UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

JULIAN SANCTON, on behalf of himself and all others similarly situated Plaintiff, v. OPENAI, INC., OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, OPENAI HOLDINGS, LLC, and MICROSOFT CORPORATION, Defendants.	Civil Action No. _____ **CLASS ACTION COMPLAINT** **JURY TRIAL DEMANDED**

Plaintiff Julian Sancton, on behalf of himself and all other similarly situated (the "Class," as defined below), for his complaint against Defendants OpenAI, Inc., OpenAI GP LLC, OpenAI, LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, (collectively "OpenAI") and Microsoft Corporation (all collectively "Defendants"), alleges as follows:

NATURE OF THE ACTION

1. OpenAI and Microsoft have built a business valued into the tens of billions of dollars by taking the combined works of humanity without permission. Rather than pay for intellectual property, they pretend as if the laws protecting copyright do not exist. Yet the United States Constitution itself protects the fundamental principle that creators deserve compensation for their works. Nonfiction authors often spend years conceiving, researching, and writing their creations. While OpenAI and Microsoft refuse to pay nonfiction authors, their AI platform is worth a fortune. The basis of the OpenAI platform is nothing less than the rampant theft of copyrighted works.

2. Plaintiff Julian Sancton is a writer and the author of the *New York Times* best-seller *Madhouse at the End of the Earth: the Belgica's Journey into the Dark Antarctic Night*, a book documenting the true story of an Antarctic polar expedition at the end of the nineteenth century. Plaintiff Sancton dedicated five years of his life and tens of thousands of dollars to completing the book, traveling around the world to Antarctica, Belgium, and Norway to complete his research. Such an investment of time and money is feasible for Plaintiff Sancton and other writers because, in exchange for their creative efforts, the Copyright Act grants them "a bundle of exclusive rights" in their works, including "the rights to reproduce the copyrighted work[s]." *Andy Warhol Found. for the Visual Arts, Inc. v. Goldsmith*, 598 U.S. 508, 526 (2023).

3. This case is about Defendants OpenAI and Microsoft's complete disregard for those exclusive rights. Defendants have made commercial reproductions of millions, maybe billions, of copyrighted works without any compensation to authors, without a license, and without permission. In doing so, they have infringed on the exclusive rights of Plaintiff Sancton and other writers and rightsholders whose work has been copied and appropriated to train their artificial intelligence models.

4. Defendants OpenAI and Microsoft collaborated closely to create and monetize the generative artificial intelligence models known as GPT-3, GPT-3.5, GPT-4, and GPT-4 Turbo. These are the computer models that power the popular ChatGPT chatbot, which Defendants have followed with a suite of other commercial offerings, like ChatGPT Enterprise, ChatGPT Plus, Bing Chat, Browse with Bing, Microsoft Copilot, and others. Defendants' GPT models have been designed to recognize and process text inputs from a user and, in response, generate text that has been calibrated to mimic a human written response.

5. That end product—a computer model and chatbot built to mimic human written

expression—came at a price. Defendants' models were calibrated (or "trained," in Defendants' parlance) by reproducing a massive corpus of copyrighted material, including, upon information and belief, tens or hundreds of thousands of nonfiction books. The only way that Defendants' models could be trained to generate text output that resembles human expression is to copy and analyze a large, diverse corpus of text written by humans. In training their models, Defendants reproduced copyrighted material to exploit precisely what the Copyright Act was designed to protect: the elements of protectible expression within them, like the style, word choice, and arrangement and presentation of facts. In OpenAI's words, the goal of the training process was to teach their model to "learn" "how words fit together grammatically," "how words work together to form higher-level ideas," and "how sequences of words form structured thoughts."[1]

6. Defendants copied and data-mined the works of writers, without compensation, to build a machine that is capable (or, as technology advances, will soon be capable) of performing the same type of work for which these writers would be paid. Without the wide corpus of copyrighted material to feed off of, there would be no ChatGPT. Defendants' commercial success was possible only because they copied and digested the protected, copyrightable expression contained in billions of pages of actual text, across millions of copyrighted works—all without paying a penny to authors and rightsholders.

7. Defendants OpenAI and Microsoft have enjoyed enormous financial gain from their exploitation of copyrighted material. OpenAI recently reported that it is "generating revenue at a pace of $1.3 billion a year."[2] Microsoft, for its part, has seen its investment in OpenAI increase

[1] Fred von Lohmann, response to Notice of Inquiry and Request for Comment 5, (Oct. 30, 2023), *available at* https://downloads.regulations.gov/COLC-2023-0006-8906/attachment_1.pdf.
[2] AJ Hess, The Biggest Challenges Facing OpenAI's Mira Murati, the Newly Minted Most Powerful Woman in Tech, FAST COMPANY (last visited Nov. 20, 2023), https://www.fastcompany.com/90985829/the-biggest-challenges-facing-openais-mira-murati-the-newly-minted-most-powerful-woman-in-tech.

many-fold and its own GPT-based products, like BingChat, succeed in the marketplace. And its stock price has increased as Microsoft has touted its ability to exploit and leverage AI across its products. Analysts project that the integration of GPT into Microsoft products could generate more than $10 billion in annualized revenue by 2026,[3] with just one version of this integration—"GitHub Copilot"—already generating more than $100 million in annual recurring revenue.[4] In developing and monetizing these AI products, Microsoft and OpenAI have been close partners every step of the way, from the training of GPT-3 to today. The OpenAI-Microsoft relationship is so close, in fact, that OpenAI's former CEO Sam Altman and former Chief Scientist Greg Brockman just left the company to lead a new artificial intelligence research team at Microsoft.

