JOSEPH SAVERI LAW FIRM, LLP
Joseph R. Saveri (SBN 130064)
601 California Street, Suite 1505
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940
jsaveri@saverilawfirm.com

CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP
Bryan L. Clobes (*pro hac vice*)
135 South LaSalle Street, Suite 3210
Chicago, IL 60603
Tel: (312) 782-4880
bclobes@caffertyclobes.com

BOIES SCHILLER FLEXNER LLP
David Boies (*pro hac vice*)
333 Main Street
Armonk, NY 10504
(914) 749-8200
dboies@bsfllp.com

Counsel for Individual and Representative Plaintiffs and the Proposed Class

[Additional counsel on signature page]

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

CATHERINE DENIAL an individual, IAN MCDOWELL, an individual, AND STEVEN SCHWARTZ, an individual v. OPENAI, INC. OPENAI, L.P., OPENAI OPCO, L.L.C., OPENAI GP, L.L.C.; OPENAI STARTUP FUND I, L.P., OPENAI STARTUP FUND GP I, L.L.C., OPENAI STARTUP FUND MANAGEMENT, LLC., and MICROSOFT CORPORATION	**COMPLAINT** **Class Action** **Demand For Jury Trial**

TABLE OF CONTENTS

Plaintiffs Catherine Denial, Ian McDowell, and Steven Schwartz, on behalf of themselves and all others similarly situated, bring this class action complaint ("Complaint") against Defendants OpenAI, Inc.; OpenAI, L.P.; OpenAI OpCo, L.L.C.; OpenAI GP, L.L.C.; OpenAI Startup Fund I, L.P.; OpenAI Startup Fund GP I, L.L.C.; and OpenAI Startup Fund Management, LLC. (collectively, "OpenAI") and Defendant Microsoft Corporation ("Microsoft").

I. INTRODUCTION

1. In the race to dominate the emerging field of generative artificial intelligence ("GenAI"), OpenAI engaged in a systematic campaign of IP theft and data piracy. In carrying out this scheme, OpenAI engaged in unlawful conduct by copying tens of millions of copyrighted works—including articles, essays, and other written works—without the consent of the authors and content creators. OpenAI copied these works from so-called "shadow libraries"[1] or pirated databases that have themselves been the target of numerous legal actions brought by government enforcers for criminal copyright infringement, money laundering, and other claims. In addition to directly downloading massive amounts of pirated data, OpenAI also obtained copies of this data via peer-to-peer file-sharing networks used to facilitate data piracy. This activity violated the rights of countless authors and content creators throughout the United States and undermined foundational principles of innovation through fair competition.

2. OpenAI's disregard of creators' rights was no oversight. OpenAI sought out and torrented, for its commercial use, tens of millions of pirated copyrighted works. OpenAI copied those works without consent, credit, or compensation, and as part of this effort, pirated authors' content through shadow libraries like Library Genesis (aka "libgen" or "LibGen").

3. OpenAI's disregard for copyright, data piracy laws, and ethical standards was not merely a case of corporate negligence. It was part of a strategy to amass a competitive advantage as fast as possible while knowingly flouting existing laws and rights that protect this country's authors and creators.

4. Microsoft was and is the key business partner of OpenAI. Microsoft played a critical role in enabling and profiting from OpenAI's unlawful activities. As a significant investor and operational

[1] In this Complaint, "shadow libraries" refers to any online repositories or large datasets containing copyrighted material of any kind, assembled and made freely accessible online without permission, and in any medium, including but not limited to any copyrighted text, images, audio, video, and programming code.

partner, Microsoft provided the financial resources, cloud infrastructure, and technical support necessary for OpenAI to acquire, process, and exploit massive amounts of stolen IP. Microsoft worked closely with OpenAI in the development, testing and commercialization of OpenAI's generative AI products. Microsoft provided OpenAI with data and environments to develop its infringing models. By integrating OpenAI's models into its own commercial products and services, Microsoft was a key participant in the development of these products and derived substantial and direct economic benefits from their joint conduct. Microsoft acted with knowledge and directly benefited from OpenAI's scheme. Microsoft acted jointly in the unlawful conduct at issue in this case and committed a series of overt acts and other conduct in furtherance of their scheme and common course of conduct. In addition, Microsoft and OpenAI, who are horizontal competitors in the market for training data for LLMs, formed an anti-competitive cartel, working together on an exchange of training data for LLMs, including direct copies of unlawfully acquired works and conspiring to deny sellers or potential sellers of the income that would have been received from the market for training data but for their joint conduct.

5. The ramifications of OpenAI's conduct extend far beyond the immediate harm to individual copyright holders. By building its GenAI models on a foundation of stolen IP, taken without compensation, OpenAI has sought to normalize copyright infringement as the leading business strategy of the GenAI industry for obtaining text data to train their models. Microsoft's support and integration of these unlawfully trained models into its own products magnify that impact, further foreclosing actual competition and future competition, ensuring financial benefits and preventing the entry and development of competitive market forces in the future.

6. Plaintiffs, representing a class of copyright owners whose works have been unlawfully acquired and exploited, seek not only to hold OpenAI and Microsoft accountable for their actions, but also to deter similar conduct in the future by other GenAI companies and bad actors who seek to exploit their works.

II. OVERVIEW

7. ChatGPT is a web-based software application created, maintained, and sold by OpenAI. ChatGPT is powered by AI software programs also known as *large language models* ("LLMs"). Vast quantities of data have been integral—indeed essential—to the development of these products and will

continue to be so in the future. There is no substitute for this data. Two of OpenAI's most popular models are called GPT-3.5 and GPT-4. More are on the way.

