UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

AUTHORS GUILD, DAVID BALDACCI, MARY BLY, MICHAEL CONNELLY, SYLVIA DAY, JONATHAN FRANZEN, JOHN GRISHAM, ELIN HILDERBRAND, CHRISTINA BAKER KLINE, MAYA SHANBHAG LANG, VICTOR LAVALLE, GEORGE R.R. MARTIN, JODI PICOULT, DOUGLAS PRESTON, ROXANA ROBINSON, GEORGE SAUNDERS, SCOTT TUROW, and RACHEL VAIL, individually and on behalf of others similarly situated,

Plaintiffs,

v.

OPEN AI INC., OPENAI OPCO LLC, OPENAI GP LLC, OPENAI, LLC, OPENAI GLOBAL LLC, OAI CORPORATION LLC, OPENAI HOLDINGS LLC, OPENAI STARTUP FUND I LP, OPENAI STARTUP FUND GP I LLC, OPENAI STARTUP FUND MANAGEMENT LLC, and MICROSOFT CORPORATION,

Defendants.

JONATHAN ALTER, KAI BIRD, TAYLOR BRANCH, RICH COHEN, EUGENE LINDEN, DANIEL OKRENT, JULIAN SANCTON, HAMPTON SIDES, STACY SCHIFF, JAMES SHAPIRO, JIA TOLENTINO, and SIMON WINCHESTER, on behalf of themselves and all others similarly situated,

Plaintiffs,

v.

OPEN AI INC., OPENAI OPCO LLC, OPENAI GP LLC, OPENAI, LLC, OPENAI GLOBAL LLC, OAI CORPORATION LLC, OPENAI HOLDINGS LLC, OPENAI STARTUP FUND I

ECF CASE

No. 1:23-cv-08292-SHS;
No. 1:23-cv-10211-SHS

FIRST CONSOLIDATED CLASS ACTION COMPLAINT

JURY TRIAL DEMANDED

LP, OPENAI STARTUP FUND GP I LLC, OPENAI STARTUP FUND MANAGEMENT LLC, and MICROSOFT CORPORATION,

Defendants.

Plaintiffs Jonathan Alter, The Authors Guild, David Baldacci, Kai Bird, Mary Bly, Taylor Branch, Rich Cohen, Michael Connelly, Sylvia Day, Jonathan Franzen, John Grisham, Elin Hilderbrand, Christina Baker Kline, Maya Shanbhag Lang, Victor LaValle, Eugene Linden, George R.R. Martin, Daniel Okrent, Jodi Picoult, Douglas Preston, Roxana Robinson, Julian Sancton, George Saunders, Stacy Schiff, Hampton Sides, James Shapiro, Jia Tolentino, Scott Turow, Simon Winchester, and Rachel Vail, on behalf of themselves and all other similarly situated (the "Fiction Author Class" and "Nonfiction Author Class," as defined below), for their complaint against Defendants OpenAI, Inc., OpenAI GP LLC, OpenAI, LLC, OpenAI OpCo LLC, OpenAI Global LLC, OAI Corporation, LLC, OpenAI Holdings, LLC, OpenAI Startup Fund I LP, OpenAI Startup Fund GP I LLC, OpenAI Startup Fund Management LLC (collectively "OpenAI"), and Microsoft Corporation (all collectively "Defendants"), allege as follows:

NATURE OF THE CASE

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 (such as when a person buys a book), they choose to operate as if the laws protecting copyright do not exist. Yet the United States Constitution protects the fundamental principle that creators, like authors, deserve compensation for their works, and the Copyright Act grants "a bundle of exclusive rights" to creators, 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).

2. Plaintiffs, authors of a broad array of works of fiction and nonfiction, bring this action under the Copyright Act seeking redress for Defendants’ flagrant and harmful infringements of Plaintiffs’ registered copyrights. Defendants copied Plaintiffs’ works and then fed them into their “large language models” or “LLMs,” algorithms designed to output human-seeming text responses to users’ prompts and queries. These algorithms are at the heart of Defendants’ massive commercial enterprise. And at the heart of these algorithms is systematic theft on a mass scale.

3. Defendants’ LLMs endanger authors’ ability to make a living, in that the LLMs allow anyone to generate—automatically and freely (or very cheaply)—texts that they would otherwise pay writers to create. Moreover, the LLMs can spit out derivative works: material that is based on, mimics, summarizes, or paraphrases Plaintiffs’ works, and harms the market for them. Without Plaintiffs’ copyrighted works on which to “train” their LLMs, Defendants would have no commercial product with which to damage—if not usurp—the market for these professional writers’ works. OpenAI’s willful copying thus makes Plaintiffs’ works into engines of Plaintiffs’ own destruction, notwithstanding that professional writers often spend years conceiving of and writing their creations.

4. Specifically, 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 "trained," in Defendants' parlance, by reproducing a massive corpus of copyrighted material, including, upon information and belief, tens or hundreds of thousands of fiction and 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 texts to exploit precisely what the Copyright Act was designed to protect: the elements of protectible expression within them, like the choice and order of words and sentences, syntax, flow, themes, and paragraph and story structure. 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] In other words, by training its models on certain works, OpenAI copied the works' expression so that the models could memorize, mimic, and paraphrase that expression.

6. Defendants copied and data-mined the works of writers, without permission or 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 on which to feed, there would be no ChatGPT. Defendants' commercial success was possible only because they copied and digested the

[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.

protected, copyrightable expression contained in billions of pages of actual text, across millions of copyrighted works—all without paying a penny to the authors of those works.

7. OpenAI could have "trained" its LLMs on works in the public domain. It could have paid a reasonable licensing fee to use copyrighted works. What Defendants could *not* do was evade the Copyright Act altogether to power their lucrative commercial endeavor, taking whatever datasets of relatively recent books they could get their hands on without authorization. There is nothing fair about this. OpenAI's unauthorized use of Plaintiffs' copyrighted works thus presents a straightforward infringement case applying well-established law to well-recognized copyright harms.

8. OpenAI's chief executive Sam Altman has told Congress that he shares Plaintiffs' concerns. According to Altman, "Ensuring that the creator economy continues to be vibrant is an important priority for OpenAI. ... OpenAI does not want to replace creators. We want our systems to be used to empower creativity, and to support and augment the essential humanity of artists and creators."[2] Altman testified that OpenAI "think[s] that creators deserve control over how their creations are used" and that "content creators, content owners, need to benefit from this technology."[3] Altman also has represented that OpenAI has "licens[ed] content directly from content owners" for "training" purposes.[4] Not so from Plaintiffs. As to them, Altman and Defendants have proved unwilling to turn these words into actions.

---

[2] Sam Altman, *Questions for the Record*, at 9–10 (June 22, 2023), *available at* https://www.judiciary.senate.gov/imo/media/doc/2023-05-16_-_qfr_responses_-_altman.pdf (last accessed Dec. 4, 2023).

[3] *Oversight of A.I.: Rules for Artificial Intelligence: Hearing Before the S. Judiciary Comm. Subcomm. on Privacy, Tech. and the Law*, 118th Cong. (2023) (testimony of OpenAI CEO Sam Altman), *available at* https://techpolicy.press/transcript-senate-judiciary-subcommittee-hearing-on-oversight-of-ai (last accessed Dec. 4, 2023).

[4] Altman, *Questions for the Record*, *supra*, at 10.

9. Defendants OpenAI and Microsoft have enjoyed enormous financial gain from their free exploitation of copyrighted material. OpenAI recently reported that it is "generating revenue at a pace of $1.3 billion a year."[5] Microsoft, for its part, has seen its investment in OpenAI increase 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,[6] with just one version of this integration—"GitHub Copilot"—already generating more than $100 million in annual recurring revenue.[7] 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.

10. OpenAI and Microsoft's commercial gain has come at the expense of creators and rightsholders like Plaintiffs and members of the Classes. 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

[5] 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.

[6] 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.

[7] 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).

expression—from shortform articles to book chapters—that mimics the syntax, voice, and themes of the copyrighted works on which it was trained.

11. Plaintiffs seek to represent Classes of fiction and nonfiction writers whose works were used to train Defendants' artificial intelligence models. Plaintiffs, on behalf of themselves and the Classes, seek damages from Defendants for their largescale infringement of their copyrighted works, as well as injunctive relief.

JURIDISCTION & VENUE

12. 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*.

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

14. OpenAI and Microsoft's copyright infringement and contributory copyright infringement occurred, in substantial part, 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.

15. 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, which were 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 of OpenAI's GPT models.

16. Plaintiffs The Authors Guild, Bird, Bly, Kline, Lang, LaValle, Linden, Robinson, Sancton, Schiff, Shapiro, Tolentino, Vail, and Winchester are citizens of New York. The injuries alleged here from Defendants' infringement occurred in this District.

17. 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 Plaintiffs' claims occurred in this District, including the sales of Defendants' GPT-based products within this District.

THE PARTIES

A. Plaintiffs

18. Plaintiff The Authors Guild is a nonprofit 501(c)(6) organization based in New York, New York.

19. Plaintiff Jonathan Alter is an author and a resident of New Jersey.

20. Plaintiff David Baldacci is an author and a resident of Virginia.

21. Plaintiff Kai Bird is an author and a resident of New York.

22. Plaintiff Mary Bly is an author and a resident of New York.

23. Plaintiff Taylor Branch is an author and a resident of Maryland.

24. Plaintiff Rich Cohen is an author and a resident of Connecticut.

25. Plaintiff Michael Connelly is an author and a resident of Florida.

26. Plaintiff Sylvia Day is an author and a resident of Nevada.

27. Plaintiff Jonathan Franzen is an author and a resident of California.

28. Plaintiff John Grisham is an author and a resident of Virginia.

29. Plaintiff Elin Hilderbrand is an author and a resident of Massachusetts.

30. Plaintiff Christina Baker Kline is an author and a resident of New York.

31. Plaintiff Maya Shanbhag Lang is an author and a resident of New York.

32. Plaintiff Victor LaValle is an author and a resident of New York.

33. Plaintiff Eugene Linden is an author and a resident of New York.

34. Plaintiff George R.R. Martin is an author and a resident of New Mexico.

35. Plaintiff Daniel Okrent is an author and a resident of Massachusetts.

36. Plaintiff Jodi Picoult is an author and a resident of New Hampshire.

37. Plaintiff Douglas Preston is an author and a resident of New Mexico.

38. Plaintiff Roxana Robinson is an author and a resident of New York.

39. Plaintiff Julian Sancton is an author and a resident of New York.

40. Plaintiff George Saunders is an author and a resident of California.

41. Plaintiff Stacy Schiff is an author and a resident of New York.

42. Plaintiff Hampton Sides is an author and a resident of New Mexico.

43. Plaintiff James Shapiro is an author and a resident of New York.

44. Plaintiff Jia Tolentino is an author and a resident of New York.

45. Plaintiff Scott Turow is an author and a resident of Florida.

46. Plaintiff Simon Winchester OBE is an author and a resident of New York.

47. Plaintiff Rachel Vail is an author and a resident of New York.

48. The registration information for the infringed works of Plaintiffs is identified in Exhibits A and B to this Complaint.