8. OpenAI and Microsoft's commercial gain has come at the expense of creators and rightsholders like Plaintiff and members of the Class. A person who reads a book typically buys it from a store. But Defendants did not even do that. Neither OpenAI nor Microsoft have paid for the books used to train their models. Nor have Defendants sought to obtain—or pay for—a license to copy and exploit the protected expression contained in the copyrighted works used to train their models. Instead, Defendants took these works; they made unlicensed copies of them; and they used those unlicensed copies to digest and analyze the copyrighted expression in them, all for commercial gain. The end result is a computer model that is not only built on the work of thousands of creators and authors, but also built to generate a wide range of expression—from shortform articles to book chapters—that mimics the syntax, style, and themes of the copyrighted works on which it was trained.

---

[3] Jordan Novet, *Microsoft Starts Selling AI Toll for Office, Which Could Generate $10 Billion a Year by 2026*, CNBC (last visited NOV. 20, 2023) https://www.cnbc.com/2023/11/01/microsoft-365-copilot-becomes-generally-available.html.

[4] Aaron Holmes, *Microsoft's GitHub AI Coding Assistant Exceeds $100 Million in Recurring Revenue*, THE INFORMATION, (last visited Nov. 20, 2023), https://www.theinformation.com/briefings/microsoft-github-copilot-revenue-100-million-ARR-ai (*available at* https://perma.cc/5S7F-4GBY).

9. Plaintiff, on behalf of himself and the proposed Class, seeks damages from Defendants for their largescale infringement of their copyrighted works, as well as injunctive relief.

JURISDICTION AND VENUE

10. The Court has subject matter jurisdiction under 28 U.S.C. §§ 1331 and 1338(a) because this action arises under the Copyright Act of 1976, 17 U.S.C. § 101, *et seq*.

11. The Court also has personal jurisdiction over Defendants because they have purposely availed themselves of the privilege of conducting business in New York.

12. OpenAI and Microsoft's copyright infringement and contributory copyright infringement, in substantial part, occurred in this District. OpenAI sold and distributed, and continues to sell and distribute, its GPT products, including ChatGPT, ChatGPT Enterprise, ChatGPT Plus, Browse with Bing, and application programming interface tools (API) within New York and to New York residents. OpenAI marketed and sold GPT-based products to New York residents and New York-based companies.

13. Microsoft distributed and sold GPT-based products, like Bing Chat and Azure products that incorporate GPT-3 and GPT-4. Upon information and belief, Microsoft also assisted OpenAI's copyright infringement from New York, including from Azure datacenters located in New York used to facilitate OpenAI's use and exploitation of the training dataset used for development of OpenAI's GPT models. Microsoft maintains offices and employs personnel in New York. Upon information and belief, Microsoft's New York personnel were involved in the creation and maintenance of the supercomputing systems that powered OpenAI's widespread infringement, as well as in the commercialization and monetization of OpenAI's GPT models.

14. Plaintiff Julian Sancton is a citizen of New York, and resides within this District. The injuries alleged here from Defendants' infringement occurred in this District.

15. Venue is proper under 28 U.S.C. § 1400(a) because Defendants or their agents reside or may be found in this District due to their infringing activities, along with their commercialization of their infringing activities, that occurred in this District. Venue is also proper under 28 U.S.C. § 1391(b)(2) because a substantial part of the events giving rise to Plaintiff's claims occurred in this District, including the sales of Defendants' GPT-based products within this District.

THE PARTIES

16. Plaintiff Julian Sancton is a writer who resides in New York. He is the author of the *New York Times* bestseller *Madhouse at the End of the Earth: The Belgica's Journey Into the Dark Antarctic Night*. He is a senior features editor of *The Hollywood Reporter* and his work has appeared in *GQ*, *Wired*, and *The New Yorker*. Mr. Sancton owns the registered copyright in *Madhouse at the End of the Earth*, register number TX0009331888.

17. Defendant OpenAI Inc. is a Delaware nonprofit corporation with a principal place of business in San Francisco, California. OpenAI Inc. was formed in December 2015. OpenAI Inc. owns and controls all other OpenAI entities.

18. Defendant OpenAI GP, LLC is a Delaware limited liability company with a principal place of business in San Francisco, California. OpenAI GP, LLC wholly owns and controls OpenAI OpCo LLC, which until recently was known as OpenAI LP. OpenAI, Inc. uses OpenAI GP LLC to control OpenAI OpCo LLC and OpenAI Global, LLC. OpenAI GP LLC was involved in the copyright infringement alleged here through its direction and control of OpenAI OpCo LLC and OpenAI Global LLC.

19. Defendant OpenAI OpCo LLC is a Delaware limited liability company with a principal place of business in San Francisco, California. OpenAI OpCo LLC was formerly known as OpenAI LP. OpenAI OpCo LLC is the sole member of OpenAI, LLC, and has been directly

involved in OpenAI's mass infringement and has directed this infringement through its control of OpenAI, LLC. OpenAI OpCo LLC serves as the for-profit arm of OpenAI.

20. Defendant OpenAI, LLC is a Delaware limited liability company with a principal place of business in San Francisco, California. OpenAI, LLC was formed in September 2020. OpenAI LLC monetizes and distributes OpenAI's GPT-based products, all of which born out of OpenAI's copyright infringement. Upon information and belief, OpenAI, LLC is owned and controlled by both OpenAI Inc. and Microsoft Corporation, through OpenAI Global LLC and OpenAI OpCo LLC.

21. Defendant OpenAI Global LLC is a Delaware limited liability company with a principal place of business in San Francisco, California. Microsoft Corporation has a minority stake in OpenAI Global LLC and OpenAI, Inc. has a majority stake in OpenAI Global LLC, indirectly through OpenAI Holdings LLC and OAI Corporation, LLC. OpenAI Global LLC was involved in the copyright infringement alleged here through its ownership, control, and direction of OpenAI LLC.