8. LLMs, rather than being programmed in the traditional way, are "trained." The so-called training process starts with copying massive amounts of text, often called *raw data.* That text is then processed and expressive information is extracted from it. The resulting corpus of text is called the *training dataset*. During the training process, computer engineers copy the text and program the LLM to ingest the text as part of the LLM training dataset. At the end of training, the LLM is able to mimic the expressive information found in the training dataset, thereby emitting convincingly naturalistic text output in response to user prompts. This process of copying and regurgitation is key to the basic function of LLMs. It is true of the LLMs at issue in this case.

9. LLM output is entirely and uniquely reliant on the material in its training dataset. In other words, every time it assembles a text output, the LLM relies on the entirety of information it extracted from its training dataset. The fact it does so is crucial to its operation.

10. Plaintiffs and Class members are authors of text materials, including articles, essays, and other written works, which OpenAI copied and used to train its LLMs.[2] Plaintiffs and Class members hold copyrights in these published works. Plaintiffs and Class members did not consent to the copying or use of their works as training data by OpenAI for its LLMs. OpenAI did not obtain permission or compensation to Plaintiffs for doing so. Instead, OpenAI copied, commercially exploited and took without compensation these valuable copyrighted materials without permission and, at times, through illegal torrenting that violates copyright and data privacy laws.

Defendants, individually and collectively, through the use of OpenAI's LLMs and ChatGPT, benefit commercially and profit significantly from their use of Plaintiffs' and Class members' copyrighted works.

[2] OpenAI's LLMs include any models in development or released commercially, even if not to public sources, and irrespective of whether those models underlie ChatGPT. For purposes of this Complaint, OpenAI LLMs includes all products derived by OpenAI or Microsoft from OpenAI's LLMs.

III. JURISDICTION AND VENUE

11. This Court has subject-matter jurisdiction under 28 U.S.C. § 1331, including because this case arises under the Copyright Act (12 U.S.C. § 101, *et seq.*).

12. Jurisdiction and venue is proper in this judicial district under 28 U.S.C. § 1391(c)(2) because Defendant OpenAI, Inc. is headquartered in this District, and thus a substantial part of the events giving rise to the claims occurred in this, and a substantial portion of the affected interstate trade and commerce was carried out in this District. Each Defendant has transacted business, maintained substantial contacts, and/or committed overt acts in furtherance of the illegal scheme and conspiracy throughout the United States, including in this District. Defendants' conduct has had the intended and foreseeable effect of causing injury to persons residing in, located in, or doing business throughout the United States, including in this District. Defendant Microsoft Corporation, for its part, maintains substantial offices and business operations in this District.

13. Pursuant to Civil Local Rule 3-2(c), assignment of this case to the San Francisco Division is proper because this case pertains to intellectual-property rights, which is a district-wide case category under General Order No. 44, and therefore venue is proper in any courthouse in this District.

IV. PARTIES

A. Plaintiffs

14. Plaintiff Catherine Denial is a writer who lives in Illinois and owns a registered copyright in multiple works, including *A proper light before the country: the shifting politics of gender and kinship among the Dakota, Ojibwe and non-native communities of the Upper Midwest, 1825-1845*.[3]

15. Plaintiff Ian McDowell is a writer who lives in North Carolina and owns copyrights in multiple works, including *Wilmington Massacre was Confederacy's Revenge*.

16. Plaintiff Steven Schwartz is a writer who lives in Arizona and owns copyrights in multiple works, including *A Comprehensive System for Item Analysis in Psychological Scale Construction*.

[3] Registration No. TX0006474253.

17. A non-exhaustive list of copyrights owned by Plaintiffs is shown in Exhibit A. Together, and for the purposes of this Complaint, these works are collectively referred to as the **Selected Infringed Works**.

B. Defendants

18. Defendant OpenAI, Inc. is a Delaware nonprofit corporation with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI Inc. owns and controls the other OpenAI entities.

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

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

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

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

23. Defendant OpenAI Startup Fund GP I, L.L.C. ("OpenAI Startup Fund GP I") is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San

Francisco, CA 94110. OpenAI Startup Fund GP I is the general partner of OpenAI Startup Fund I. OpenAI Startup Fund GP I is a party to the unlawful conduct alleged herein. OpenAI Startup Fund GP I manages and operates the day-to-day business and affairs of OpenAI Startup Fund I.

24. Defendant OpenAI Startup Fund Management, LLC ("OpenAI Startup Fund Management") is a Delaware limited liability company with its principal place of business located at 3180 18th Street, San Francisco, CA 94110. OpenAI Startup Fund Management is a party to the unlawful conduct alleged herein. OpenAI Startup Fund Management was aware of the unlawful conduct alleged herein and exercised control over OpenAI, L.P. throughout the Class Period.

25. Defendant Microsoft Corporation is a Washington corporation with its principal place of business located at One Microsoft Way, Redmond, Washington 98052. It also maintains multiple offices and facilities, key employees, and a sizable customer population within this District, and it conducts business in this District.

C. Agents and Co-Conspirators

26. The unlawful acts alleged against Defendants were authorized, ordered, or performed by Defendants' respective officers, agents, employees, representatives, or shareholders while actively engaged in the management, direction, or control of Defendants' businesses or affairs. Defendants' agents operated under the explicit and apparent authority of their principals. Each Defendant, and its subsidiaries, affiliates, and agents, operated as a single unified entity.

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

V. FACTUAL ALLEGATIONS

A. Background on OpenAI's LLMs

28. OpenAI creates, markets and sells artificial intelligence ("AI") software products. Generally, AI software is designed to attempt to algorithmically simulate human reasoning or inference, often using statistical methods. AI models do not think or reason like humans. AI models mimic certain human interactions, including, for example, by providing answers to questions or user prompts.