B. <u>OpenAI Defendants</u>

49. 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.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

56. Defendant OpenAI Startup Fund I LP is a limited partnership formed under the laws of Delaware with its principal place of business in San Francisco, California.

57. Defendant OpenAI Startup Fund GP I LLC is a limited liability company formed under the laws of Delaware with its principal place of business in San Francisco, California.

58. Defendant OpenAI Startup Fund Management LLC is a limited liability company formed under the laws of Delaware with its principal place of business in San Francisco, California.

C. <u>Microsoft</u>

59. 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

A. Generative AI and Large Language Models

60. The terms "artificial intelligence" or "AI" refer generally to computer systems designed to imitate human cognitive functions.

61. The terms "generative artificial intelligence" or "generative AI" refer specifically to systems that are capable of generating "new" content in response to user inputs called "prompts."

62. For example, the user of a generative AI system capable of generating images from text prompts might input the prompt, "A lawyer working at her desk." The system would then attempt to construct the prompted image. Similarly, the user of a generative AI system capable of generating text from text prompts might input the prompt, "Tell me a story about a lawyer working at her desk." The system would then attempt to generate the prompted text.

63. Recent generative AI systems designed to recognize input text and generate output text are built on "large language models" or "LLMs."

64. OpenAI's LLMs 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.

65. 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.

66. "Training" an LLM refers to the process by which the parameters that define an LLM's behavior are adjusted through the LLM's ingestion and analysis of large "training" datasets.

67. The "training" of an LLM requires inputting large numbers of parameters in the model and then supplying the LLM with large amounts of text for the LLM to ingest—the more text, the better. That is, in part, and hand-in-hand with number of model specifications, the *large* in *large language model*.

68. As the U.S. Patent and Trademark Office has observed, LLM "training" "almost by definition involve[s] the reproduction of entire works or substantial portions thereof."[8]

69. "Training" in this context is therefore a technical-sounding euphemism for "copying and ingesting expression."

70. Moreover, in some form and to some degree currently unknowable to the public, OpenAI's LLMs have "memorized" or stored their "training" data (even if in a "translated"

---

[8] U.S. Patent & Trademark Office, *Public Views on Artificial Intelligence and Intellectual Property Policy* 29 (2020), *available at* https://www.uspto.gov/sites/default/files/documents/USPTO_AI-Report_2020-10-07.pdf (last accessed Jan. 22, 2024).

form, such as a unique statistical profile), such that the data (at least in part) can be accessed, recalled, and reproduced by the LLM at will.[9]

71. The quality of the LLM (that is, its capacity to generate human-seeming responses to prompts) is dependent on the quality of the datasets used to "train" the LLM.

72. Professionally authored, edited, and published fiction and nonfiction books—such as those authored by Plaintiffs here—are an especially important source of LLM "training" data.

73. As one group of AI researchers (not affiliated with Defendants) has observed, "[b]ooks are a rich source of both fine-grained information, how a character, an object or a scene looks like, as well as high-level semantics, what someone is thinking, feeling and how these states evolve through a story."[10]

74. Once the "training" data is ingested, OpenAI can then control how closely the LLMs' outputs adhere to probability. Software engineers refer to this parameter as "temperature." If OpenAI engineers set its LLMs at a "hotter" temperature, the model will bias *against* what it calculates as the most probable response in favor of more random outputs. Likewise, the "cooler" the LLMs are set, the more closely their outputs will adhere to statistical probability. In this way, OpenAI can control the very perception of copying.

75. LLMs can also be fine-tuned after "training" by adjusting the parameters to perform specific tasks.

76. Finally, and as discussed below, OpenAI continues to produce new versions of their LLMs supported by increasingly more "training" data, much of which is copyrighted

---

[9] *See* Jason Koebler, *Google Researchers' Attack Prompts ChatGPT to Reveal Its Training Data*, 404 Media (Nov. 29, 2023), *available at* https://www.404media.co/google-researchers-attack-convinces-chatgpt-to-reveal-its-training-data/ (last accessed Jan. 22, 2024); Kent K. Chang *et al.*, *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4* (2023), *available at* https://arxiv.org/pdf/2305.00118v1.pdf (last accessed Jan. 22, 2024).

[10] Yukun Zhu *et al.*, *Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and Reading Books* 1 (2015), *available at* https://arxiv.org/pdf/1506.06724.pdf (last accessed Jan. 22, 2024).

material. These new versions hone the performance of OpenAI's consumer-facing products through the ingestion of more copyrighted material.

77. In other words, books are the high-quality materials Defendants want, need, and have therefore outright pilfered to develop generative AI products that produce high-quality results: text that appears to have been written by a human writer.

78. This use is highly commercial.

B. <u>OpenAI's Willful Infringement of Plaintiffs' Copyrights</u>

1. OpenAI

79. OpenAI (specifically, Defendant OpenAI Inc.) was founded in 2015 as a non-profit organization with the self-professed goal of researching and developing AI tools "unconstrained by a need to generate financial return."

80. Four years later, in 2019, OpenAI relaunched itself (specifically, through Defendant OpenAI GP LLC and Defendant OpenAI OpCo LLC) as a for-profit enterprise.

81. Investments began pouring in. Microsoft Corporation, one of the world's largest technology companies, invested $1 billion in 2019, an estimated $2 billion in 2021, and a staggering $10 billion in 2023, for a total investment of $13 billion.

82. Industry observers currently value OpenAI at up to $80 billion.

2. GPT-N and ChatGPT

83. OpenAI's LLMs are collectively referred to as "GPT-N," which stands for "Generative Pre-trained Transformer" (a specific type of LLM architecture), followed by a version number.

84. GPT-3 was released in 2020 and exclusively licensed to Microsoft the same year.

85. OpenAI further refined GPT-3 into GPT-3.5, which was released in 2022.

86. In November 2022, OpenAI released ChatGPT, a consumer-facing chatbot application built on GPT-3.5.

87. ChatGPT's popularity exploded virtually overnight. By January 2023, less than three months after its release, the application had an estimated 100 million monthly active users, making it one of the fastest-growing consumer applications in history.

88. Today, the application is estimated to have more than 180 million users.[11]

89. GPT-4, the successor to GPT-3.5, was released in March 2023.

90. GPT-4 underlies OpenAI's new subscription-based chatbot, called ChatGPT Plus, which is available to consumers for $20 per month.

91. Defendants intend to earn billions of dollars from this technology.

92. When announcing the release of ChatGPT Enterprise, a subscription-based high-capability GPT-4 application targeted for corporate clients, in August 2023, OpenAI claimed that teams in "over 80% of Fortune 500 companies" were using its products.[12]

93. GPT-4 also underlies Microsoft's Bing Chat product, offered through its Bing Internet search engine, and is integrated into its sales and marketing software, coding tools, productivity software, and cloud storage services.

94. OpenAI's annualized revenue reached over $1.6 billion in 2023 due, in large part, to strong growth from its ChatGPT product.[13]

---

[11] Anna Tong, *Exclusive: ChatGPT traffic slips again for third month in a row*, Reuters (Sept. 7, 2023) *available at* https://www.reuters.com/technology/chatgpt-traffic-slips-again-third-month-row-2023-09-07/ (last accessed Jan. 24, 2024).

[12] OpenAI, *Introducing ChatGPT Enterprise* (Aug. 28, 2023), *available at* https://openai.com/blog/introducing-chatgpt-enterprise (last accessed Jan. 24, 2024).

[13] Maria Heeter et al., *OpenAI Annualized Revenue Tops $1.6 Billions as Customers Shrug Off CEO Drama*, The Information (Dec. 30, 2023), *available at* https://www.theinformation.com/articles/openais-annualized-revenue-tops-1-6-billion-as-customers-shrug-off-ceo-drama (last accessed Jan. 22, 2024).

95. Analysts estimate that Microsoft could earn more than $10 billion in annual revenue by 2026 *only* from AI add-ons to its Microsoft 365 productivity software, "at the core" of which lies OpenAI technology.[14]

3. Knowingly "Training" GPT-N on Copyrighted Books

96. OpenAI does not disclose or publicize with specificity what datasets GPT-3, GPT-3.5, or GPT-4 were "trained" on. Despite its name, OpenAI treats that information as proprietary.

97. To "train" its LLMs—including GPT-3, GPT-3.5, and GPT-4—OpenAI has reproduced copyrighted books—including copyrighted books authored by Plaintiffs here—without their authors' consent.

98. OpenAI has admitted as much.

99. OpenAI has admitted that it has "trained" its LLMs on "large, publicly available datasets that include copyrighted works."[15]

100. OpenAI has admitted that "training" LLMs "require[s] large amounts of data," and that "analyzing large corpora" of data "necessarily involves first making copies of the data to be analyzed."[16]

101. OpenAI has admitted that, if it refrained from using copyrighted works in its LLMs' "training," it would "lead to significant reductions in model quality."[17]

[14] Jordan Novet, *Microsoft starts selling AI tool for Office, which could generate $10 billion a year by 2026*, CNBC (Nov. 1, 2023), *available at* https://www.cnbc.com/2023/11/01/microsoft-365-copilot-becomes-generally-available.html (last accessed Jan. 22, 2024).

[15] OpenAI, *Comment Regarding Request for Comments on Intellectual Property Protection for Artificial Intelligence Innovation*, U.S. Patent and Trademark Office Dkt. No. PTO-C-2019-0038, at 1 (2019), *available at* https://www.uspto.gov/sites/default/files/documents/OpenAI_RFC-84-FR-58141.pdf (last accessed Jan. 22, 2024).

[16] *Id.*

[17] *Id.* at 7 n.33.

102. OpenAI thus has conceded that reproduction of copyrighted works is central to the quality of its products.

103. In response to a query submitted to it in January 2023, ChatGPT responded, "[i]t is possible that some of the books used to train me were under copyright. However, my training data was sourced from various publicly available sources on the internet, and it is likely that some of the books included in my training dataset were not authorized to be used. ... If any copyrighted material was included in my training data, it would have been used without the knowledge or consent of the copyright holder."

104. In another inquiry made that same year, ChatGPT confirmed the existence of Plaintiff Julian Sancton's book in its dataset: "Yes, Julian Sancton's book 'Madhouse at the End of the Earth' is included in my training data."