22. Defendant OAI Corporation, LLC is a Delaware limited liability company with a principal place of business in San Francisco, California. OAI Corporation, LLC's sole member is OpenAI Holdings, LLC. OAI Corporation, LLC was and is involved in the unlawful conduct alleged herein through its ownership, control, and direction of OpenAI Global LLC and OpenAI LLC.

23. Defendant OpenAI Holdings, LLC is a Delaware limited liability company, whose sole members are OpenAI, Inc. and Aestas, LLC. The sole member of Aestas, LLC is Aestas Management Company, LLC. Aestas Management Company, LLC is a Delaware company created to facilitate a half-billion-dollar capital raise for OpenAI. OpenAI Holdings LLC was involved in

the infringement alleged herein through its indirect ownership, control, and direction of OpenAI OpCo LLC.

24. Microsoft Corporation is a Washington corporation with a principal place of business and headquarters in Redmond, Washington. Microsoft has invested at least $13 billion in OpenAI, and reportedly owns a 49% stake in the company's for-profit operations. Microsoft has described its relationship with the OpenAI Defendants as a "partnership." This Microsoft-OpenAI partnership has included the creation, development, and maintenance of the supercomputing systems that the OpenAI Defendants used to house and make copies of copyrighted material in the training set for OpenAI's large language models. In course of designing and maintaining these tailored supercomputing systems for OpenAI's needs, upon information and belief, Microsoft was both directly involved in making reproductions of copyrighted material and facilitated the copyright infringement committed by OpenAI.

FACTUAL ALLEGATIONS

I. OpenAI's History and Collaboration With Microsoft to Build Commercial Large Language Models

A. OpenAI's Early Years and Shift To For-Profit Status

25. OpenAI was formed in December 2015 with $1 billion in investments.

26. At the outset, OpenAI described itself as a "non-profit artificial intelligence research company." OpenAI assured the public that its research and work was drive purely by altruism. In a December 11, 2015 blog post, co-founders Greg Brockman and Ilya Sutskever wrote: "Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact."

27. OpenAI also made a commitment to keep its research and development open for

the public to enjoy and benefit from. In particular, OpenAI promised that both its work and intellectual property would be open and available to the public and "shared with the world."

28. Both commitments were short-lived. By March 2019, OpenAI created a for-profit arm, OpenAI LP, which was formed to manage OpenAI's commercial operations—including, critically, product development. OpenAI LP has since been renamed OpenAI OpCo LLC.

29. In the months and years that followed, OpenAI morphed into a complex (and secretive) labyrinth of for-profit corporate entities to manage OpenAI's day-to-day operations, product development, for-profit research, and billion-dollar capital raises. Today, OpenAI is valued at $29 billion, collecting revenues north of $100 million per month.

30. OpenAI has also abandoned its commitment to openness. Shortly after its formation, OpenAI kept the public informed of its research and data sets. For example, both GPT-1 and GPT-2 were released on a largely open-source basis, with substantial documentation of their dataset and methodology.

31. But after OpenAI formed a for-profit entity and accepted billion-dollar investments from Microsoft and others, that changed. Beginning with GPT-3, which was leaps and bounds more sophisticated than GPT-1 and GPT-2, OpenAI disclosed far less information about the technical details of the model and how it was trained. And it released even less information when it announced the releases of GPT-4 in 2023. For example, the GPT-4 "technical report" said: "this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."[5]

32. Why this sudden secrecy? The motivations are entirely commercial. According to

[5] OpenAI, GPT-4 Technical Report 2 (March 27, 2023), https://cdn.openai.com/papers/gpt-4.pdf.

OpenAI's Chief Scientist Sutskever: "It's competitive out there."[6]

33. Upon information and belief, OpenAI had another reason to keep its training data and development of GPT-3, GPT-3.5, and GPT-4 secret: To keep rightsholders like Plaintiff and members of the Class in the dark about whether their works were being infringed and used to train OpenAI's models. It also appears that OpenAI have made it more difficult to determine whether any particular book is in their training set—sacrificing rights-holders for the sake of covering up its actions.

B. The Development of GPT-3 and Subsequent Commercialization of GPT-based Products

34. OpenAI announced its completion of GPT-3 in May of 2020. In its technical report describing GPT-3, OpenAI previewed that GPT-3 was a far more complex, sophisticated large language model than any that preceded it. The technical paper described GPT-3 as a "language model with 175 billion parameters, 10x more than any previous non-sparse language model." It went on to on to confirm that "GPT-3 achieves strong performance on many NLP [natural language processing] datasets, including translation, question-answering, and cloze tasks, as well as several tasks that require on-the-fly reasoning or domain adaption."

35. OpenAI's GPT-3 paper also acknowledged that GPT-3 was able to generate text material that successfully mimicked some of the nonfiction copyrighted works on which it was trained. For example, GPT-3 was able to "generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans."

36. Upon the release of the GPT-3 technical paper and a limited beta release, GPT-3 made a splash among the engineering and artificial intelligence community. In July 2020, the *MIT*

[6] James Vincent, OpenAI Co-founder on Company's Past Approach to Openly Sharing Research: 'We Were Wrong' THE VERGE (last visited Nov. 20, 2023), https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview.

Technology Review reported that "OpenAI's new language generator GPT-3 is shockingly good" and quoted one developer as saying that "GPT-3 feels like seeing the future."[7] Nevertheless, OpenAI's GPT-3 remained largely obscure to the public for another two years.

37. In November 2022, OpenAI released ChatGPT, a generative AI chatbot powered by GPT-3.5. As OpenAI described it at the time, ChatGPT is "an AI-powered chatbot developed by OpenAI, based on the GPT (Generative Pretrained Transformer) language model. It uses deep learning techniques to generate human-like responses to text inputs in a conversational manner." ChatGPT was released free to the public.