29. Certain AI products created and sold by OpenAI are known as *large language models*, or LLMs for short. An LLM is AI software designed to parse and emit natural-sounding text generally in response to user inquiries or prompts. Though an LLM is a software program, written by computer scientists or engineers, it is not created in the way most software programs are—that is, by human software engineers writing code. Rather, LLMs rely on training by copying massive amounts of text data from various sources and feeding these copies into a computer model at various stages of the LLM process.

30. The training of an LLM begins with the collection of *raw data*. Raw data includes textual material collected from various sources—some legal (e.g., Project Gutenberg, an online repository of out-of-copyright books)—and some not (e.g., notorious shadow libraries or pirated material like LibGen). Once gathered, raw data is processed—for instance, processing can include removing low-quality raw data and organizing the dataset to make training easier. The resulting processed data comprises the *training dataset* that is fed to the LLM.

31. During training, the LLM copies each piece of text in the training dataset and extracts expressive information from it. The LLM progressively adjusts its output to more closely resemble the sequences of words copied from the training dataset. Once the LLM has copied and ingested all this text, it is frequently able to emit convincing simulations of natural written language as it appears in the training dataset.

32. Much of the raw data OpenAI acquired and uses in its training datasets comes from copyrighted material—encompassing a range of text data such as articles, essays, and other written works authored by Plaintiffs and other copyright holders—that were copied by OpenAI without consent, credit, or compensation, including through illegal torrenting from shadow libraries like LibGen or by crawling and scraping the internet with little to no regard for the copyright status of the scraped materials or any terms and conditions proscribing such scraping. OpenAI and Microsoft could have obtained this material legally, in compliance with copyright and other laws, but chose not do so.

33. Authors, including Plaintiffs, typically publish their works with certain copyright management information, or "CMI." This information generally includes the title of the work, the ISBN number or copyright number, the author's name, the copyright holder's name, and terms and

conditions of use. This information is usually displayed prominently in the introductory or bibliographic sections of published materials, including articles, essays, and other written works.

34. OpenAI made a series of LLMs, including but not limited to GPT-1 (released June 2018), GPT-2 (February 2019), GPT-3 (May 2020), GPT-3.5 (March 2022), GPT-4 (March 2023) and other variations still in development and set to be released. "GPT" is an abbreviation for "generative pre-trained transformer," where *pre-trained* refers to the use of text data for training, *generative* refers to the model's ability to emit text, and *transformer* refers to the underlying training algorithm. OpenAI offers certain language models in variant forms: for instance, the GPT-4 family of models includes publicly accessible variants called 'gpt-4-0125-preview,' 'gpt-4-turbo-preview,' and 'gpt-4-32k;' the GPT-3.5 Turbo family of models includes publicly accessible variants called 'gpt-3.5-turbo-0125,' 'gpt-3.5-turbo-1106,' and 'gpt-3.5-turbo-instruct.' Starting in December 2024, OpenAI also began releasing a series of "reasoning" LLMs (LLMs designed to accomplish more complex reasoning tasks like solving puzzles or riddles): o1, o1-mini, o3, and o3-mini. There are other models as well (https://platform.openai.com/docs/models), and OpenAI continues to develop more: In an interview with the Financial Times in November 2023, OpenAI CEO Sam Altman confirmed OpenAI was developing GPT-5. More than a year later, in a February 12, 2025 post on the social media platform X, Altman confirmed GPT-5 is still under development and said OpenAI will first release GPT-4.5.

35. While some of OpenAI's LLMs and GPT language-model variants are publicly available and free to download, others require paid monthly or annual subscriptions. OpenAI has also made other language-model variants that are in commercial use and integrated into products manufactured and sold by others.

36. OpenAI may use many kinds of materials to train its AI systems and models. But copyrighted text data has always been a key ingredient used by OpenAI and Microsoft in training datasets for its LLMs.

37. In addition to the data necessary to train the LLMs, another key input is computing power. LLMs require large, fast, sophisticated computing. It has been reported that the training dataset for GPT-4 contained over 1.4 trillion tokens. Training on this data, which sometimes occurs over one or more epochs, can take days or weeks.

38. There is already a substantial market for AI training data with many willing buyers and sellers. The market is valued by some analysts at approximately 2.92 billion USD in 2024 and projected to exceed 17 billion USD by 2032. There is also a market for LLM training data, which includes copyrighted literary works such as fiction and non-fiction. Recognizing the economic value copyrighted works have as training data, GenAI companies have negotiated and entered into licensing agreements to use copyrighted as training data.

39. OpenAI and Microsoft are major players in the relevant market for LLM training data. They are also horizontal competitors in the market for LLM training data, whether it be for registered copyrighted works, or for unregistered textual works. OpenAI has entered into deals with a variety of organizations such as Axel Springer, the Financial Times, Reddit and the Associated Press to license their content as training data for its LLMs. Microsoft has also entered into licensing deals with organizations for licensing training data for LLMs, including a November 2024 deal with book publisher HarperCollins to use nonfiction works (and almost certainly copyrighted) as training data.

40. OpenAI and Microsoft recognize the value of the material used to train its LLMs, whether it is copyrighted works or unregistered textual material. OpenAI and Microsoft recognize that the textual material is protected by copyright and other laws that protect authors and prohibit taking of the textual material without permission or compensation. OpenAI and Microsoft knowingly and willfully violated those laws. OpenAI and Microsoft knowingly and willfully made the crass business decision to take what they could.