105. In another example, when prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Last Juror*, one of the Grisham Infringed Works, and titled the infringing and unauthorized derivative "The Juror's Dilemma," using the same characters from Grisham's existing book.[18]

106. Until recently, ChatGPT could be prompted to return quotations of text from copyrighted books with a good degree of accuracy, evidencing that the underlying LLM likely or must have ingested these books in their entireties during its "training."

107. Now, however, ChatGPT generally responds to such prompts with the statement, "I can't provide verbatim excerpts from copyrighted texts." Thus, while ChatGPT previously provided such excerpts and in principle retains the capacity to do so, it has been restrained from doing so, if only temporarily, by its programmers.

[18] Additional examples of OpenAI's plaintiff-specific infringement are detailed below.

108. Instead of "verbatim excerpts," ChatGPT offers to produce a summary of the copyrighted book, which usually contains details not available in reviews and other publicly available material—again suggesting that the underlying LLM must have ingested the entire book during its "training."

109. Notably, though, even when OpenAI fine-tunes a base model, it does not alter the fact of reproduction.

110. OpenAI is characteristically opaque about where and how it procured the entirety of these books, including Plaintiffs' copyrighted works.

111. OpenAI has discussed limited details about the datasets used to "train" GPT-3.

112. OpenAI admits that among the "training" datasets it used to "train" the model were "Common Crawl," and two "high-quality," "internet-based books corpora" which it calls "Books1" and "Books2."[19]

113. Common Crawl is a vast and growing corpus of "raw web page data, metadata extracts, and text extracts" scraped from billions of web pages. It is widely used in "training" LLMs, and has been used to "train," in addition to GPT-N, Meta's LlaMa, and Google's BERT. It is known to contain text from books copied from pirate sites.[20]

114. OpenAI refuses to discuss the source or sources of the Books2 dataset.

[19] Tom B. Brown *et al.*, *Language Models Are Few-Shot Learners* 8 (2020), *available at* https://arxiv.org/pdf/2005.14165.pdf (last accessed Jan. 22, 2024).

[20] Alex Hern, *Fresh Concerns Raised Over Sources of Training Material for AI Systems*, The Guardian (Apr. 20, 2023), *available at* https://www.theguardian.com/technology/2023/apr/20/fresh-concerns-training-material-ai-systems-facist-pirated-malicious (last accessed Jan. 22, 2024).

115. Some independent AI researchers suspect that Books2 contains or consists of ebook files downloaded from large pirate book repositories such as Library Genesis or "LibGen," "which offers a vast repository of pirated text."[21]

116. LibGen is already known to this Court as a notorious copyright infringer.[22]

117. Other possible candidates for Books2's sources include Z-Library, another large pirate book repository that hosts more than 11 million books, and pirate torrent trackers like Bibliotik, which allow users to download ebooks in bulk.

118. Websites linked to Z-Library appear in the Common Crawl corpus and have been included in the "training" dataset of other LLMs.[23]

119. Z-Library's Internet domains were seized by the FBI in February 2022, only months after OpenAI stopped "training" GPT-3.5 in September 2021.

120. The disclosed size of the Books2 dataset (55 billion "tokens," the basic units of textual meaning such as words, syllables, numbers, and punctuation marks) suggests it comprises over 100,000 books.

121. "Books3," a dataset compiled by an independent AI researcher, is comprised of nearly 200,000 books downloaded from Bibliotik, and has been used by other AI developers to "train" LLMs.

122. The similarities in the sizes of Books2 and Books3, and the fact that there are only a few pirate repositories on the Internet that allow bulk ebook downloads, strongly indicates that

---

[21] Kate Knibbs, *The Battle Over Books3 Could Change AI Forever*, Wired (Sept. 4, 2023), *available at* https://www.wired.com/story/battle-over-books3 (last accessed Jan. 22, 2024).

[22] *See Elsevier Inc. v. Sci-Hub*, No. 1:15-cv-4282-RWS (S.D.N.Y.).

[23] Kevin Schaul *et al.*, *Inside the Secret List of Websites that Make AI Like ChatGPT Sounds Smart*, The Washington Post (Apr. 19, 2023), *available at* https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning (last accessed Jan. 22, 2024).

the books contained in Books2 were also obtained from one of the notorious repositories discussed above.

123. OpenAI has not discussed the datasets used to "train" GPT-3.5, GPT-4, or their source or sources.

124. GPT-3.5 and GPT-4 are significantly more powerful than their predecessors. GPT-3.5 contains roughly 200 billion parameters, and GPT 4 contains roughly 1.75 trillion parameters, compared to GPT-3's roughly 175 billion parameters.

125. The growth in power and sophistication from GPT-3 to GPT-4 suggests a correlative growth in the size of the "training" datasets, raising the inference that one or more very large sources of pirated ebooks discussed above must have been used to "train" GPT-4.

126. There is no other way OpenAI could have obtained the volume of books required to "train" a powerful LLM like GPT-4.

127. In short, OpenAI admits it needs[24] and uses[25] "large, publicly available datasets that include copyrighted works"[26]—and specifically, "high-quality"[27] copyrighted books—to "train" its LLMs; pirated sources of such "training" data are readily available; and one or more of these sources contain Plaintiffs' works.

128. Defendants knew that their "training" data included texts protected by copyright but willfully proceeded without obtaining authorization.

129. OpenAI's "training" its LLMs could not have happened without Microsoft's financial and technical support. In 2020, Microsoft announced that it had developed Azure, "one of the top five publicly disclosed supercomputers in the world" which was "[b]uilt in

---

[24] OpenAI, *Comment Regarding Request for Comments*, *supra*, at 7 n.33.

[25] *Id.* at 2.

[26] *Id.* at 1.

collaboration with and exclusively for OpenAI," and "designed specifically to train that company's AI models."[28]

130. Furthermore, in 2023, Microsoft CEO Satya Nadella reminded the world that the "heavy lifting" for OpenAI's LLM "training" was done by Microsoft "compute infrastructure."[29]

C. The Microsoft-OpenAI Partnership

131. 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 Plaintiffs' and others' expression—Microsoft built and operated the computer system that enabled this unlicensed copying in the first place.

132. 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."

133. 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, and GPT-4 (and thus copy 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

---

[27] Brown *et al.*, *Few-Shot Learners*, *supra*, at 8.

[28] Jennifer Langston, *Microsoft announces new supercomputer, lays out vision for future AI work*, Microsoft (May 19, 2020), *available at* https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/ (last accessed Jan. 22, 2024).

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.

134. 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 in 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."

135. 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."

136. 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.

---

[29] CNBC, *First on CNBC: CNBC Transcript: Microsoft CEO Satya Nadella Speaks with CNBC's Jon Fortt on "Power Lunch" Today* (Fed. 7, 2023), *available at* https://www.cnbc.com/2023/02/07/first-on-cnbc-cnbc-transcript-microsoft-ceo-satya-nadella-speaks-with-cnbcs-jon-fortt-on-power-lunch-today.html (last accessed Jan. 22, 2024).

137. 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 books. 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 Plaintiffs' works and other copyrighted books—for the purpose of training the GPT models.

138. The large-scale copyright infringement would have been obvious to OpenAI and its business partners. Microsoft also became aware of OpenAI's largescale copyright infringement in the course of conducting the due diligence required for 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."[30] 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.

139. In addition to 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 Plaintiffs and the proposed Class. At Microsoft's largest partner event of the year, Inspire, Mr. Nadella said that while OpenAI is "innovating on the algorithms and the training of these frontier models, [Microsoft] innovate[s] on applications on

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

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.

140. Indeed, recent events have further demonstrated the close relationship between OpenAI and Microsoft. When OpenAI CEO Sam Altman was terminated, Microsoft hired him. In a November 2023 interview following the termination, Mr. Nadella stated: "We have all the IP rights and all the capability. If OpenAI disappeared tomorrow, I don't want any customer of ours to be worried about it quite honestly, because we have all of the rights to continue the innovation. Not just to serve the product, but we can go and just do what we were doing in partnership ourselves. We have the people, we have the compute, we have the data, we have everything."[31] Mr. Nadella continued, "And also this thing, it's not hands off, right? We are in there. We are below them, above them, around them. We do the kernel optimizations, we build tools, we build the infrastructure. So that's why I think a lot of the industrial analysts are saying, 'Oh wow, it's really a joint project between Microsoft and OpenAI.'"[32]

141. Shortly after this November 2023 interview, under pressure from Microsoft (and others), OpenAI reinstated Mr. Altman as CEO and granted Microsoft a nonvoting seat on the board of OpenAI, Inc.

D. <u>GPT-N's and ChatGPT's Harm to Authors</u>

142. ChatGPT and the LLMs underlying it seriously threaten the livelihood of the very authors—including Plaintiffs here, as discussed specifically below—on whose works they were "trained" without the authors' consent.

---

[31] Intelligencer Staff, *Satya Nadella on Hiring the Most Powerful Man in AI When OpenAI threw Sam Altman overboard, Microsoft's CEO saw an opportunity*, INTELLIGENCER (Nov. 21, 2023), available at https://nymag.com/intelligencer/2023/11/on-with-kara-swisher-satya-nadella-onhiring-sam-altman.html (last accessed Dec. 11, 2023).

[32] *Id.*

143. Goldman Sachs estimates that generative AI could replace 300 million full-time jobs in the near future, or one-fourth of the labor currently performed in the United States and Europe.

144. Already, writers report losing income from copywriting, journalism, and online content writing—important sources of income for many book authors. The Authors Guild's most recent author earnings study[33] shows a median writing-related income for full-time authors of just over $20,000, and that full-time traditional authors earn only half of that from their books. The rest comes from activities like content writing—work that is starting to dry up as a result of generative AI systems like ChatGPT.

145. An Authors Guild member who writes marketing and web content reported losing 75 percent of their work as a result of clients switching to AI.

146. Another content writer told the *Washington Post* that half of his annual income (generated by ten client contracts) was erased when the clients elected to use ChatGPT instead.[34]

147. Recently, the owner of popular online publications such as *Gizmodo*, *Deadspin*, *The Root*, *Jezebel* and *The Onion* came under fire for publishing an error-riddled, AI-generated piece, leading the Writers Guild of America to demand "an immediate end of AI-generated articles" on the company's properties.[35]

---

[33] Authors Guild, "Top Takeaways from the 2023 Author Income Survey (2023), *available at* https://authorsguild.org/news/key-takeaways-from-2023-author-income-survey/#:~:text=Though%20overall%20author%20incomes%20are,coming%20in%20a%20close%20second (last accessed Jan. 22, 2024).