38. ChatGPT gained over 100 million users within three months, becoming the fastest growing internet service of all time.[8]

39. Building on ChatGPT's success, OpenAI now offers a range of services powered by its GPT models. Along with ChatGPT, which is free to use, OpenAI sells the subscription service ChatGPT Plus at $20 per month, and ChatGPT Enterprise, a subscription service aimed at business. OpenAI offers ChatGPT API tools that allow software developers to create new applications building on ChatGPT. OpenAI licenses its technology to corporate clients for licensing fees.

40. Both ChatGPT and OpenAI's commercial offerings have been widely adopted. OpenAI reports, from its internal user statistics, that over 90 percent of Fortune 500 companies are using ChatGPT. OpenAI has also reported that it expects to reach $1 billion in revenue in less than a year.

---

[7] Will Douglas Heaven, Open AI's New Language Gnererator GPT-3 is Shockingly Good—and Completely Mindless, MIT TECHNOLOGY REVIEW, (last accessed Nov. 20, 2023), https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/.

[8] Will Douglass Heaven, ChatGPT is Everywhere. Here's Where it Came From, MIT TECHNOLOGY REVIEW, (last visited Nov. 20, 2023) https://www.technologyreview.com/2023/02/08/1068068/chatgpt-is-everywhere-heres-where-it-came-from/.

41. Credit for OpenAI's success can be attributed to its largescale copyright infringement. ChatGPT's popularity is driven by its ability to produce believable natural language text that mimics what a human would create. As *CNBC* reported shortly after ChatGPT's release: "What makes ChatGPT so impressive is its ability to produce human-like responses, thanks in no small part to the vast amount of data it is trained on."[9]

42. That data, as explained in more detail below, consists largely of copyrighted material. Since its release, the internet has been flooded with guides on how to use ChatGPT to write academic papers, grant proposals, novels, nonfiction books, and other pieces.[10] ChatGPT could only be trained to generate this range of text and style by making unlicensed reproductions of a massive corpus of copyrighted content, including Plaintiff's work and works owned by the Class. Without this largescale infringement—without a large corpus of copyrighted material from which to mine expression—Defendants' GPT models could not have been trained to perform their intended function, that is, to mimic human written expression.

C. The Microsoft-OpenAI Partnership

43. Microsoft, for the last four years, has been deeply involved in the training, development, and commercialization of OpenAI's GPT products. Microsoft CEO Satya Nadella has called its relationship with OpenAI a "great commercial partnership."

44. Given the volume of the training corpus—the equivalent of nearly four billion pages of single-spaced text—and the complexity of OpenAI's large language models, OpenAI required a specialized supercomputing system to train GPT-3, GPT-3.5, or GPT-4 (and thus copy

[9] Ryan Brown, *All You Need To Know About ChatGPT, the A.I. Chatbot That's Got the World Talking and Tech Giants Clashing*, CNBC Online (last visited Nov. 20, 2023), http://tiny.cc/luyevz.
[10] *See, e.g.*, Rob Cubbon, How I Wrote My New Book With ChatGPT, (last accessed Nov. 20, 2023), https://robcubbon.com/how-i-wrote-my-new-book-with-chatgpt/; Aaron Stark, How to Use ChatGPT to Write Non-Fiction Books, (last visited Nov. 20, 2023), https://www.griproom.com/fun/how-to-use-chatgpt-to-write-non-fiction-books.

and exploit the copyrighted material in its training set). That is where Microsoft came in. Microsoft's Azure provided the cloud computing systems that powered the training process, and continues to power OpenAI's operations to this day. Microsoft and OpenAI worked together to design this system, which was used to train all of OpenAI's GPT models. Without these bespoke computing systems, OpenAI would not have been able to execute and profit from the mass copyright infringement alleged herein.

45. Microsoft has in public statements acknowledged its intimate involvement in the development of OpenAI's GPT models. In a 2020 press release, Microsoft announced that it had built a bespoke supercomputing infrastructure "in collaboration and exclusively for OpenAI," "designed specifically to train that company's AI models. It represents a key milestone a partnership announced last year to jointly create new supercomputing technologies in Azure." The press release went on to describe the computer "developed for OpenAI" as "top five" in the "world," and "a single system with more than 285,000 CPU cores, 10,000 GPUs and 4,000 gigabits per second of network connectivity for each GPU server."

46. After the release of ChatGPT, Microsoft also took credit for its substantial role in the training process. In a February 2023 interview with Fortt Knox on CNBC, Mr. Nadella said that "beneath what OpenAI is putting out as large language models, remember, the heavy lifting was done by the Azure team to build the compute infrastructure." A few months later, in his keynote speech at the Microsoft Inspire conference, Mr. Nadella acknowledged that Microsoft "buil[t] the infrastructure to train [OpenAI's] models."

47. Upon information and belief, that "heavy lifting" involved developing, maintaining, troubleshooting, and supporting OpenAI's supercomputing system. Microsoft employees worked closely with OpenAI personnel to understand the training process and training

dataset used for OpenAI's GPT models.

48. Through that process, Microsoft would have known that OpenAI's training data was scraped indiscriminately from the internet and included a massive quantity of pirated and copyrighted material, including a trove of copyrighted nonfiction works. Through its creation and maintenance of the supercomputing system, Microsoft directly made unlicensed copies and provided critical assistance to OpenAI in making unlicensed copies of copyrighted material—including Plaintiff Sancton's work *Madhouse at the End of the Earth* and other nonfiction books—for the purpose of training the GPT models.