B. OpenAI targets and steals copyrighted works

41. In its June 2018 paper introducing GPT-1 (called "Improving Language Understanding by Generative Pre-Training"), OpenAI revealed that it trained GPT-1 on BookCorpus, a collection of "over 7,000 unique unpublished books from a variety of genres including Adventure, Fantasy, and Romance." OpenAI confirmed why a dataset of books was so valuable: "Crucially, it contains long stretches of contiguous text, which allows the generative model to learn to condition on long-range information." Hundreds of LLMs have been trained on BookCorpus, including those made by OpenAI, Google, Amazon, and others.

42. BookCorpus, however, is an illicit dataset of pirated books. It was assembled in 2015 by a team of AI researchers for the purpose of training language models. They copied the books from the website www.smashwords.com, which makes unpublished novels available online at no cost. Those novels are largely under copyright and were copied into the BookCorpus dataset without consent, credit, or compensation to the authors.

43. Despite these known issues, OpenAI proceeded to copy and use BookCorpus for training its LLMs, including GPT-1. Their decision underscores a pattern of negligence and disregard for the legal and ethical standards governing the use of copyrighted materials.

44. OpenAI also accessed and copied vast quantities of copyrighted works, including Plaintiffs' works, through various other illegal sources, including from notorious shadow libraries such as LibGen. Sometimes, OpenAI did so by torrenting and seeding these pirated works—in other words, downloading and sharing Plaintiffs' and others' copyrighted works using peer-to-peer networks.

45. In the July 2020 paper introducing GPT-3 (called "Language Models are Few-Shot Learners"), OpenAI disclosed that 15% of its enormous GPT-3 training dataset came from "two internet-based books corpora," which OpenAI pretextually referred to as "Books1" and "Books2," concealing the names used internally by OpenAI employees when referring to these datasets: Libgen1 and Libgen 2 (collectively, "LibGen Datasets"). At the time, the true source and provenance of "Books1" and "Books2" was a mystery, which OpenAI knowingly concealed.

46. Tellingly, OpenAI never publicly revealed which copyrighted books and other works are part of the LibGen Datasets—though there are some clues. First, OpenAI admitted these are "internet-based books corpora." Second, the LibGen Datasets are apparently much larger than BookCorpus. Microsoft was aware that the material described had in fact been obtained from pirate websites or other illicit sources.

47. The only "internet-based books corpora" that have ever made that quantity of material available are shadow libraries like LibGen, Z-Library (aka B-ok), Sci-Hub, Internet Archive, and Bibliotik. These datasets are large collections of pirated materials stolen from authors around the world. *See Cengage Learning, Inc. v. Library Genesis*, Case No. 23-cv-08136 (S.D.N.Y. Sep. 24, 2024), Dkt. 36 (permanently enjoining LibGen due to copyright infringement); *Hachette Book Group, Inc. v.*

Internet Archive, Case No. 20-cv-04160-JGK-OTW, (S.D.N.Y. Aug. 11, 2023), Dkt. 213 (permanently enjoining Internet Archive due to copyright infringement). OpenAI accessed these pirated databases and illegally downloaded and torrented mass quantities of copyrighted works.

48. After accessing and copying these stolen works, OpenAI compiled them into training datasets.

49. On information and belief, one or more of the Selected Infringed Works for each Plaintiff are found in OpenAI's datasets.

C. OpenAI accessed and copied vast amounts of copyrighted works using peer-to-peer file sharing

50. OpenAI's use of LibGen demonstrates that OpenAI knowingly and intentionally torrented large volumes of digital files containing pirated copyrighted works, including Plaintiffs' works.

51. OpenAI's reliance on torrenting is especially alarming because obtaining data through peer-to-peer sharing generally involves not just copying and hosting pirated data, but uploading, distributing or "seeding" pirated data. In other words, to acquire a torrented file, a user must typically participate in a data exchange: data is downloaded from fellow pirates while simultaneously uploaded to fellow pirates. Thus, it is plausible that OpenAI was not only downloading and copying massive amounts of pirated copyrighted works but also distributing them to other IP pirates in the swarm.

52. Microsoft knew, or should have known, that OpenAI had obtained pirated copyrighted works by torrenting.

D. OpenAI attempts to conceal its use of torrented copyrighted data

53. OpenAI tried to hide its piracy in at least two ways.

54. *First*, at the individual file level, OpenAI wanted to conceal and obscure its reliance on copyrighted data by stripping copyright-identifying information from the files it stole.

55. *Second*, and more broadly, OpenAI sought to obscure the origins of the pirated data it accessed and copied for use with its LLMs. For example, OpenAI coined the term "Books1" and "Books2," concealing the true names for these datasets: Libgen1 and Libgen 2 (collectively, "LibGen Datasets").

56. OpenAI’s concealment of the data it acquired and processed into training datasets for its LLMs continued for years. In March 2023, OpenAI’s paper introducing GPT-4 contained no information about its dataset at all, claiming that “[g]iven both the competitive landscape and the safety implications of large-scale models like GPT-4, this report contains no further details about … dataset construction.”

57. Microsoft knew, or should have known, that OpenAI concealed its torrenting of pirated copyrighted data and provided incorrect pretextual explanations. Despite this knowledge or information, Microsoft did not disclose these facts.

E. OpenAI steals additional copyrighted material by crawling and scraping the internet

58. OpenAI’s LLM training datasets have included a wide array of other copyrighted materials, such as articles, essays, and other written works, taken without permission. OpenAI has used data repositories such as Common Crawl, a publicly available dataset that scrapes vast amounts of internet content, including copyrighted material from websites, blogs, and news articles. This extensive use of Common Crawl derived data underscores the breadth of copyrighted materials taken by OpenAI and ingested by its models, which extends far beyond books to encompass a diverse range of written works.