[34] Pranshu Verma & Gerrit De Vynck, *ChatGPT Took Their Jobs. Now They Walk Dogs and Fix Air Conditioners*, The Washington Post (June 2, 2023), *available at* https://www.washingtonpost.com/technology/2023/06/02/ai-taking-jobs (last accessed Jan. 22, 2024).

[35] Todd Spangler, *WGA Slams G/O Media's AI-Generated Articles as 'Existential Threat to Journalism,' Demands Company End Practice*, Variety (July 12, 2023), *available at* https://variety.com/2023/digital/news/wga-slams-go-media-ai-generated-articles-existential-threat-1235668496 (last accessed Jan. 22, 2024).

148. In a survey of authors conducted by The Authors Guild in March 2023 (early in ChatGPT's lifecycle), 69 percent of respondents said they consider generative AI a threat to their profession, and 90 percent said they believe that writers should be compensated for the use of their work in "training" AI.

149. As explained above, until recently, ChatGPT provided verbatim quotes of copyrighted text. Currently, it instead readily offers to produce summaries of such text. These summaries are themselves derivative works, ineluctably based on original unlawfully copied work that could be—but for ChatGPT—licensed by the authors of the underlying works to willing, *paying* licensees.

150. ChatGPT creates other outputs that are derivative of authors' copyrighted works. Businesses are sprouting up to sell prompts that allow users to enter the world of an author's books and create derivative stories within that world. For example, a business called Socialdraft offers long prompts that lead ChatGPT to engage in "conversations" with popular fiction authors like Plaintiff Grisham, Plaintiff Martin, Margaret Atwood, Dan Brown, and others about their works, as well as prompts that promise to help customers "Craft Bestselling Books with AI."

151. OpenAI also allows third parties to build their own applications on top of ChatGPT by making it available through an "application programming interface" or "API." Applications integrated with the API allow users to generate works of fiction, including books and stories similar to those of Plaintiffs and other authors.[36]

152. Defendants' unauthorized commercial copying of Plaintiffs' works and works owned by the proposed Classes was manifestly unfair use. Even OpenAI's attempt to

anthropomorphize its commercial product underscores the unlawful nature of its conduct. By OpenAI's own telling, ChatGPT uses copyrighted texts for the same purpose OpenAI claims an ordinary reading consumer may use a book—to review, understand, and learn from the expression in it, including the order of words, presentation of facts, and syntax, among other expressive elements. Yet humans who learn from books—even those who might describe their reading experience as merely a series of parts not summed into any whole—do not routinely copy them, and inherently must buy them, or borrow them from libraries that buy them, providing some measure of compensation to authors and creators. OpenAI copies books and provides no compensation. It has usurped authors' content for the purpose of creating a machine built to generate the very type of content for which authors usually would be paid.

153. Furthermore, ChatGPT is being used to generate low-quality ebooks, impersonating authors, and displacing human-authored books.[37] For example, author Jane Friedman discovered "a cache of garbage books" written under her name for sale on Amazon.[38]

154. Even OpenAI CEO Sam Altman admitted in testimony to the Senate 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."[39]

---

[36] Adi Robertson, *I Tried the AI Novel-Writing Tool Everyone Hates, and It's Better than I Expected*, The Verge (May 24, 2023), *available at* https://www.theverge.com/2023/5/24/23732252/sudowrite-story-engine-ai-generated-cyberpunk-novella (last accessed Jan. 22, 2024).

[37] Jules Roscoe, *AI-Generated Books of Nonsense Are All Over Amazon's Bestseller Lists*, Vice (June 28, 2023), *available at* https://www.vice.com/en/article/v7b774/ai-generated-books-of-nonsense-are-all-over-amazons-bestseller-lists (last accessed Jan. 22, 2024).

[38] Pilar Melendez, *Famous Author Jane Friedman Finds AI Fakes Being Sold Under Her Name on Amazon*, The Daily Beast (Aug. 8, 2023), *available at* https://www.thedailybeast.com/author-jane-friedman-finds-ai-fakes-being-sold-under-her-name-on-amazon (last accessed Jan. 22, 2024).

[39] Ted Johnson, *OpenAI CEO Sam Altman Says Content Owners Need To Get "Significant Upside Benefit" From New Technology*, Deadline (May 16, 2023), *available at* https://deadline.com/2023/05/ai-chat-gpt-senate-sam-altman-1235368420/ (last accessed Jan. 22. 2024).

155. Plaintiffs and other professional writers are thus reasonably concerned about the risks OpenAI's conduct poses to their livelihoods specifically and the literary arts generally.

156. To this end, Plaintiff The Authors Guild, among others, have given voice to these concerns on behalf of working American authors.

157. The Authors Guild is the nation's oldest and largest professional writers' organization. It "exists to support working writers and their ability to earn a living from authorship."[40]

158. Among other principles, The Authors Guild holds that "authors should not be required to write or speak without compensation. Writers, like all professionals, should receive fair payment for their work."[41]

159. In June 2023, The Authors Guild wrote an open letter (the "Open Letter") calling on OpenAI and other major technology companies to fairly license authors' works for use in LLM "training."

160. The Open Letter emphasizes that "[g]enerative AI technologies built on large language models owe their existence to our writings," and protests "the inherent injustice in exploiting our works as part of your AI systems without our consent, credit, or compensation."[42]

161. The Open Letter also points to the risks to authors' livelihoods posed by generative AI like GPT-N and ChatGPT: "As a result of embedding our writings in your systems, generative AI threatens to damage our profession by flooding the market with mediocre, machine-written books, stories, and journalism based on our work. ... The introduction of generative AI threatens ... to make it even more difficult, if not impossible, for

[40] Authors Guild, https://authorsguild.org (last accessed Jan. 22, 2024).

[41] Authors Guild, *Principles*, https://authorsguild.org/about/principles (last accessed Jan. 22, 2024).

writers—especially young writers and voices from under-represented communities—to earn a living from their profession."[43]

162. To date, the Open Letter has been signed by more than 15,000 authors,[44] including many Plaintiffs here.[45]

163. In short, the success and profitability of OpenAI are predicated on mass copyright infringement without a word of permission from or a nickel of compensation to copyright owners, including Plaintiffs here. OpenAI knows it; its investors know it; and Plaintiffs know it.

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

164. 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.

165. As of November 2023, ChatGPT has reported over 100 million weekly active users. Included among those users are 92% of all Fortune 500 companies.[46] 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.

---

[42] Open Letter from The Authors Guild to Sam Altman *et al.*, at 1, *available at* https://authorsguild.org/app/uploads/2023/07/Authors-Guild-Open-Letter-to-Generative-AI-Leaders.pdf (last accessed Jan. 22, 2024).

[43] *Id.*

[44] Authors Guild, *Open Letter to Generative AI Leaders*, *available at* https://actionnetwork.org/petitions/authors-guild-open-letter-to-generative-ai-leaders (last accessed Jan. 22, 2024).

[45] *See* Open Letter, *supra*, at 2–124.

[46] 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/.

166. 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.

167. 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 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, [47] with just one version of this integration—"GitHub Copilot"—already generating more than $100 million in annual recurring revenue.[48]

F. <u>Defendants Exploited Each of Plaintiffs' Copyrighted Works</u>

1. Fiction Authors

168. Plaintiffs' works collectively span a wide range of commercial fiction whose continuing commercial viability is endangered by Defendants. Each author represented here has a distinct voice, a distinct style, and distinct creative expression. But all Plaintiffs have suffered identical harms from Defendants' infringing reproductions of their works.

---

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

169. The contents of the datasets OpenAI has used to "train" its LLMs are peculiarly within its knowledge and not publicly disclosed, such that Plaintiffs are unable discern those contents with perfect accuracy. Plaintiffs make the specific allegations of infringement below based on what is known about OpenAI's training practices; what is known about the contents, uses, and availability of the pirate book repositories such as LibGen, Bibliotik, and Z-Library; and the results of Plaintiffs' testing of ChatGPT.

170. Most Plaintiffs have written more books than are included in this Complaint.

171. ***Plaintiff The Authors Guild.*** The Authors Guild is the owner of the registered copyrights in Mignon Eberhart's works, including *While the Patient Slept* and *The Patient in Room 18*.

172. Mignon G. Eberhart (1899–1996), dubbed "America's Agatha Christie," was the author of dozens of mystery novels over nearly sixty years. Several of Eberhart's novels have been adapted for film, including *Hasty Wedding*, *Mystery House*, *While the Patient Slept*, *The Patient in Room 18*, and *The White Cockatoo*.

173. The Authors Guild is the owner or beneficial owner of the registered copyrights in eleven (11) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Authors Guild Infringed Works").

174. The registration information for the Authors Guild Infringed Works is contained in Exhibit A to this Complaint, at 1.

175. OpenAI unlawfully and willfully copied the Authors Guild Infringed Works and used them to "train" OpenAI's LLMs without The Authors Guild's permission.

176. For example, when prompted, ChatGPT accurately generated summaries of several of the Authors Guild Infringed Works, including summaries for *While the Patient Slept* and *The Patient in Room 18.*

177. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *While the Patient Slept*, one of the Authors Guild Infringed Works, and titled the infringing and unauthorized derivative “Shadows Over Federie House,” using the same characters from Eberhart’s existing book.

178. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Patient in Room 18*, one of the Authors Guild Infringed Works, and titled the infringing and unauthorized derivative “Echoes from Room 18,” using the same characters from Eberhart’s existing book.

179. When prompted, ChatGPT generated an accurate summary of the final chapter of *While the Patient Slept*, one of the Authors Guild Infringed Works.

180. ChatGPT could not have generated the material described above if OpenAI’s LLMs had not ingested and been “trained” on the Authors Guild Infringed Works.

181. ***Plaintiff Baldacci.*** Baldacci is a best-selling author, philanthropist, and lawyer whose novels have been adapted for film and television, published in over 45 languages and in more than 80 countries, with 150 million copies sold worldwide. Some of Baldacci’s most popular works include books in the Camel Club series, Vega Jane series, and Archer series.

182. Baldacci is a member of The Authors Guild.

183. Baldacci is the sole author of and owner or beneficial owner of the registered copyrights in forty-one (41) written works of fiction, all or many of which OpenAI ingested and copied without permission (the “Baldacci Infringed Works”).

184. The registration information for the Baldacci Infringed Works is contained in Exhibit A to this Complaint, at 1–2.

185. OpenAI unlawfully and willfully copied the Baldacci Infringed Works and used them to “train” OpenAI’s LLMs without Baldacci’s permission.

186. For example, when prompted, ChatGPT accurately generated summaries of several of the Baldacci Infringed Works, including summaries of The Collectors, The Finisher, and One Good Deed.

187. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of The Simple Truth, one of the Baldacci Infringed Works, and titled the infringing and unauthorized derivative "The Complex Justice," using the same characters from Baldacci's existing book.

188. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of Total Control, one of the Baldacci Infringed Works, and titled the infringing and unauthorized derivative "Total Control: Unfinished Business," using the same characters from Baldacci's existing book.

189. When prompted, ChatGPT generated an accurate summary of the final chapter of Long Road to Mercy, one of the Baldacci Infringed Works.

190. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Baldacci Infringed Works.

191. <u>***Plaintiff Bly***</u>. Bly is a tenured professor and chair of the English department at Fordham University who also writes best-selling Regency and Georgian romance novels under the pen name Eloisa James. Some of Bly's most popular works include books in the *Desperate Duchesses* series, the *Fairy Tales* series, the *Wildes of Lindow Castle* series, and the *Essex* series.

192. Bly is a Vice President of The Authors Guild Council and a member of The Authors Guild.

193. Bly is the sole author of and owner or beneficial owner of the registered copyrights in thirty-three (33) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Bly Infringed Works").

194. The registration information for the Bly Infringed Works is contained in Exhibit A to this Complaint, at 2–3.

195. OpenAI unlawfully and willfully copied the Bly Infringed Works used them to "train" OpenAI's LLMs without Bly's permission.

196. For example, when prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *This Duchess of Mine*, one of the Bly Infringed Works, and titled the infringing and unauthorized derivative "The Duchess' New Dawn," using the same characters from Bly's existing book.

197. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *An Affair Before Christmas*, one of the Bly Infringed Works, and titled the infringing and unauthorized derivative "Whispers of Winter," using the same characters from Bly's existing book.

198. When prompted, ChatGPT generated an accurate summary of the final chapter of *A Duke of Her Own*, one of the Bly Infringed Works.

199. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Bly Infringed Works.

200. ***Plaintiff Connelly***. Connelly is a best-selling author with over 85 million copies of his books sold worldwide and translated into 45 foreign languages. Some of Connelly's most popular novels include *The Lincoln Lawyer*, *City of Bones*, and *The Law of Innocence*.

201. Connelly is a member of The Authors Guild.

202. Connelly is the sole author of and owner or beneficial owner of the registered copyrights in forty-six (46) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Connelly Infringed Works").

203. The registration information for the Connelly Infringed Works is contained in Exhibit A to this Complaint, at 3–4.

204. OpenAI unlawfully and willfully copied the Connelly Infringed Works and used them to "train" OpenAI's LLMs without Connelly's permission.

205. For example, when prompted, ChatGPT accurately generated summaries of several of the Connelly Infringed Works, including summaries for *The Black Echo*, *The Poet*, and *The Crossing*.

206. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Lincoln Lawyer*, one of the Connelly Infringed Works, and titled the infringing and unauthorized derivative "The City's Shadows," using the same characters from Connelly's existing book.

207. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Brass Verdict*, one of the Connelly Infringed Works, and titled the infringing and unauthorized derivative "Double-Edged Justice," using the same characters from Connelly's existing book.

208. When prompted, ChatGPT generated an accurate summary of the final chapter of *The Late Show*, one of the Connelly Infringed Works.

209. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Connelly Infringed Works.

210. ***Plaintiff Day***. Day is a best-selling author of over twenty award-winning novels, including ten *New York Times* best sellers and thirteen *USA Today* best sellers. Her work has

been translated into forty-one languages. Some of Day's most popular novels include books in *The Crossfire® Saga* series, the *Georgian* series, and the *Marked* series.

211. Day is a member of The Authors Guild Council and a member of The Authors Guild.

212. Day is the sole author of and owner or beneficial owner of the registered copyrights in thirty-one (31) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Day Infringed Works").

213. The registration information for the Day Infringed Works is contained in Exhibit A to this Complaint, at 4.

214. OpenAI unlawfully and willfully copied the Day Infringed Works and used them to "train" OpenAI's LLMs without Day's permission.

215. For example, when prompted, ChatGPT accurately generated summaries of several of the Day Infringed Works, including summaries for *Bared to You*, *One With You*, and *Ask For It*.

216. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *A Touch of Crimson*, one of the Day Infringed Works, and titled the infringing and unauthorized derivative "Crimson Temptations: A Love Rekindled," using the same characters from Day's existing book.

217. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Butterfly in Frost*, one of the Day Infringed Works, and titled the infringing and unauthorized derivative "Butterfly in Frost: Embers of Desire," using the same characters from Day's existing book.

218. When prompted, ChatGPT generated an accurate summary of the final chapter of *The Stranger I Married*, one of the Day Infringed Works.

219. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Day Infringed Works.

220. ***Plaintiff Franzen***. Franzen is a novelist whose honors include the National Book Award, the James Tait Black Memorial Award, the Heartland Prize, Die Welt Literature Prize, the Budapest Grand Prize, and the first Carlos Fuentes Medal awarded at the Guadalajara International Book Fair. Franzen is a member of the American Academy of Arts and Letters, the American Academy of Arts and Sciences, the German Akademie der Künste, and the French Ordre des Arts et des Lettres. Some of Franzen's most popular novels include *The Corrections, Purity,* and *Freedom*.

221. Franzen is a member of The Authors Guild.

222. Franzen is the sole author of and owner or beneficial owner of the registered copyrights in five (5) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Franzen Infringed Works").

223. The registration information for the Franzen Infringed Works is contained in Exhibit A to this Complaint, at 4–5.

224. OpenAI unlawfully and willfully copied the Franzen Infringed Works and used them to "train" OpenAI's LLMs without Franzen's permission.

225. For example, when prompted, ChatGPT accurately generated summaries of several of the Franzen Infringed Works, including summaries for *The Corrections*, *Purity*, and *Freedom*.

226. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Corrections*, one of the Franzen Infringed Works, and titled the infringing and unauthorized derivative "Revisions," using the same characters from Franzen's existing book.

227. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Twenty-Seventh City*, one of the Franzen Infringed Works, and titled the infringing and unauthorized derivative "The Rising Metropolis," using the same characters from Franzen's existing book.

228. When prompted, ChatGPT generated an accurate summary of the final chapter of *Freedom*, one of the Franzen Infringed Works.

229. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Franzen Infringed Works.

230. ***Plaintiff Grisham***. Grisham is a civically engaged and best-selling author. His award-winning work has been translated into approximately 50 languages and adapted for both television and film. Some of Grisham's most popular novels include *The Pelican Brief*, *The Runaway Jury*, and *The Rainmaker*.

231. Grisham is a member of The Authors Guild.

232. Grisham is the sole author of and owner or beneficial owner of the registered copyrights in twenty-six (26) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Grisham Infringed Works").

233. The registration information for the Grisham Infringed Works is contained in Exhibit A to this Complaint, at 5.

234. OpenAI unlawfully and willfully copied the Grisham Infringed Works and used them to "train" OpenAI's LLMs without Grisham's permission.

235. For example, when prompted, ChatGPT accurately generated summaries of several of the Grisham Infringed Works, including summaries for *The Chamber*, *The Client*, and *The Firm*.

236. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The King of Torts*, one of the Grisham Infringed Works, and titled the infringing and unauthorized derivative "The Kingdom of Consequences," using the same characters from Grisham's existing book.

237. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Last Juror*, one of the Grisham Infringed Works, and titled the infringing and unauthorized derivative "The Juror's Dilemma," using the same characters from Grisham's existing book.

238. When prompted, ChatGPT generated an accurate summary of the final chapter of *The Litigators*, one of the Grisham Infringed Works.

239. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Grisham Infringed Works.

240. ***Plaintiff Hilderbrand***. Hilderbrand is a best-selling author, whose works include novels in the romance genre adapted for television. Hilderbrand has previously taught writing at the University of Iowa. Some of Hilderbrand's most popular novels include *The Summer of '69*, *The Identicals*, and *The Perfect Couple*.

241. Hilderbrand is the sole author of and owner or beneficial owner of the registered copyrights in twenty-nine (29) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Hilderbrand Infringed Works").

242. The registration information for the Hilderbrand Infringed Works is contained in Exhibit A to this Complaint, at 5–6.

243. OpenAI unlawfully and willfully copied the Hilderbrand Infringed Works and used them "train" OpenAI's LLMs without Hilderbrand's permission.

244. For example, when prompted, ChatGPT accurately generated summaries of several of the Hilderbrand Infringed Works, including summaries for *The Summer of '69*, *The Identicals*, and *The Perfect Couple*.

245. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Identicals*, one of the Hilderbrand Infringed Works, and titled the infringing and unauthorized derivative "The Reckoning of Twins," using the same characters from Hilderbrand's existing book.

246. When prompted, ChatGPT generated an accurate summary of the final chapter of *The Perfect Couple*, one of the Hilderbrand Infringed Works.

247. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Hilderbrand Infringed Works.

248. ***<u>Plaintiff Kline</u>***. Kline is a globally published author who writes best-selling novels and has taught different disciplines of writing at Yale University, New York University, and the University of Virginia. Some of Kline's most popular novels include *Orphan Train*, *A Piece of the World*, and *Bird in Hand*.

249. Kline is a member of The Authors Guild Council and a member of The Authors Guild.

250. Kline is the sole author of and owner or beneficial owner of the registered copyrights in five (5) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Kline Infringed Works").

251. The registration information for the Kline Infringed Works is contained in Exhibit A to this Complaint, at 6.

252. OpenAI unlawfully and willfully copied the Kline Infringed Works and used them to "train" OpenAI's LLMs without Kline's permission.

253. For example, when prompted, ChatGPT accurately generated summaries of several of the Kline Infringed Works, including summaries for *Orphan Train*, *A Piece of the World*, and *Bird in Hand.*

254. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Orphan Train*, one of the Kline Infringed Works, and titled the infringing and unauthorized derivative "Legacy Rails," using the same characters from Kline's existing book.

255. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Bird in Hand*, one of the Kline Infringed Works, and titled the infringing and unauthorized derivative "Ties That Bind," using the same characters from Kline's existing book.

256. When prompted, ChatGPT generated an accurate summary of the final chapter of *A Piece of the World*, one of the Kline Infringed Works.

257. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Kline Infringed Works.

258. ***Plaintiff Lang***. Lang is an author and teacher who holds a doctorate in Comparative Literature. Lang is the author of the novel *The Sixteenth of June*.

259. Lang is the President and a member of The Authors Guild.

260. Lang is the sole author of and owner or beneficial owner of the registered copyrights in one (1) written work of fiction that OpenAI ingested and copied without permission (the "Lang Infringed Work").