49. The large-scale copyright infringement was right in the open for OpenAI and its business partners. Microsoft also became aware of OpenAI's largescale copyright infringement in the course of conducting due diligence required before making its multibillion-dollar investments in OpenAI. As Andreesen Horowitz, another OpenAI investor, put it: "the only practical way generative AI models can exist is if they can be trained on an almost unimaginably massive amount of content, much of which . . . will be subject to copyright."[11] As the public company made its decision to invest $13 billion into OpenAI, surely Microsoft—like Andreesen Horowitz—was fully aware that OpenAI was taking a massive corpus of copyrighted content, without compensation to rightsholders, and copying it for the purpose of training and developing its GPT models to mimic the human writing.

50. In addition to its integral role in facilitating the training process, Microsoft has played a key role in commercializing OpenAI's GPT-based technology, and in doing so has profited from OpenAI's infringement of content owned by Plaintiff and the proposed Class. At Microsoft's largest partner event of the year, Inspire, Mr. Nadella said that while OpenAI is

[11] Andreeson Horowitz, Notice of Inquiry on Artificial Intelligence and Copyright, (Oct. 30, 2023), *available at* https://s3.documentcloud.org/documents/24117939/a16z.pdf.

"innovating on the algorithms and the training of these frontier models, [Microsoft] innovate[s] on applications on top of it." For example, Microsoft unveiled Bing Chat, a generative AI chatbot feature on its search engine powered by GPT-4, and, in turn, ChatGPT integrated a "Browse with Bing" feature on paid ChatGPT Plus offering.

51. Indeed, recent events have further demonstrated the close relationship between OpenAI and Microsoft. When OpenAI CEO Sam Altman was terminated, Microsoft hired him.

II. OpenAI and Microsoft Engaged in Largescale Copyright Infringement in Training the GPT Models

1. GPT Models and the Training Process

52. OpenAI's GPT models are a species of a large language model or "LLM." Large language models are designed to mimic human use of language. LLMs attempt to mimic human understanding of language by processing input text, and attempt to mimic human use of language by generating output text. LLMs like GPT-3, GPT-3.5, and GPT-4 are often described as "neural networks" because they are designed to operate like the neural networks that make up the human brain.

53. OpenAI's GPT-based models are complex mathematical functions comprised of a series of algorithms that break down input text into smaller pieces—words or portions of words, called "tokens"—then translate those pieces into "vectors," or a sequence of numbers that is used to identify the token within the series of algorithms. Those vectors help place each token on a map, by identifying other tokens closely associated with the word. According to OpenAI, "the process begins by breaking text down into roughly word-length 'tokens,' which are converted to numbers. The model then calculates each token's proximity to other tokens in the training data—essentially, how near one word appears in relation to any other word. These relationships between words reveal which words have similar meanings . . . and functions." As the model trains and digests more

expression, the algorithms depicting the relationship between various tokens changes with it.

54. The model is trained on a massive corpus of text. The model takes text inputs in the form of an incomplete phrase or passage, and attempts to complete the phrase, essentially a fill-in-blank quiz. The model compares its predicted phrase completion with the actual "correct" answer. The model then adjusts its internal algorithms to "learn" from its mistakes—in other words, it adjusts its algorithms to reduce the likelihood of making the same mistake again and thus minimize the delta between any given text input and the "correct" text output.

55. The model then repeats this same cycle millions, possibly billions, of times across the entire training corpus, adjusting its algorithms each time to reflect the text input from the training set. As OpenAI describes it, "pre-training teaches language to the model, by showing the model a wide range of text, and, utilizing sophisticated statistical and computational analysis, having it try to predict the word that comes next in each of a huge range of sequences" and from this process "gain[s] fluency in predicting the next word." In this way, the GPT model effectively mines and feeds on the expression contained in the training set, adjusting its algorithms such that it can mirror and mimic the ordering of words, style, syntax, and presentation of facts, concepts, and themes.

56. After the pre-training process, the generative model must undergo a further post-training process. At this point, the model is capable of completing phrases and predicting the next word or words that come next after a particular text input, but is not capable of responding to questions or providing human-like responses. The post-training process is sometimes referred to as "fine-tuning," and involves more human supervision and, according to OpenAI, making "targeted changes to the model, using relatively small (compared to pre-training) and carefully engineered datasets that represent ideal behavior." For both the post- and pre-training processes,

OpenAI creates multiple, unlicensed copies of the training data.

57. The quality and quantity of the training data is critical to the quality of the resulting model. With respect to LLM development, the phrase "garbage in, garbage out" carries particular weight. ChatGPT, for example, has shown the capability of coherently processing large tranches of text input, and generating coherent, clearly-written passages in response—responses that mimic an understanding not just of the proper ordering of words and syntax, but also higher-level themes and ideas. ChatGPT could only develop this capability from training on high-quality prose and complex, longer pieces. To this end, books serve a particularly critical role in the training process. In a recent paper discussing the training of the GPT-2 model, for example, OpenAI described the importance of books in its training set: "We use the BooksCorpus dataset for training the language model. . . . Crucially, it contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information."

2. The Largescale Unlicensed Copying of Nonfiction Books to Train the GPT Models

58. OpenAI and Microsoft created unlicensed reproductions of copyrighted works owned by Plaintiff and the putative class in the course of training and fine-tuning their models.

59. A massive quantity of copyrighted works, including copyrighted nonfiction works, were included in the "training set" that OpenAI used to develop GPT-3, GPT-3.5, and GPT-4. The training set for GPT-3, GPT-3.5, and GPT-4 was enormous. It included 45 terabytes of data—i.e., several billion pages of single-spaced text—and was comprised of datasets called Common Crawl, WebText2, Books1 and Books2, and Wikipedia.

