLC
3 are vicariously liable for the infringement alleged herein because they had the right and ability
4 to supervise and control the infringing activity but failed to stop the infringing conduct.

5 82. Furthermore, Defendants have a direct financial interest in the infringing conduct
6 and received revenue in connection with the development, deployment, and advancement of the
7 GPT models and ChatGPT. Each entity profited from the advancement of GPT models and
8 ChatGPT.

9
10 83. These committed acts of copyright infringement were willful, intentional, and
11 malicious and thus subjects Defendants to liability for statutory damages under Section
12 504(c)(2) of the Copyright Act of up to \$150,000 per infringement.

13
14 84. Plaintiffs and Class members have been injured by Defendants’ acts of vicarious
15 copyright infringement. Plaintiffs and the Class are entitled to statutory damages, actual
16 damages, restitution of profits, and other remedies provided by law.

17 **THIRD CAUSE OF ACTION**

18 **DIGITAL MILLENNIUM COPYRIGHT ACT – REMOVAL OF COPYRIGHT**
19 **MANAGEMENT INFORMATION**
20 **17 U.S.C. § 1202(B)**

21 85. Plaintiffs incorporate by reference all allegations of the preceding paragraphs as
22 though fully set forth herein.

23 86. Plaintiffs bring this claim on behalf of herself and on behalf of the Class against
24 Defendants.

25 87. Plaintiffs and Class members included one or more forms of copyright-
26 management information in each of Plaintiffs’ and Class members’ infringed works, including:
27 copyright notice, title and other identifying information, the name or other identifying
28 information about the owners of each book, terms and conditions of use, and identifying
numbers or symbols referring to the copyright-management information.

1 88. Without the authority of Plaintiffs and the Class, OpenAI copied Plaintiffs' and
2 Class members' works and used them as training data for its GPT software. By design, the
3 training process does not preserve any copyright-management information. Therefore, OpenAI
4 intentionally removed copyright-management information from Plaintiffs' and Class members'
5 works in violation of 17 U.S.C. § 1202(b)(1).

6 89. OpenAI's removal or alteration of Plaintiffs' and the Class's copyright-
7 management information has been done knowingly and with the intent to induce, enable,
8 facilitate, or conceal infringement of Plaintiffs' and the Class's copyrights.

9 90. Without the authority of Plaintiffs and the Class, Defendants created derivative
10 works based on Plaintiffs' and Class members' works. By distributing these works without their
11 copyright-management information, OpenAI violated 17 U.S.C. § 1202(b)(3).

12 91. OpenAI knew or had reasonable grounds to know that this removal of copyright-
13 management information would facilitate copyright infringement by concealing the fact that
14 every output from ChatGPT is an infringing derivative work, synthesized entirely from
15 expressive information found in the training data.

16 92. Plaintiffs and the Class have been injured by OpenAI's removal of copyright-
17 management information. Plaintiffs and the Class are entitled to statutory damages, actual
18 damages, restitution of profits, and other remedies provided by law, including full costs and
19 attorneys' fees.

20 **FOURTH CAUSE OF ACTION**

21 **VIOLATIONS OF THE CALIFORNIA UNFAIR COMPETITION LAW**
22 **CAL. BUS. & PROF. CODE §§ 17200, *ET SEQ.***

23 93. Plaintiffs and the Class incorporate by reference each preceding and succeeding
24 paragraph as though fully set forth at length herein.

25 94. Plaintiffs bring this claim on behalf of herself and on behalf of the Class against
26 Defendants.

1 95. The California Unfair Competition Law (“UCL”) prohibits acts of “unfair
2 competition,” including any “unlawful, unfair or fraudulent business act or practice” and “unfair,
3 deceptive, untrue or misleading advertising.” Cal. Bus. & Prof. Code § 17200.

4 96. Defendants have engaged in unfair competition and unfair, unlawful or
5 fraudulent business practices by the conduct, statements, and omissions described above because
6 it illegally collected and used Plaintiffs’ and the Class’s copyrighted works to train its GPT
7 models.

8 97. The unlawful business practices described herein violate the UCL because
9 Defendants used Plaintiffs’ and the Class’s protected works to train its GPT software for
10 Defendants’ own commercial profit without Plaintiffs’ and the Class’s authorization.
11 Defendants further knowingly designed ChatGPT to include portions or summaries of Plaintiffs’
12 copyrighted works without attributions in its responses, and they unfairly profit from and take
13 credit for developing a commercial product based on unattributed reproductions of those stolen
14 writing and ideas.

15 98. These acts and practices have deceived Plaintiffs and are likely to deceive the
16 public into believing that Plaintiffs and the Class have granted OpenAI the right to use its
17 copyrighted materials. In failing to disclose the sources of its training datasets and suppressing
18 other material facts from Plaintiffs and Class members as well as the public, Defendant breached
19 its duties to disclose these facts, violated the UCL, and caused injuries to Plaintiffs and Class
20 members. The omissions and acts of concealment by Defendants pertained to information that
21 was material to Plaintiffs and Class members, as it would have been to all reasonable consumers.

22 99. The injuries suffered by Plaintiffs and Class members are not greatly outweighed
23 by any potential countervailing benefit to consumers or to competition, nor are they injuries that
24 Plaintiffs and Class members should have reasonably avoided.

25 100. Defendant’s acts and practices are unlawful because they violate California Civil
26 Code §§ 1668, 1709, 1710, and 1750 *et seq.*, and California Commercial Code § 2313.