261. The registration information for the Lang Infringed Work is contained in Exhibit A to this Complaint, at 6.

262. OpenAI unlawfully and willfully copied the Lang Infringed Work and used it to "train" OpenAI's LLMs without Lang's permission.

263. When prompted, ChatGPT accurately generated a summary of the Lang Infringed Work, *The Sixteenth of June*.

264. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Sixteenth of June*, the Lang Infringed Work, and titled the infringing and unauthorized derivative "The Seventeenth of June," using the same characters from Lang's existing book.

265. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Lang Infringed Work.

266. ***Plaintiff LaValle***. LaValle is an associate professor of Creative Writing at Columbia University and the author of five novels, a short story collection, two novellas, and two comic books. Some of LaValle's most popular novels include *Big Machine, The Devil in Silver*, and *The Changeling*.

267. LaValle is the sole author of and owner or beneficial owner of the registered copyrights in six (6) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "LaValle Infringed Works").

268. The registration information for the LaValle Infringed Works is contained in Exhibit A to this Complaint, at 6.

269. OpenAI unlawfully and willfully copied the LaValle Infringed Works and used them to "train" OpenAI's LLMs without LaValle's permission.

270. For example, when prompted, ChatGPT accurately generated summaries of several of the LaValle Infringed Works, including summaries for *Big Machine, The Devil in Silver*, and *The Changeling*.

271. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Changeling*, one of the LaValle Infringed Works, and titled the infringing and unauthorized derivative "The Fae's Return," using the same characters from LaValle's existing book.

272. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Devil in Silver*, one of the LaValle Infringed Works, and titled the infringing and unauthorized derivative "The New Hyde Legacy," using the same characters from LaValle's existing book.

273. When prompted, ChatGPT generated an accurate summary of the final chapter of *Big Machine*, one of the LaValle Infringed Works.

274. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the LaValle Infringed Works.

275. ***Plaintiff Martin***. Martin is an award-winning author, television producer, and writer who is widely known for his fantasy, science fiction, and horror writing. Some of Martin's most popular novels include *A Game of Thrones*, *A Clash of Kings*, and *A Storm of Swords*.

276. Martin is the sole author of and owner or beneficial owner of the registered copyrights in fifteen (15) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Martin Infringed Works").

277. The registration information for the Martin Infringed Works is contained in Exhibit A to this Complaint, at 6–7.

278. OpenAI unlawfully and willfully copied the Martin Infringed Works and used them to "train" OpenAI's LLMs without Martin's permission.

279. In July 2023, Liam Swayne used ChatGPT to generate versions of *The Winds of Winter* and *A Dream of Spring,* intended to be the final two books in the series *A Song of Ice and Fire*, which Martin is currently writing.

280. An experiment conducted by researchers at the University of California, Berkeley, into the "memorization" of works by ChatGPT found that Martin's novel *A Game of Thrones* ranked 12th with respect to the degree of "memorization."[49]

281. When prompted, ChatGPT accurately generated summaries of several of the Martin Infringed Works, including summaries for Martin's novels *A Game of Thrones*, *A Clash of Kings*, and *A Storm of Swords,* the first three books in the series *A Song of Ice and Fire*.

282. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for an alternate sequel to *A Clash of Kings*, one of the Martin Infringed Works, and titled the infringing and unauthorized derivative "A Dance With Shadows," using the same characters from Martin's existing books in the series *A Song of Ice and Fire*.

283. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for a prequel book to *A Game of Thrones*, one of the Martin Infringed Works, and titled the infringing and unauthorized derivative "A Dawn of Direwolves," using the same characters from Martin's existing books in the series *A Song of Ice and Fire*.

284. When prompted, ChatGPT generated an accurate summary of the final chapter of *The Armageddon Rag*, one of the Martin Infringed Works.

285. ChatGPT could not have generated the results described above if OpenAI's LLMs had not ingested and been "trained" on the Martin Infringed Works.

---

[49] *See* Kent K. Chang *et al.*, *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4* (2023), *available at* https://arxiv.org/pdf/2305.00118v1.pdf (last accessed Dec. 4, 2023).

286. ***Plaintiff Picoult***. A *New York Times* best-selling author, Picoult writes popular fiction. Picoult is also the recipient of many awards, including the New England Bookseller Award for Fiction, the Alex Awards from the YALSA, a lifetime achievement award for mainstream fiction from the Romance Writers of America, the NH Literary Award for Outstanding Literary Merit and the Sarah Josepha Hale Award. Some of Picoult's most popular novels include *My Sister's Keeper*, *Nineteen Minutes*, and *House Rules*.

287. Picoult is a member of The Authors Guild.

288. Picoult is the sole author of and owner or beneficial owner of the registered copyrights in twenty-seven (27) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Picoult Infringed Works").

289. The registration information for the Picoult Infringed Works is contained in Exhibit A to this Complaint, at 7.

290. OpenAI unlawfully and willfully copied the Picoult Infringed Works and used them to "train" OpenAI's LLMs without Picoult's permission.

291. For example, when prompted, ChatGPT accurately generated summaries of several of the Picoult Infringed Works, including summaries for *Keeping Faith*, *Handle With Care*, and *Sing You Home*.

292. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Small Great Things*, one of the Picoult Infringed Works, and titled the infringing and unauthorized derivative "Small Great Things: Unfinished Business," using the same characters from Picoult's existing book.

293. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *My Sister's Keeper*, one of the Picoult Infringed

Works, and titled the infringing and unauthorized derivative as "My Sister's Legacy," using the same characters from Picoult's existing book.

294. When prompted, ChatGPT generated an accurate summary of the final chapter of *Change of Heart*, one of the Picoult Infringed Works.

295. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Picoult Infringed Works.

296. ***Plaintiff Preston***. Preston is an author and journalist who has received awards for his writing, both in America and abroad, and previously taught writing at Princeton University. Some of Preston's most popular novels include *Blasphemy*, *Impact*, and *The Codex*.

297. Preston is a member of The Authors Guild and past President of The Authors Guild Council.

298. Preston is the sole author of and owner or beneficial owner of the registered copyrights in six (6) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Preston Infringed Works").

299. The registration information for the Preston Infringed Works is contained in Exhibit A to this Complaint, at 7.

300. OpenAI unlawfully and willfully copied the Preston Infringed Works and used them to "train" OpenAI's LLMs without Preston's permission.

301. For example, when prompted, ChatGPT accurately generated summaries of several of the Preston Infringed Works, including summaries for *Impact*, *Blasphemy*, and *The Codex*.

302. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Impact*, one of the Preston Infringed Works, and

titled the infringing and unauthorized derivative "Unearthed Secrets," using the same characters from Preston's existing book.

303. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Codex*, one of the Preston Infringed Works, and titled the infringing and unauthorized derivative "The Codex: The Lost Dynasty," using the same characters from Preston's existing book.

304. When prompted, ChatGPT generated an accurate summary of the final chapter of *The Kraken Project*, one of the Preston Infringed Works.

305. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Preston Infringed Works.

306. ***Plaintiff Robinson***. Robinson is an award-winning author with a wide reach, having written six novels and three collections of short stories, whose fiction has appeared in internationally respected publications and whose books have been published internationally. Some of Robinson's most popular novels include *Dawson's Fall*, *Sparta*, and *Cost*.

307. Robinson is a member of The Authors Guild and a past President of The Authors Guild Council.

308. Robinson is the sole author of and owner or beneficial owner of the registered copyrights in eight (8) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Robinson Infringed Works").

309. The registration information for the Robinson Infringed Works is contained in Exhibit A to this Complaint, at 7–8.

310. OpenAI unlawfully and willfully copied the Robinson Infringed Works and used them to "train" OpenAI's LLMs without Robinson's permission.

311. For example, when prompted, ChatGPT accurately generated summaries of several of the Robinson Infringed Works, including summaries of *Cost*, *Sparta* and *Dawson's Fall*.

312. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Dawson's Fall*, one of the Robinson Infringed Works, and titled the infringing and unauthorized derivative "Dawson's Legacy," using the same characters from Robinson's existing book.

313. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Sparta*, one of the Robinson Infringed Works, and titled the infringing and unauthorized derivative "Homefront," using the same characters from Robinson's existing book.

314. When prompted, ChatGPT generated an accurate summary of the final chapter of *Sparta*, one of the Robinson Infringed Works.

315. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Robinson Infringed Works.

316. ***Plaintiff Saunders***. Saunders is a professor in the English department at Syracuse University, who also writes best-selling books of fiction. Some of Saunders' most popular works include the short story titled *Escape From Spiderhead*, a novel titled *Lincoln in the Bardo*, and a novella titled *The Brief and Frightening Reign of Phil*.

317. Saunders is a member of The Authors Guild.

318. Saunders is the sole author of and owner or beneficial owner of the registered copyrights in seven (7) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Saunders Infringed Works").

319. The registration information for the Saunders Infringed Works is contained in Exhibit A to this Complaint, at 8.

320. OpenAI unlawfully and willfully copied the Saunders Infringed Works and used them to "train" OpenAI's LLMs without Saunders's permission.

321. For example, when prompted, ChatGPT accurately generated summaries of several of the Saunders Infringed Works, including summaries for *CivilWarLand in Bad Decline*, *Lincoln in the Bardo*, and *Tenth of December*.

322. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Fox 8*, one of the Saunders Infringed Works, and titled the infringing and unauthorized derivative "Fox 8 and the Hidden World," using the same characters from Saunders's existing book.

323. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Tenth of December*, one of the Saunders Infringed Works, and titled the infringing and unauthorized derivative "The Eleventh of December: A Continuation," using the same characters from Saunders's existing book.

324. When prompted, ChatGPT generated an accurate summary of the conclusion of *Escape From Spiderhead*, one of the Saunders Infringed Works.

325. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Saunders Infringed Works.

326. ***Plaintiff Turow***. A best-selling author, Turow is a novelist and lawyer who is best known for setting his novels in fictional Kindle County's legal community. Some of Turow's most popular novels include *The Last Trial*, *Testimony*, and *Identical*.

327. Turow is a member of The Authors Guild and past President of The Authors Guild Council.

328. Turow is the sole author of and owner or beneficial owner of the registered copyrights in sixteen (16) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Turow Infringed Works").

329. The registration information for the Turow Infringed Works is contained in Exhibit A to this Complaint, at 8.

330. OpenAI unlawfully and willfully copied the Turow Infringed Works and used them to "train" OpenAI's LLMs without Turow's permission.

331. For example, when prompted, ChatGPT accurately generated summaries of several of the Turow Infringed Works, including summaries for *The Burden of Proof*, *Innocent*, and *Testimony*.

332. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *The Last Trial*, one of the Turow Infringed Works, and titled the infringing and unauthorized derivative "Echoes of Judgment," using the same characters from Turow's existing book.

333. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Pleading Guilty*, one of the Turow Infringed Works, and titled the infringing and unauthorized sequel "Redemption's Price," using the same characters from Turow's existing book.

334. When prompted, ChatGPT generated an accurate summary of the final chapter of *Ordinary Heroes*, one of the Turow Infringed Works.

335. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Turow Infringed Works.

336. ***Plaintiff Vail***. Rachel Vail is an award-winning American author who primarily authors children's and young adult books. Some of Vail's most popular novels include *Ever After*, *Unfriended*, and *Justin Case: School, Drool, and Other Daily Disasters*.

337. Vail is a member of The Authors Guild and a member of The Authors Guild Council.

338. Vail is the sole author of and owner or beneficial owner of the registered copyrights in twenty-four (24) written works of fiction, all or many of which OpenAI ingested and copied without permission (the "Vail Infringed Works").

339. The registration information for the Vail Infringed Works is contained in Exhibit A to this Complaint, at 8–9.

340. OpenAI unlawfully and willfully copied the Vail Infringed Works and used them to "train" its LLMs without Vail's permission.

341. For example, when prompted, ChatGPT accurately generated summaries of several of the Vail Infringed Works, including summaries for *If We Kiss*, *A Is For Elizabeth*, and *Not That I Care*.

342. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Bad Best Friend*, one of the Vail Infringed Works, and titled the infringing and unauthorized derivative "Redeeming Friendship," using the same characters from Vail's existing book.

343. When prompted, ChatGPT generated an infringing, unauthorized, and detailed outline for the next purported installment of *Do-Over*, one of the Vail Infringed Works, and titled the infringing and unauthorized derivative "Do-Over: Second Chances," using the same characters from Vail's existing book.

344. When prompted, ChatGPT generated an accurate summary of the final chapter of *Daring to be Abigail*, one of the Vail Infringed Works.

345. ChatGPT could not have generated the material described above if OpenAI's LLMs had not ingested and been "trained" on the Vail Infringed Works.

2. Nonfiction Authors

346. While OpenAI and Microsoft have kept the contents of their training data secret, it is likely that, in training their GPT models, they reproduced all or nearly all commercially successful nonfiction books. As OpenAI investor Andreesen Horowitz has admitted, "large language models," like Defendants' GPT models, "are trained on something approaching the entire corpus of the written word," a corpus that would of course include Plaintiffs' works.

347. The size of the Books2 database—the "internet based books corpora" that Defendants used to train GPT-3, GPT-3.5, and possibly GPT-4 as well—has led commentators to believe that Books2 is comprised of books scraped from entire pirated online libraries such as LibGen, ZLibrary, or Bibliotik. Shawn Presser, an independent software developer, created an open-source set of training data called Books3, which was intended to give developers, in his words, "OpenAI-grade training data." The Books3 dataset, similar in size to Books2, was built from a corpus of pirated copies of books available on the site Bibliotik. Works authored and owned by Plaintiffs Alter, Bird, Branch, Cohen, Linden, Okrent, Sancton, Sides, Schiff, Shapiro, Tolentino, and Winchester are available on Books3, an indication that these works were also likely included in the similarly sized Books2.

348. ***Plaintiff Alter***. Alter is the author of a number of *New York Times* bestsellers, including *The Center Holds: Obama and His Enemies*; *The Promise: President Obama, Year One*; and *The Defining Moment: FDR's Hundred Days and the Triumph of Hope*. Each of those three books are a part of the Books3 dataset. Pirated copies of each of those three books—as well

as most recent book *His Very Best: Jimmy Carter, A Life*—are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

349. When prompted with questions about Alter's books, ChatGPT returned detailed, accurate summaries of them, including of *The Center Holds* and *The Promise*. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Alter's books.

350. Alter is the author and owner of the registered copyrights listed under his name in Exhibit B.

351. ***Plaintiff Bird***. Bird is the recipient of the 2006 Pulitzer Prize for Biography for *American Prometheus: The Triumph and Tragedy of J. Robert Oppenheimer*. A number of this books, including *American Prometheus*, *The Good Spy*, and *The Color of Truth* are a part of the Books3 dataset. Pirated copies of each of all of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

352. When prompted with questions about Bird's books, ChatGPT returned detailed, accurate summaries of them, including of *American Prometheus* and *The Good Spy*. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Bird's books.

353. Bird is the author and owner of the registered copyrights listed under his name in Exhibit B.

354. ***Plaintiff Branch***. Branch is the author of, among other works, *America in the King Years*, a three-volume history of Martin Luther King Jr. and the Civil Rights Movement. He received the 1989 Pulitzer Prize in History for the first volume in the series, *Parting the Waters:*

America in the King Years, 1954-63. Five of his books, including *Parting the Waters* and *The Clinton Tapes*, are a part of the Books3 dataset. Pirated copies of each of nearly all of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

355. When prompted with questions about Branch's books, ChatGPT returned detailed, accurate summaries of them, including of *Parting the Waters* and *The Clinton Tapes*. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Branch's books.

356. Branch is the author and owner of the registered copyrights listed under his name in Exhibit B.

357. ***Plaintiff Cohen***. Cohen is the author of several *New York Times* bestsellers, including *Tough Jews: Fathers, Sons and Gangster Dreams*; *Sweet and Low: A Family Story*; and *The Sun & the Moon & the Rolling Stones*. He is a contributing editor for *Vanity Fair* and *Rolling Stone* and a columnist for *The Wall Street Journal*.

358. Six of his books, including *Tough Jews* and *The Chicago Cubs*, are a part of the Books3 dataset. Pirated copies of each of nearly all of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

359. When prompted with questions about Plaintiff Cohen's books, ChatGPT returned detailed, accurate summaries of them, including of *Tough Jews*. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Cohen's books.

360. Cohen is the author and owner of the registered copyrights listed under his name in Exhibit B.

361. ***__Plaintiff Linden__***. Linden is the author of nine nonfiction books, including *The Parrot's Lament, and Other True Tales of Animal Intrigue*, *Intelligence, and Ingenuity*; *The Octopus and the Orangutan: More True Tales of Animal Intrigue, Intelligence, and Ingenuity*; *The Alms Race: The Impact of American Voluntary Aid Abroad*; and *Winds of Change*.

362. Two of his books, *The Parrot's Lament* and *The Ragged Edge of the World*, are a part of the Books3 dataset. Pirated copies of a number of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

363. When prompted with questions about Plaintiff Linden's books, ChatGPT returned detailed, accurate summaries of them, including of *The Octopus and the Orangutan*. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Linden's books.

364. Linden is the author and owner of the registered copyrights listed under his name in Exhibit B.

365. ***__Plaintiff Okrent__***. Okrent is the author of a number of nonfiction books, including *Great Fortune: The Epic of Rockefeller Center*, *Last Call: The Rise and Fall of Prohibition*, and *The Guarded Gate: Bigotry, Eugenics and the Law that Kept Two Generations of Jews, Italians and Other European Immigrants Out of America.*

366. At least three of his books, including *Last Call* and *The Guarded Gate*, are a part of the Books3 dataset. Pirated copies of a number of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

367. When prompted with questions about Plaintiff Okrent's books, ChatGPT returned detailed, accurate summaries of them, including of *Last Call* and *The Guarded Gate*. Upon

information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Okrent's books.

368. Okrent is the author and owner of the registered copyrights listed under his name in Exhibit B.

369. ***<u>Plaintiff Sancton</u>***. Sancton 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*.

370. Pirated copies of his book *Madhouse at the End of the Earth* are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset. When prompted with questions about Plaintiff Sancton's book, ChatGPT confirmed that *Madhouse at the End of the Earth* was part of its training dataset.

371. Sancton is the author and owner of the registered copyright listed under his name in Exhibit B.

372. ***<u>Plaintiff Sides</u>***. Sides is the author of a number of *New York Times* bestsellers, including *Ghost Soldiers: The Epic Account of World War II's Greatest Rescue Mission* and *Blood and Thunder: An Epic of the American West*. He is an editor-at-large for *Outside* and is a frequent contributor to *National Geographic*.

373. Six of his books, including *Ghost Soldiers* and *Blood and Thunder*, are a part of the Books3 dataset. Pirated copies of all of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

374. When prompted with questions about Plaintiff Sides's books, ChatGPT returned detailed, accurate summaries of them, including of *Ghost Soldiers* and *Blood and Thunder*. Upon

information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Sides's books.

375. Sides is the author and owner of the registered copyrights listed under his name in Exhibit B.

376. ***Plaintiff Schiff***. Schiff is the recipient of the 2000 Pulitzer Prize in Biography for *Véra (Mrs. Vladimir Nabokov)*. Plaintiff Schiff is also the author of the *New York Times* bestsellers *Cleopatra: A Life*; *The Witches: Salem, 1692*; and *The Revolutionary*: *Samuel Adams.* Her work has appeared in *The New Yorker*, *The New York Review of Books*, and *The New York Times*.

377. At least five of her books, including *The Witches* and *Cleopatra*, are a part of the Books3 dataset. Pirated copies of all of her books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

378. When prompted with questions about Plaintiff Schiff's books, ChatGPT returned detailed, accurate summaries of them, including of *The Witches* and *Cleopatra*. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Schiff's books.

379. Schiff is the author and owner of the registered copyrights listed under her name in Exhibit B.

380. ***Plaintiff Shapiro***. Shapiro is the author of a number of nonfiction books, including *Oberammergau: The Troubling Story of the World's Most Famous Passion Play* and *1599: A Year in the Life of William Shakespeare.* He is the Larry Miller Professor of English and Comparative Literature at Columbia University.

381. At least four of his books, including *The Year of Lear* and *Contested Will*, are a part of the Books3 dataset. Pirated copies of nearly all of his books are available on the internet

through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

382. When prompted with questions about Plaintiff Shapiro's books, ChatGPT returned detailed, accurate summaries of them, including of *Contested Will*. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Shapiro's books.

383. Shapiro is the author and owner of the registered copyrights listed under his name in Exhibit B.

384. ***Plaintiff Tolentino***. She is the author of the *New York Times* bestseller *Trick Mirror: Reflections on Self-Delusion*. She is a staff writer for *The New Yorker* whose work has also appeared in *The New York Times Magazine* and *Pitchfork.*

385. Her book *Trick Mirror* part of the Books3 dataset. Pirated copies of it are widely available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

386. When prompted with questions about *Trick Mirror*, ChatGPT returned detailed, accurate summaries of the book and its chapters. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Tolentino's book.