60. The Common Crawl dataset is a "copy of the Internet" made available by an eponymous 501(c)(3) organization that was created and managed by wealthy technology investors. WebText2 was a dataset created by OpenAI that includes text from the internet, primarily from

Reddit posts.

61. Books1 and Books2 comprised around 15 percent of the training set for GPT-3, and was also used for training GPT-3.5, and GPT-4. OpenAI has described these datasets as "internet-based books corpora"—in other words, a large mass of books that OpenAI obtained from internet sources. In total, the Books1 and Books2 datasets are approximately 100 million pages of single-spaced text.

62. Though OpenAI has gone to great lengths to conceal the contents of its training datasets—especially Books2—what is known about the training data indicates that OpenAI's GPT models were trained on a mass of copyrighted books and other copyrighted material.

63. For one, OpenAI has publicly acknowledged that its models were trained on "large, publicly available datasets that include copyrighted works."[12] OpenAI has also admitted that its training process "necessarily involves first making copies of the data to be analyzed," including the large volume of copyrighted works in its dataset.

64. Furthermore, a recent academic study by researchers at the University of California at Berkeley tested whether the GPT-4 model was capable of exhibiting "memorization," i.e., returning exact passages, of a number of popular (and copyrighted) fiction books. If passages of a book are memorized, then it is likely the results showed that hundreds of copyrighted books were memorized in the models. The research confirmed that GPT-4 had memorized hundreds of copyrighted books.

65. OpenAI has since re-calibrated ChatGPT to avoid divulging the details of its training dataset and the extent of its copyright infringement.

66. In the early days after its release, however, ChatGPT, in response to an inquiry,

[12] Christopher T. Zirpoli, Cong. Rsch. Serv., LSB10922, Generative Artificial Intelligence and Copyright Law 3 (2023).

confirmed: "Yes, Julian Sancton's book 'Madhouse at the End of the Earth' is included in my training data." OpenAI has acknowledged that material that was incorporated in GPT-3 and GPT-4's training data was copied during the training process.

67. Upon information and belief, in the course of the training process, Defendants made hundreds of copies of copyrighted content owned by Plaintiff and/or the Class. In order to calibrate the GPT models to produce human-like expression, OpenAI and Microsoft collaborated to develop a complex, bespoke supercomputing system that was made to house and reproduce copies of the training dataset. Millions of copyrighted works were copied—including at least tens of thousands of nonfiction books—and then ingested for the purpose of "training" Defendants' GPT models. Those works were used as inputs into the GPT models, then copied several times again for the purpose of gauging how well the output mimicked human expression—that is, mimicked the style, expression, and content of the copyrighted works that the GPT models exploited in the calibration process.

68. Thousands, maybe more, copyrighted works—including nonfiction books—were then used for the purposes of fine-tuning the models. For fine tuning, OpenAI and Microsoft used their supercomputing systems to generate even more unlicensed copies, as OpenAI's personnel copied works—including nonfiction books—to use as inputs to the model to test the quality of the outputs, then re-calibrate the weights of the GPT model to better mimic—and in many cases, to generate outright copies—the content, expression, and style reflected in the training data.

69. While OpenAI was responsible for designing the calibration and fine-tuning of the GPT models—and thus, the largescale copying of this copyrighted material involved in generating a model programmed to accurately mimic Plaintiff's and others' styles—Microsoft built and operated the computer system that enabled this unlicensed copying in the first place.

70. Upon information and belief, Microsoft and OpenAI continue to make largescale, unlicensed copies to calibrate and fine-tune GPT-3, GPT-3.5, and GPT-4, and forthcoming generations of the GPT models, like GPT-5.

71. Defendants' commercial copying of Plaintiff's work and works owned by the proposed Class was manifestly unfair use, for several reasons. For starters, even by OpenAI's own description, the use is of the same kind and purpose that an ordinary reading consumer may use a book—to review the expression in it, that is, the order of words, presentation of facts, and syntax, among others. OpenAI has suggested that it uses the training data to "learn" how words and concepts fit to together, much in the way a human learns. While OpenAI's anthropomorphizing of its models is up for debate, at a minimum, humans who learn from books buy them, or borrow them from libraries that buy them, providing at least some measure of compensation to authors and creators. OpenAI does not, and it has usurped authors' content for the purpose of creating a machine built to generate the very type of content for which authors would usually be paid.

72. Even OpenAI has acknowledged that its use is unfair to creators. In his testimony before the Senate, former OpenAI CEO and current Microsoft employee Sam Altman admitted that "creators deserve control over how their creations are used, and what happens sort of beyond the point of releasing it into the world" and that "creators, content owners need to benefit from this technology." Yet OpenAI has given creators and copyright owners zero control over how their works are used in the training process—and zero compensation for it.

73. OpenAI and Microsoft's appropriation of nonfiction works is especially egregious. Nonfiction books take an enormous personal investment of time and resources. Nonfiction authors often dedicate countless hours to poring over primary source material or interviewing witnesses. Case in point: Plaintiff Sancton over the course of the five years he dedicated to writing *Madhouse*

at the End of the Earth, traveled across the world to inspect remote icescapes and uncover photographs from the Belgian Antarctic Expedition of 1898. Plaintiff's Sancton's experience is hardly unique. The labor of nonfiction authors also delivers a substantial public benefit. Books like Mr. Sancton's provide a digestible presentation of facts based on research and evidence. OpenAI and Microsoft, in reproducing and training their AI models off of these nonfiction works for free, have unfairly appropriated the fruits of nonfiction authors' labor.