59. In contrast with its circumspection about shadow libraries, OpenAI *has* publicly admitted to using Common Crawl to develop its LLMs. In a paper authored by several AI researchers, including OpenAI engineers who worked directly on GPT-3, the downloading and use of Common Crawl is discussed openly. This paper, *Language Models are Few-Shot Learners* by Tom B. Brown et al.,[4] admits that “The CommonCrawl data was downloaded from 41 shards of monthly CommonCrawl covering 2016 to 2019, constituting 45TB of compressed plaintext . . .” This extensive dataset includes a vast array of copyrighted text from websites, blogs, and news articles, highlighting the breadth of sources used to train OpenAI’s models.

60. Common Crawl publishes insights into each of the “crawls’ they conduct, including the top domains included in each dataset. The insights published by Common Crawl on the May through

[4] https://arxiv.org/pdf/2005.14165

June 2018 crawl reveal that the domains WordPress.com and BlogSpot.com were both in the top 20 domains crawled for data.[5] Those platforms host millions of blogs and articles, many of which are copyrighted. The inclusion of such domains in the Common Crawl dataset underscores OpenAI's extensive copying and use of diverse copyrighted materials, including articles, essays, and other written works with its LLMs.

61. Because Common Crawl copies essentially the entire internet, on information and belief, the Selected Infringed Works (or parts of them) can be found in OpenAI's common crawl datasets and similar datasets that are the product of scraping and crawling.

F. OpenAI and Microsoft knowingly profit from stealing copyrighted material

62. OpenAI's practices in acquiring text data for training its LLMs have come under significant scrutiny, particularly regarding its use of peer-to-peer file-sharing networks to acquire massive quantities of copyrighted material from shadow libraries such as LibGen. Downloading, including by torrenting, pirated IP is not only unlawful but also negates any attempt to claim fair use.

63. At every turn, when faced with what it saw as a choice to respect copyright law and intellectual property rights or gain competitive advantage, OpenAI knowingly chose the latter.

VI. MICROSOFT'S INVOLVEMENT

64. Defendant Microsoft has played a significant role in the development and operation of OpenAI, both financially and operationally. Microsoft has invested billions of dollars into OpenAI. This financial backing not only provided OpenAI with the resources necessary to develop its AI systems, but has also given Microsoft knowledge and awareness of OpenAI's conduct. It has also provided Microsoft substantial influence over, and even the ability to control, OpenAI's operations and strategic decisions. Microsoft has sought, obtained and maintained such influence because Microsoft attempted and failed to design its own LLM and other AI products. In order to support its business strategy, Microsoft provided OpenAI with infrastructure such as access to its cloud network and computing power in order for OpenAI to develop its infringing technology. Without Microsoft, OpenAI would not have been able to have done so.

[5] https://commoncrawl.github.io/cc-webgraph-statistics/

65. In addition, incorporation of OpenAI’s GenAI technology into Microsoft’s commercial products and services, such as Azure, Microsoft Office, and other enterprise solutions was a key component of the partnership. Microsoft and OpenAI were also commercializing other products, including Microsoft’s Bing search engine. Microsoft’s incorporation of OpenAI’s LLM technology was a central pillar of both companies’ commercialization of their LLM technology. This integration has allowed Microsoft to benefit directly from the AI systems developed by OpenAI, including the use of copyrighted material obtained unlawfully in developing and training its LLMs. This required substantial integration of the software engineers and computer scientists of the two companies. Indeed, particular teams of officers and other employees of the companies were formed to work together. Among other things, these joint endeavors provided the knowledge and information, shared by the two companies, of the unlawful conduct at issue in this case. By incorporating OpenAI’s models into its products, Microsoft has effectively directly endorsed and profited from OpenAI’s activities.

66. Given Microsoft’s deep involvement in OpenAI’s operations, Microsoft was aware of, and upon information and belief, approved, OpenAI’s piracy and other unlawful data acquisition practices. Public statements and internal communications suggest that Microsoft had access to information about the sources of OpenAI’s training data.

67. Microsoft was not just aware of OpenAI’s approach; Microsoft encouraged it and profited from it because Microsoft received OpenAI’s training data. On information and belief, Microsoft received that data as part of a trade in which Microsoft provided OpenAI with Bing crawl data and OpenAI provided Microsoft with its training data. On information and belief, OpenAI’s side of the trade included LibGen.

68. Incredibly, in a joint press release in July 2019, Microsoft and OpenAI stated: “We are dedicated to ensuring that our AI technologies are developed and used in a manner that is ethical and respects the rights of all individuals.” (Microsoft-OpenAI Press Release, July 2019.) Hardly. On information and belief, at this very time, OpenAI, with Microsoft’s knowledge, was illegally torrenting massive amounts of copyrighted works from shadow libraries such as LibGen, in addition to web crawling and scraping additional copyrighted works, in violation of the rights of copyright owners like

Plaintiffs. Microsoft condoned and supported OpenAI's mass IP piracy, fully aware of its legal implications.

69. Worse, by blessing—and profiting from—OpenAI's data theft, Microsoft and OpenAI artificially limited the market for training data. In a well-functioning market, OpenAI and Microsoft would compete to buy and license high-quality training data from copyright holders. But OpenAI and Microsoft suppressed that market by sharing stolen data—essentially agreement to a price of zero.

70. The companies independently and together recognized that one of their shared common purposes was to limit the market for AI training data. As part of their agreement to work together—including OpenAI's agreement to provide Microsoft stolen copyrighted works—Microsoft agreed to reduce, prevent and foreclose competition in the AI training space. Had the two companies not embarked on their common course of conduct, the market would have grown and developed, and owners of copyrights could have, and would have, been able to obtain fair value for their work in a marketplace where prices would have been set by willing buyers and sellers.