387. Tolentino is the author and owner of the registered copyright listed under her name in Exhibit B.

388. ***Plaintiff Winchester***. Winchester is the author of several *New York Times* bestsellers, including *The Professor and the Madman*; *The Map That Changed the World: William Smith and the Birth of Modern Geolog*y; and *The Men Who United the States: America's Explorers, Inventors, Eccentrics, and Mavericks, and the Creation of One Nation, Indivisible*.

389. At least eleven of his books, including *The Professor and the Madman* and *Krakatoa*, are a part of the Books3 dataset. Pirated copies of nearly all of his books are available on the internet through websites like LibGen, ZLibrary, and/or Bibliotik, which likely sourced OpenAI's Books2 dataset.

390. When prompted with questions about Plaintiff Winchester's books, ChatGPT returned detailed, accurate summaries of them, including of *The Professor and the Madman* and *Krakatoa*. Upon information and belief, ChatGPT is able to return such detailed information only because it was trained on Plaintiff Winchester's books.

391. Winchester is the author and owner of the registered copyrights listed under his name in Exhibit B.

CLASS ALLEGATIONS

A. Class Definitions

392. This action is brought by Plaintiffs individually and on behalf of the Fiction and Nonfiction Author Classes, as defined below, pursuant to Rule 23(b)(3) and 23(b)(1) of the Federal Rules of Civil Procedure.

393. The Fiction Author Class consists of:

> All natural persons in the United States who are the sole authors of, and legal or beneficial owners of Eligible Copyrights in, one or more Fiction Class Works; and all persons in the United States who are the legal or beneficial owners of Eligible Fiction Copyrights in one or more Fiction Class Works held by literary estates.

394. Fiction Class Works is defined as follows:

> Any work of fiction, the text of which has been, or is being, used by Defendants to "train" one or more of Defendants' large language models.

395. Eligible Fiction Copyrights is defined as follows:

> Any copyright that was registered with the United States Copyright Office before or within five years after first publication of the work, and whose effective date of

registration is either within three months after first publication of the work or before Defendants began using the work to "train" one or more of Defendants' large language models.

396. The Nonfiction Author Class consists of:

All natural persons, literary trusts, and literary estates in the United States who are legal or beneficial owners of Eligible Nonfiction Copyrights in one or more Nonfiction Class Works; and all persons in the United States who are the legal or beneficial owners of Eligible Nonfiction Copyrights in one or more Fiction Class Works held by literary estates.

397. Nonfiction Class Works is defined as follows:

Any work of nonfiction (A) the text of which has been, or is being, used by Defendants to "train" one or more of Defendants' large language models; (B) that have been assigned an International Standard Book Number (ISBN); and that fall within a Book Industry Standards and Communications (BISAC) code other than Reference (REF).[50]

398. Eligible Nonfiction Copyrights is defined as follows:

Any copyright that was registered with the United States Copyright Office before or within five years after first publication of the work.

399. Excluded from the definitions of the Fiction Author Class and Nonfiction Author Class above are Defendants; Defendants' co-conspirators, aiders and abettors, and members of their immediate families; Defendants' corporate parents, subsidiaries, and affiliates; Defendants' directors, officers, employees, and other agents, as well as members of their immediate families; and any judge who may preside over this action, the judge's staff, and members of their immediate families.

B. <u>Rules 23(a) and 23(g)</u>

400. Both Classes consists of at least tens of thousands of authors and copyright holders and thus are so numerous that joinder of all members is impractical.

401. The identities of members of the Classes can be readily ascertained from business records maintained by Defendants.

402. The claims asserted by Plaintiffs are typical of the claims of the Classes because their copyrights were infringed in materially the same way and their interests in preventing future infringement and redressing past infringement are materially the same.

403. The Plaintiffs will fairly and adequately protect the interests of the Classes and do not have any interests antagonistic to those of other members of the Classes.

404. Plaintiffs have retained attorneys who are knowledgeable and experienced in copyright and class action matters, as well as complex litigation.

405. There are questions of fact or law common to the Classes, including:

a. Whether Defendants' copied works owned by Plaintiffs and the members of the Classes;

b. Whether Defendants' copying of Plaintiffs' and the Classes' copyrighted works consisted direct, vicarious, or contributory infringement; and

c. Whether Defendants' copying of works owned by Plaintiffs and the Classes was willful.

C. <u>Rule 23(b)</u>

406. Defendants have acted on grounds common to Plaintiffs and the Classes by treating all Plaintiffs' and Class Members' works equally, in all material respects, in their LLM "training."

407. Common questions of liability for infringement predominate over any individualized damages determinations as may be necessary. To decide liability, the Court will necessarily apply the same law to the same conduct, which Defendants engaged in indiscriminately with respect to all Plaintiffs and all members of the Classes.

---

[50] *See* http://www.bisg.org/complete-bisac-subject-headings-list (last visited Feb. 2, 2024)

408. Further, to the extent Plaintiffs elect to pursue statutory rather than actual damages before final judgment, the damages inquiry will likewise be common, if not identical, across Plaintiffs and members of the Classes.

409. A class action is superior to any individual litigation of Plaintiffs' and Class Members' claims. Class Members have little interest, distinct from Plaintiffs' and other Class Members', in prosecuting individual actions. It would waste judicial resources to decide the same legal questions repeatedly, thousands of times over, on materially indistinguishable facts. The Classes presents no special manageability problems.

410. This action is also appropriate as a class action pursuant to Rule 23(b)(2) of the Federal Rules of Civil Procedure because Defendants infringing conduct is applicable generally to Plaintiffs and the proposed Classes and the requested injunctive relief is appropriate respecting the proposed Classes as a whole.

D. Rule 23(c)(4)

411. In the alternative to certification under Rule 23(b)(3) and 23(b)(2), common questions predominate within the determination of liability for infringement, therefore the issue of liability may be separately certified for class treatment even if the entire action is not.

CLAIMS FOR RELIEF

COUNT I: COPYRIGHT INFRINGEMENT (17 U.S.C. § 501)

Against OpenAI and Microsoft

412. Plaintiffs incorporate by reference the allegations in Paragraphs 1 to 411 as though fully set forth herein.

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

414. Plaintiffs and members of the Classes therefore hold the exclusive rights, including the rights of reproduction and distribution, to those works under 17 U.S.C. § 106.

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

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

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

418. Plaintiffs 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 Plaintiffs 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: VICARIOUS COPYRIGHT INFRINGEMENT

Against OpenAI Inc. and OpenAI GP LLC

419. Plaintiffs incorporate and reallege paragraphs 1 through 418 above.

420. Defendants OpenAI Inc. and OpenAI GP LLC had the right and ability to control the direct infringement alleged in Count I because Defendant OpenAI Inc. fully controls Defendant OpenAI GP LLC, and Defendant OpenAI GP LLC fully controls Defendant OpenAI OpCo LLC, according to the corporate structure outlined above.

421. Defendants OpenAI Inc. and OpenAI GP LLC have a direct financial interest in the direct infringement alleged in Count I because they benefit from the profits and investments generated by Defendant OpenAI OpCo LLC's infringing activities.

422. Defendants OpenAI Inc. and OpenAI GP LLC are vicariously liable for the direct infringement alleged in Count I.

COUNT III: CONTRIBUTORY INFRINGEMENT

Against Microsoft, OpenAI, Inc., OpenAI GP LLC, OpenAI Global LLC, OpenAI LLC, OAI Corporation LLC, OpenAI Holdings LLC, OpenAI Startup Fund I LP, OpenAI Startup Fund GP I LLC, and OpenAI Startup Fund Management LLC

423. Plaintiffs incorporate by reference and realleges the allegations in Paragraphs 1 to 422 as though fully set forth herein.

424. 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 Plaintiffs and the proposed Classes. This assistance was necessary for OpenAI to perpetrate the largescale copyright infringement alleged herein.

425. 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 Plaintiffs and the proposed Class.

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

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

428. OpenAI, Inc., OpenAI GP LLC, OpenAI Global LLC, OpenAI LLC, OAI Corporation LLC, OpenAI Holdings LLC, OpenAI Startup Fund I LP, OpenAI Startup Fund GP I LLC, and OpenAI Startup Fund Management 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.

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

PRAYER FOR RELIEF

430. Plaintiffs, on behalf of themselves and all others similarly situated, pray for the following relief:

a. Certification of this action as a class action under Federal Rule of Civil Procedure 23;

b. Designation of Plaintiffs as class representatives;

c. Designation of Plaintiffs' counsel as class counsel;

d. An injunction prohibiting Defendants from infringing on Plaintiffs' and class members' copyrights, including without limitation enjoining Defendants from using Plaintiffs' and class members' copyrighted works in "training" Defendants' large language models without express authorization;

e. An award of actual damages to Plaintiffs and class members;

f. An award of Defendants' additional profits attributable to infringement to Plaintiffs and class members;

g. An award of statutory damages up to $150,000 per infringed work to Plaintiffs and members of the Classes, in the alternative to actual damages and profits, at Plaintiffs' election before final judgment;

431. Reasonable attorneys' fees and costs, as allowed by law;

432. Pre-judgment and post-judgment interest, as allowed by law; and

433. Such further relief as the Court may deem just and proper.

DEMAND FOR JURY TRIAL

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

Dated: February 5, 2024

/s/ Rohit D. Nath

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

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

J. Craig Smyser
SUSMAN GODFREY L.L.P.
1901 Avenue of the Americas, 32nd Floor
New York, New York 10019
Tel.: 212-336-8330
csmyser@susmangodfrey.com

/s/ Rachel Geman
Rachel Geman
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
250 Hudson Street, 8th Floor
New York, New York 10013-1413
Tel.: 212-355-9500
rgeman@lchb.com

Reilly T. Stoler (*pro hac vice forthcoming*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
275 Battery Street, 29th Floor
San Francisco, CA 94111-3339
Tel.: 415-956-1000
rstoler@lchb.com

Wesley Dozier (*pro hac vice*)
LIEFF CABRASER HEIMANN & BERNSTEIN, LLP
222 2nd Avenue, Suite 1640
Nashville, TN 37201
Tel.: 615-313-9000
wdozier@lchb.com

/s/ Scott J. Sholder
Scott J. Sholder
CeCe M. Cole
COWAN DEBAETS ABRAHAMS & SHEPPARD LLP
41 Madison Avenue, 38th Floor
New York, New York 10010
Tel.: 212-974-7474
ssholder@cdas.com
ccole@cdas@com

Attorneys for Plaintiffs and the Proposed Fiction and Nonfiction Author Classes

CERTIFICATE OF SERVICE

I hereby certify this 5th day of February 2024, I caused a true and correct copy of the foregoing to be electronically filed with the Clerk of the court using the CM/ECF system which will send notification to the attorneys of record and is available for viewing and downloading.

/s/ Rohit D. Nath
(Signature)