27 101. Plaintiffs seek to enjoin further unlawful, unfair and/or fraudulent acts or
28 practices by Defendants, to obtain restitutionary disgorgement of all monies and revenues

1 generated as a result of such practices, and all other relief allowed under California Business &
2 Professions Code § 17200.

3 **FIFTH CAUSE OF ACTION**

4 **NEGLIGENCE**

5 102. Plaintiffs incorporate by reference the allegations of all foregoing paragraphs as
6 if they had been set forth in full herein.

7 103. Plaintiffs bring this claim on behalf of themselves and on behalf of the Class
8 against Defendants.

9 104. Defendants owed a duty of care toward Plaintiffs and the Class in (1) obtaining
10 data to train its GPT models and (2) not using Plaintiffs’ and the Class’s protected works to train
11 its GPT models.

12 105. Defendants have a common law duty to prevent foreseeable harm to others,
13 including Plaintiffs and members of the Class, who were foreseeable and probable victims of
14 Defendants’ unlawful practices.

15 106. Defendants breached their duty to exercise due care by negligently, carelessly,
16 and recklessly collecting, maintaining, and controlling Plaintiffs’ and Class members’ works
17 and engineering, designing, maintaining, and controlling systems—including ChatGPT—that
18 are trained on Plaintiffs’ and Class members’ works without their authorization.

19 107. The damages suffered by Plaintiffs and the Class were the direct and reasonably
20 foreseeable result of Defendants’ negligent breach of their duties to adequately design,
21 implement, and maintain reasonable practices to avoid infringing protected works without
22 consent of copyright holders.

23 108. Defendants’ negligence directly caused significant harm to Plaintiffs and the
24 Class.

25 **SIXTH CAUSE OF ACTION**

26 **UNJUST ENRICHMENT**

27 109. Plaintiffs incorporate by reference all allegations of the preceding paragraphs as
28 though fully set forth herein.

1 110. By virtue of the unlawful, unfair, and deceptive conduct alleged herein,
2 Defendants knowingly realized substantial revenue from the use of Plaintiffs' and Class
3 members' copyrighted works for the commercial training of its GPT models used to power its
4 ChatGPT product.

5 111. Defendant knew or should have known that Plaintiffs and the Class have invested
6 substantial time and energy creating the works in which they hold a copyright.

7 112. Defendants were conferred significant benefits when they downloaded and
8 copied Plaintiffs' and the Class's copyrighted works to train their GPT software without
9 Plaintiffs' and the Class's permission. Defendant knowingly and willingly accepted and enjoyed
10 those benefits.

11 113. By using Plaintiffs' and the Class's copyrighted works to train ChatGPT,
12 Defendants caused Plaintiffs and the Class to suffer actual damages from the deprivation of the
13 benefits of their work, including monetary damages.

14 114. Defendants derived profit and other economic benefits from the use of Plaintiffs'
15 and the Class's copyrighted works to train ChatGPT.

16 115. It would be inequitable and unjust to permit Defendants to retain the enormous
17 economic benefits it has obtained from and/or at the expense of Plaintiffs and Class members.

18 116. As a direct and proximate cause of Defendants' unjust enrichment, Plaintiffs and
19 the Class are entitled to restitution, attorneys' fees, costs and interest.

20 117. Defendants' conduct is causing and, unless enjoined and restrained by this Court,
21 will continue to cause Plaintiffs and the Class irreparable injury that cannot be compensated or
22 measured in money.

23 **REQUEST FOR RELIEF**

24 WHEREFORE, Plaintiffs, individually and on behalf of members of the Class defined
25 above, respectfully request that the Court enter judgment against Defendants and award the
26 following relief:

27 A. Certification of this action as a class action pursuant to Rule 23 of the Federal
28 Rules of Civil Procedure, declaring Plaintiffs as the representative of the Class, and Plaintiffs'

1 counsel as counsel for the Class;

2 B. An order awarding declaratory relief and temporarily and permanently enjoining
3 Defendant from continuing the unlawful and unfair business practices alleged in this Complaint
4 and to ensure that all applicable information set forth in 17 U.S.C. § 1203(b)(1) is included when
5 appropriate;

6 C. An award of statutory and other damages under 17 U.S.C. § 504 for violations of
7 the copyrights of Plaintiff and the Class by Defendants.

8 D. An award of statutory damages under 17 U.S.C. § 1203(b)(3) and 17 U.S.C. §
9 1203(c)(3), or in the alternative, an award of actual damages and any additional profits under 17
10 U.S.C. § 1203(c)(2);

11 E. A declaration that Defendant is financially responsible for all Class notice and
12 the administration of Class relief;

13 F. An order awarding any applicable statutory and civil penalties;

14 G. An order requiring Defendant to pay both pre- and post-judgment interest on any
15 amounts awarded;

16 H. An award of costs, expenses, and attorneys’ fees as permitted by law; and

17 I. Such other or further relief as the Court may deem appropriate, just, and
18 equitable.

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

DEMAND FOR JURY TRIAL

Plaintiffs hereby demand a jury trial for all claims so triable.

DATED: September 8, 2023

Respectfully submitted,

/s/ Daniel J. Muller
DANIEL J. MULLER, SBN 193396
dmuller@venturahersey.com
VENTURA HERSEY & MULLER, LLP
1506 Hamilton Avenue
San Jose, California 95125
Telephone: (408) 512-3022
Facsimile: (408) 512-3023
dmuller@venturahersey.com

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

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

Attorneys for Plaintiffs