71. Microsoft eventually began to assist OpenAI in its data acquisition efforts. Microsoft ultimately obtained and shared with OpenAI training data derived, on information and belief, from scraping and crawling the public internet.

72. Microsoft's stated goal was to acquire AI training data legally through business partnerships.

73. Yet, on information and belief, Microsoft also sought to acquire and use vast quantities of pirated works from LibGen for training purposes.

74. Microsoft's Azure cloud platform has also been a critical infrastructure component for OpenAI, providing the computing power necessary to train OpenAI's LLMs. By providing this infrastructure, Microsoft facilitated the processing and storage of the vast amounts of data acquired and used by OpenAI, including the vast quantities of copyrighted works that OpenAI illegally torrented from shadow libraries of pirated material. Microsoft's role in providing the technological backbone for OpenAI's operations implicates it in the unlawful data-acquisition scheme employed by OpenAI. Not only is Microsoft providing critical infrastructure to store pirated works, but it intends to allow OpenAI

the use of the Azure cloud platform for the next decade without any indication that it will stop storing these pirated works.

75. Microsoft has made several other public statements which provide pretextual and inaccurate versions of OpenAI's practices of acquiring AI training data. For example, Microsoft CEO Satya Nadella wrote, "We are committed to the highest standards of data ethics and transparency in all our AI endeavors." (Microsoft Blog, January 2021) (a line that OpenAI CEO Sam Altman echoed when he told the Financial Times in November 2023, "We are committed to using only publicly available data and data we have the right to use"). Such statements were knowingly deceptive and misleading given the evidence that OpenAI, under Microsoft's substantial influence and support, purposely, repeatedly, and illegally obtained and used pirated copyrighted material to train its AI models.

76. Additionally, Microsoft claimed that its partnership with OpenAI was built on a foundation of "trust and integrity" (Microsoft Annual Report, 2022), despite knowing full well the illegal data-acquisition methods employed by OpenAI. Microsoft internal personnel knew, or should have known, that this was at odds with the truth, and Microsoft's own internal strategy and designs. By providing the financial resources, technological infrastructure, and strategic support necessary for OpenAI to develop its models, Microsoft has contributed materially to the unauthorized accessing and infringement of Plaintiffs' copyrighted works and conspired alongside OpenAI to violate Plaintiffs' rights.

77. Put another way, OpenAI and Microsoft shared a common plan and purpose: to access Plaintiffs' copyrighted works without authorization, infringe on Plaintiffs' copyrights, and conceal both the origins of the data pirated to train OpenAI's and Microsoft's AI models and the means with which it was acquired. Microsoft knew, or should have known, the reasonable and foreseeable results of this conduct.

VII. INTERROGATING THE OPENAI LANGUAGE MODELS USING CHATGPT

78. ChatGPT is an LLM created and sold by OpenAI. As its name suggests, ChatGPT is designed to offer a conversational style of interaction with a user. OpenAI offers ChatGPT through a web interface to users for free and via paid subscriptions.

79. OpenAI also offers ChatGPT to software developers through an application-programming interface (or "API"). The API allows developers to write programs that exchange data with ChatGPT. Access to ChatGPT through the API is billed on the basis of usage.

80. Regardless of how it is accessed—either through the web interface or through the API—ChatGPT allows users to enter text prompts, which ChatGPT then attempts to respond to in a "natural" way, *i.e.*, ChatGPT can generate coherent and fluent answers that closely mimic human language. If a user prompts ChatGPT with a question, ChatGPT will answer. If a user prompts ChatGPT with a command, ChatGPT will obey. And if a user prompts ChatGPT to summarize a copyrighted book, ChatGPT will do so.

81. ChatGPT's output, like other LLMs' outputs, relies on the data upon which it is trained to generate "new" content. LLMs generate output using patterns and connections drawn from the training data. For example, if an LLM is prompted to generate writing in the style of a certain author, the LLM will attempt to generate content based on the patterns and connections it learned from analyzing that author's works in its training data.

82. ChatGPT can accurately summarize and even quote from copyrighted materials because those materials were copied by OpenAI using peer-to-peer file-sharing networks or crawling and scraping the internet and ingested by the underlying OpenAI model as part of its training data.

VIII. CLASS ALLEGATIONS

83. The 'Class Period' as defined in this Complaint begins on at least January 1, 2018, and runs through the present, during which time OpenAI engaged in the unauthorized acquisition and use of copyrighted text data, including but not limited to books, articles, essays, and other written works. On information and belief, that unlawful conduct may have begun earlier than January 1, 2018, and Plaintiffs reserve the right to amend the Class Period to comport with the facts and evidence uncovered during further investigation or through discovery.

84. Plaintiffs bring this action for damages and injunctive relief as a class action under Federal Rules of Civil Procedure 23(a), 23(b)(2), and 23(b)(3), on behalf of the following Class ("Non-Book Infringement Class"):

> **All persons or entities domiciled in the United States that own a United States copyright in any textual work, where the work is registered with the United States Copyright Office, but are not assigned one or more International Standard Books Number(s) (ISBN) or Amazon Standard Identification Number(s) (ASIN).**

This Class definition excludes:

a. any of the Defendants named herein;

b. any of the Defendants' co-conspirators;

c. any of Defendants' parent companies, subsidiaries, and affiliates;

d. any of Defendants' officers, directors, management, employees, subsidiaries, affiliates, or agents;

e. all governmental entities; and

f. the judges and chambers staff in this case, including on appeal, as well as any members of their immediate families. This exclusion applies regardless of the type of copyrighted text material involved.