74. OpenAI, in taking authors' works without compensation, has deprived authors of books sales and licensing revenues. There is, and has been, an established market for the sale of books and e-books, yet OpenAI ignored it and chose to scrape a massive corpus of copyrighted books from the internet, without even paying for an initial copy. OpenAI has also usurped a licensing market for copyright owners. In the short time since ChatGPT's release, there has been significant evidence that a licensing market is likely to be—and has been—developing for AI training datasets. Case in point: OpenAI itself has reached deals with content creators like the Associated Press and, on information and belief, others, in connection with the use of their copyrighted content in AI training. Similarly, OpenAI has admitted that it "paid for" training data from certain third parties, yet in the case of Plaintiff and the proposed Class, OpenAI has not compensated them at all.

75. There is also substantial reason to believe that, without OpenAI's largescale copyright infringement, blanket licensing practices would be possible through clearinghouses, like the Copyright Clearance Center, which has announced its own role in helping develop legal mechanisms for the licensed use of copyrighted material for AI training.[13]

76. OpenAI, however, has chosen to use Plaintiff Sancton's works and the works

[13] Copyright Clearance Center, The Intersection of AI & Copyright, (last visited Nov. 20, 2023), https://www.copyright.com/resource-library/insights/intersection-ai-copyright/.

owned by the proposed Class free of charge, and in doing so has harmed the market for the copyrighted works by depriving them of book sales and licensing revenue.

III. Defendants Have Profited From Their Unlicensed Exploitation of Copyrighted Material At the Expense of Authors

77. Microsoft and OpenAI have enjoyed substantial commercial gain from their GPT-based commercial offerings, including ChatGPT Plus, ChatGPT Enterprise, Bing Chat, and the licensing of the OpenAI API for businesses seeking to develop their own generative AI systems built on top of GPT-3, GPT-3.5, or GPT-4.

78. As of November 2023, ChatGPT has reported over 100 million weekly active users. Included among those users are 92% of all Fortune 500 companies.[14] OpenAI has generated revenue through its subscription services, ChatGPT Plus ($20/month) and its business-focused ChatGPT Enterprise. OpenAI is currently generating revenue of more than $100 million per month, on pace for $1.3 billion per year.

79. Microsoft has also reaped the benefits from its investment and development of ChatGPT. Since incorporating GPT-3 into its Bing search engine, Bing surpassed more than 100 million daily active users for the first time in its history. That surge was in large part attributable to the incorporation of OpenAI's GPT models, as large percentage of Bing's new users are using Bing Chat daily.

80. Microsoft has also been integrating ChatGPT into Azure and Office 365 products, and charging add-on fees for users seeking to take advantage of generative AI offerings. Microsoft Teams is charging an additional license for use of AI features. Microsoft has also unveiled a GPT-4-powered product called Microsoft 365 Copilot, which, according to Microsoft, "combines the

[14] Aisha Malik, OpenAI's ChatGPT Now Has 100 Million Weekly Active Users, (last visited Nov. 20, 20230, https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/.

power of large language models (LLMs) with your data in the Microsoft Graph and the Microsoft 365 apps to turn your words into the most powerful productivity tool on the planet." Microsoft Copilot is $30 per month. Analysts project that the integration of GPT into Microsoft products could generate more than $10 billion in annualized revenue by 2026, [15] with just one version of this integration—"GitHub Copilot"—already generating more than $100 million in annual recurring revenue.[16]

CLASS ALLEGATIONS

81. This action is brought by Plaintiff individually and on behalf of a class pursuant to Rule 23(b)(3) 23(b)(1) of the Federal Rules of Civil Procedure. The Class consists of:

> All owners of copyrighted literary works that: (a) are registered with the United States Copyright Office; (b) were or are used by Defendants in training their generative artificial intelligence models, including but not limited to GPT-3, GPT-3.5, GPT-4, and GPT-5; and (c) are works of nonfiction and either are or have been assigned an International Standard Book Number (ISBN) or are published in an academic journal. The Class excludes Defendants, their officers and directors, members of their immediate families, and the heirs, successors or assigns of any of the foregoing.

82. The Class consists of at least thousands of authors and copyright holders and thus is so numerous that joinder of all members is impractical. The identities of members of the Class can be readily ascertained from business records maintained by Defendants.

83. The claims asserted by Plaintiff are typical of the claims of the Class, all of whose works were also copied as part of the GPT training process.

---

[15] Novet, *supra* Note 3, (last visited Nov. 20, 2023).
[16] Holmes, *supra* Note 4, (last visited Nov. 20, 2023).

84. The Plaintiff will fairly and adequately protect the interests of the Class and does not have any interests antagonistic to those of other members of the Class.

85. The Plaintiff has retained attorneys who are knowledgeable and experienced in copyright and class action matters, as well as complex litigation.

86. Plaintiff requests that the Court afford Class members notice and the right to opt-out of any Class certified in this action.

87. This action is appropriate as a class action pursuant to Rule 23(b)(3) of the Federal Rules of Civil Procedure because common questions of law and fact affecting the Class predominate over those questions affecting only individual members. Those common questions include:

a. Whether Defendants' reproduction of the Class's copyrighted work constituted copyright infringement;

b. Whether Defendants' reproduction of the Class's copyrighted work in the course of training their generative AI models was fair use;

c. Whether Defendants' reproduction of the Class's copyrighted work harmed Class member and whether Class member is entitled to damages, including statutory damages and the amount of statutory damages;

d. Whether Defendant Microsoft substantially facilitated the copyright infringement committed by OpenAI; and

e. Whether Defendants Microsoft knew or should have known that OpenAI was making copies of copyrighted content, including the Class's copyrighted works, in the course of training their AI models.

88. This action is also appropriate as a class action pursuant to Rule 23(b)(1) of the

Federal Rules of Civil Procedure because prosecution of separate actions by individual class members risks inconsistent adjudication and because the resolution of claims for individual class members may be dispositive of the actions of other class members.