85. Plaintiffs bring this action for damages and injunctive relief as a class action under Federal Rules of Civil Procedure 23(a), 23(b)(2), and 23(b)(3), on behalf of the following Class ("Unregistered Copyright Holders Class"):

> **All persons or entities domiciled in the United States that own a United States copyright in any text data, who have not registered their works with the United States Copyright Office, including but not limited to books, articles, essays, and other written works, that was accessed, copied, or used by OpenAI during the Class Period.**

This Class definition excludes:

a. any of the Defendants named herein;

b. any of the Defendants' co-conspirators;

c. any of Defendants' parent companies, subsidiaries, and affiliates;

d. any of Defendants' officers, directors, management, employees, subsidiaries, affiliates, or agents;

e. all governmental entities; and

f. the judges and chambers staff in this case, including on appeal, as well as any members of their immediate families. This exclusion applies regardless of the type of copyrighted text material involved.

86. **Numerosity**. Plaintiffs do not know the exact number of members in each Class. This information is in the exclusive control of Defendants. On information and belief, there are at least hundreds of thousands of members in the Classes geographically dispersed throughout the United States, encompassing owners of a wide range of copyrighted text materials, including articles, essays, and other written works. Therefore, joinder of all members of each Class in the prosecution of this action is impracticable.

87. **Typicality**. Plaintiffs' claims are typical of the claims of other members of each Class because Plaintiffs and all members of each Class were damaged by the same wrongful conduct of Defendants as alleged herein, including the unauthorized access and CMI stripping. The relief sought herein also is common to all members of each Class.

88. **Adequacy**. Plaintiffs will fairly and adequately represent the interests of the members of the respective Classes because the Plaintiffs have experienced the same harms as the members of the Class, including the unauthorized access and CMI stripping, and Plaintiffs have no conflicts with any other members of the Class. Furthermore, Plaintiffs retained and are represented by sophisticated and competent counsel who are experienced in prosecuting federal and state class actions, as well as other complex litigation.

89. **Commonality and predominance**. Numerous questions of law or fact common to each Class arise from Defendants' conduct:

a. whether Defendants' conduct alleged herein, including but not limited to the unlawful use of peer-to-peer file-sharing networks to access and copy pirated copyrighted works and the misrepresentation of data sources used to train OpenAI's models, constitutes Unfair Competition under California Business and Professions Code § 17200 *et seq.*;

b. whether this Court should enjoin Defendants from engaging in the unlawful conduct alleged herein, including the unauthorized access, copying, and use of Plaintiffs' copyrighted material, and what the scope of that injunction would be, including but not limited to whether OpenAI and Microsoft should be allowed to continue offering their suite of products trained on Plaintiffs' works unlawfully;

c. whether Defendants' actions in copying mass quantities of text material from the internet, including but not limited to Plaintiff Catherine Denial's work, *A proper light before the country: the shifting politics of gender and kinship among the Dakota, Ojibwe and non-native communities of the Upper Midwest, 1825-1845*, without Plaintiffs' permission, constitute direct copyright infringement under 17 U.S.C. § 501;

d. whether Microsoft, by providing financial resources, technical infrastructure, and strategic support to OpenAI, had the right and ability to supervise and control OpenAI's infringing activity and failed to exercise such supervision and control;

e. whether Defendants' conduct, including but not limited to, the unauthorized use of peer-to-peer file-sharing networks to access and copy pirated copyrighted works and the misrepresentation of data sources used to train OpenAI's models, constitutes Unfair Competition under California Business and Professions Code § 17200 *et seq.*;

f. whether OpenAI's unauthorized access and use of Plaintiffs' copyrighted Infringed Works, including but not limited to by torrenting digital copies from shadow libraries such as LibGen, constitute violations of the California Comprehensive Computer Data Access and Fraud Act (CDAFA), Cal. Penal Code § 502;

g. whether OpenAI circumvented technological measures that control access to Plaintiffs' copyrighted works, including by torrenting these and other copyrighted materials from shadow libraries such as LibGen, in violation of the Digital Millennium Copyright Act (DMCA), 17 U.S.C. § 1201, and whether OpenAI's removal of copyright management information (CMI) from copyrighted works, including but not limited to the Selected Infringed Works, constitutes a violation of the DMCA, 17 U.S.C. § 1201(b)(1);

h. whether OpenAI's unauthorized acquisition and use of Plaintiffs' copyrighted works, including but not limited to, by torrenting them from shadow libraries such as LibGen constitute conversion under California law;

i. whether OpenAI has been unjustly enriched by its unauthorized access, copying, and use of Plaintiffs' copyrighted material to train its GenAI models, deriving significant commercial benefits and profits from this use, and whether Microsoft directly benefited from this unjust enrichment by integrating OpenAI's models into its own products and services, leveraging Plaintiffs' unlawfully obtained intellectual property to enhance its offerings and increase its profits;

k. whether OpenAI's unauthorized access and use of data from websites that prohibit such activities in their terms and conditions constitute violations of the Computer Fraud and Abuse Act (CFAA), 18 U.S.C. § 1030, and whether Microsoft contributed to these actions by providing infrastructure, financial support, and resources necessary for OpenAI to engage in these unlawful activities; and

l. whether Defendants unlawfully acquired copyrighted material from the internet, including but not limited to by torrenting works from shadow libraries and violating websites' terms and conditions, and used the material to develop GenAI models and products in order to generate profits, in violation of California Penal Code § 496(a), (c).

These and other questions of law and fact are common to each Class and predominate over any questions affecting the Class members individually, particularly given the widespread and systematic nature of Defendants' unauthorized access, copying, and use of a vast amount of copyrighted text data and works.