COUNT I: Copyright Infringement (17 U.S.C. § 501)

Against OpenAI and Microsoft

89. Plaintiff incorporates by reference the allegations in Paragraphs 1 to 88 as though fully set forth herein.

90. Plaintiffs and members of the proposed Class own the registered copyrights in the works that Defendants reproduced and appropriated to train their artificial intelligence models.

91. Plaintiff and members of the proposed Class therefore hold the exclusive rights, including the rights of reproduction and distribution, to those works under 17 U.S.C. § 106.

92. Defendants infringed on the exclusive rights, under 17 U.S.C. § 106, of Plaintiff and members of the proposed Class by, among other things, reproducing the works owed by Plaintiff and the proposed Class in datasets used to train their artificial intelligence models.

93. On information and belief, Defendants' infringing conduct alleged herein was and continues to be willful. Defendants infringed on the exclusive rights of Plaintiff and members of the proposed Class knowing that they were profiting from mass copyright infringement.

94. Plaintiff and members of the proposed Class are entitled to statutory damages, actual damages, disgorgement, and other remedies available under the Copyright Act.

95. Plaintiff and members of the proposed Class have been and continue to be irreparably injured due to Defendants' conduct, for which there is no adequate remedy at law. Defendants will continue to infringe on the exclusive right of Plaintiff and the proposed class unless their infringing activity is enjoined by this Court. Plaintiffs are therefore entitled to permanent injunctive relief barring Defendants' ongoing infringement.

COUNT II: Contributory Infringement

Against Microsoft, OpenAI Inc., OpenAI GP LLC, OpenAI Global LLC, OpenAI LLC, OAI Corporation LLC, and OpenAI Holdings LLC

96. Plaintiff incorporates by reference and realleges the allegations in Paragraphs 1 to 94 as though fully set forth herein.

97. Microsoft materially contributed and facilitated OpenAI's direct infringement alleged in Count I by providing billions of dollars in investments and designing, creating, and maintaining the bespoke supercomputing system that OpenAI used to maintain and copy the copyrighted works owned by Plaintiff and the proposed Class. This assistance was necessary for OpenAI to perpetrate the largescale copyright infringement alleged herein.

98. Microsoft knew, or had reason to know, of the direct infringement alleged in Count I because OpenAI, upon information and belief, informed Microsoft as part of the due diligence process that it was copying and scraping copyrighted material in order to train its generative artificial intelligence models. Furthermore, in the course of designing and maintain its bespoke supercomputing system, Microsoft became aware of OpenAI's direct infringement and directly assisted the copying of copyrighted content owned by Plaintiff and the proposed Class.

99. Microsoft profited from its OpenAI's direct infringement through its investment in OpenAI and its monetization of GPT-based products.

100. Microsoft is liable for contributing to the direct infringement alleged in Count I.

101. Microsoft is liable for contributing to the direct infringement alleged in Count I.

102. OpenAI Inc., OpenAI GP LLC, OpenAI Global LLC, OpenAI LLC, OAI Corporation LLC, and OpenAI Holdings LLC, each directly and indirectly control, direct, and manage other OpenAI entities, including OpenAI OpCo LLC, that are and were responsible for the direct infringement alleged in Count I. These OpenAI entities, through their direction and

control of other OpenAI entities, were aware of the direct infringement perpetrated by a variety of OpenAI entities, as alleged in Count I. These OpenAI entities profited from the infringement perpetrated by OpenAI as a whole through their ownership of other OpenAI entities.

103. OpenAI Inc., OpenAI GP LLC, OpenAI Global LLC, OpenAI LLC, OAI Corporation LLC, and OpenAI Holdings LLC, are each liable for contributing to the direct infringement alleged in Count I

PRAYER FOR RELIEF

WHEREFORE, Plaintiff demands judgment against each Defendant as follows:

1. Declaring this action can be properly maintained pursuant to Rule 23 of the Federal Rules of Civil Procedure;

2. Awarding Plaintiff and the proposed Class statutory damages, compensatory damages, restitution, disgorgement, and any other relief that may be permitted by law or equity pursuant to the first and second claims for relief;

3. Permanently enjoining Defendants from the infringement and contributory infringement alleged herein;

4. Awarding Plaintiff and the proposed Class pre-judgment and post-judgment interest pursuant to the first and second claims for relief;

5. Awarding Plaintiff and the proposed Class costs, expenses, and attorneys' fees as permitted by law; and

6. Awarding Plaintiff and the proposed Class further relief as the Court may deem just and proper under the circumstances.

DEMAND FOR JURY TRIAL

Pursuant to Rule 38 of the Federal Rules of Civil Procedure, Plaintiff hereby demands a jury trial for all claims so triable.

Dated: November 21, 2023

/s/ J. Craig Smyser

J. Craig Smyser
SUSMAN GODFREY L.L.P
1301 Avenue of the Americas, 32nd Floor
New York, NY 10019
Telephone: (212) 336-8330
Facsimile: (212) 336-8340
csmyser@susmangodfrey.com

Justin A. Nelson (*pro hac vice forthcoming*)
Alejandra C. Salinas (*pro hac vice forthcoming*)
SUSMAN GODFREY L.L.P
1000 Louisiana Street, Suite 5100
Houston, TX 77002-5096
Telephone: (713) 651-9366
Facsimile: (713) 654-6666
jnelson@susmangodfrey.com
asalinas@susmangodfrey.com

Rohit D. Nath (*pro hac vice forthcoming*)
SUSMAN GODFREY L.L.P
1900 Avenue of the Stars, Suite 1400
Los Angeles, CA 90067-2906
Telephone: (310) 789-3100
Facsimile: (310) 789-3150
rnath@susmangodfrey.com

Attorneys for Plaintiff and the Proposed Class

ClassAction.org