90. **Other class considerations**. Defendants acted on grounds generally applicable to the Class. This class action is superior to alternatives, if any, for the fair and efficient adjudication of this controversy. Prosecuting the claims pleaded herein as a class action will eliminate the possibility of repetitive litigation and inconsistent results. There will be no material difficulty in the management of

this case as a class action. Furthermore, final injunctive relief is appropriate with respect to the Class as a whole, given the systematic and widespread nature of Defendants' unauthorized and unlawful access, copying, and use of copyrighted text data and works.

91. The prosecution of separate actions by individual Class members would create the risk of inconsistent or varying adjudications, establishing incompatible standards of conduct for Defendants, particularly given the widespread and systematic nature of Defendants' unauthorized and unlawful access, copying, and use of a vast amount of copyrighted text data and works.

IX. CLAIMS FOR RELIEF

COUNT 1

Direct Copyright Infringement

17 U.S.C. § 501

(Against OpenAI and Microsoft on Behalf of the Non-Book Infringement Class)

92. Plaintiffs incorporate by reference the preceding factual allegations.

93. Plaintiffs hold the exclusive rights to works, including but not limited to the Selected Infringing Works, under 17 U.S.C. § 106.

94. Plaintiffs never authorized OpenAI or Microsoft to make copies of these texts, including but not limited to the Selected Infringed Works, or any portion thereof, to make derivative works, to publicly display copies (or derivative works), or to distribute copies (or derivative works). All those rights belong exclusively to Plaintiffs under copyright law.

95. On information and belief, in connection with training its LLMs, OpenAI and Microsoft copied mass quantities of text material, including but not limited to the Plaintiff Catherine Denial's work, *A proper light before the country: the shifting politics of gender and kinship among the Dakota, Ojibwe and non-native communities of the Upper Midwest, 1825-1845*, in digital formats, including by torrenting them from one or more shadow libraries or pirate websites, such as LibGen.

96. To the extent not already specified, Plaintiffs incorporate by reference Exhibit A, which identifies by title, author, and (where applicable), the registration number for the copyrighted work at issue. Plaintiffs allege, on information and belief, that each of these works, or substantial portions thereof, were included in the datasets used by Defendants to train their large language models, as

evidenced by the ability of those models to generate verbatim or near-verbatim excerpts from said works upon prompt. Plaintiff Catherine Denial further alleges that Defendants' acts of copying, storing, and using these works in the course of training and deploying their commercial AI products constitute unauthorized reproductions in violation of 17 U.S.C. § 106.

97. OpenAI and Microsoft made additional copies of and/or from texts including, but not limited to, the Selected Infringed Works during its LLM training process without Plaintiffs' permission.

98. Licensing copyrighted material to train AI models is plainly feasible. It already happens. Indeed, OpenAI itself has licensed copyrighted material for training its LLMs. For instance, OpenAI reached agreements with the Associated Press and Axel Springer to license text data and material for its LLM training. OpenAI has reportedly been in negotiations with other publishers as well. Microsoft has also sought to negotiate and obtain licenses for text works as training data for training LLMs.

99. Microsoft played a pivotal role in facilitating OpenAI's infringing activities by providing the financial resources, technical infrastructure, and strategic support necessary for OpenAI to develop and expand its AI systems. This includes the use of Microsoft Azure, which powered the large-scale training of OpenAI's models using Plaintiffs' (and others') copyrighted material without authorization.

100. Microsoft and OpenAI also engaged in exchanges of training data. OpenAI received training data from Microsoft based on various web scrapes of the internet. Microsoft also received training data, including direct copies of the copyrighted works, directly from OpenAI for its own use.

101. On information and belief, OpenAI and Microsoft's infringing conduct was and continues to be willful, continuing to infringe on Plaintiff Catherine Denial and Plaintiffs and members of the Non-Book Infringement Class' exclusive rights knowing they were profiting from widescale copyright infringement.

102. Plaintiff Catherine Denial and members of the Non-Book Infringement Class have been injured by Defendants' acts of direct copyright infringement. Plaintiff Catherine Denial and members of the Non-Book Infringement Class are entitled to statutory damages, actual damages, restitution of profits, and/or other remedies provided by law.

COUNT 2

Vicarious Copyright Infringement

17 U.S.C. § 501

(Against Microsoft, OpenAI Inc. and OpenAI GP LLC on Behalf of the Non-Book Infringement Class)

103. Plaintiffs incorporate by reference the preceding factual allegations.

104. As explained above in Count 1, OpenAI directly infringed copyrights owned by Plaintiff Catherine Denial and members of the Non-Book Infringement Class.

105. Microsoft directly benefitted from that infringement both because of its partnership with OpenAI and because it incorporates OpenAI's products—which infringe Plaintiffs' copyrights—into Microsoft's products.

106. Defendant Microsoft, by virtue of its substantial investment, contractual arrangements, and operational integration with OpenAI, had both the legal right and practical ability to supervise and control the infringing activities of OpenAI, including but not limited to the acquisition and use of copyrighted works in training datasets. Microsoft derived a direct financial benefit from the infringement by incorporating the resulting AI models into its own commercial products and services, thereby increasing its revenues and market share. Similarly, OpenAI Inc. and OpenAI GP LLC exercised day-to-day control over the operations of OpenAI OpCo LLC and directly benefited from the infringing activities through increased valuation, licensing revenues, and other commercial advantages.

107. Microsoft—based on its partnership with, and significant investment in, OpenAI—had the right and ability to supervise and control OpenAI's infringing activity.

108. Essentially, instead of Microsoft entering the market and buying copyrighted works for its own training data, it received those same works from OpenAI, who had previously stolen them, and then provided them to Microsoft in exchange for data Microsoft had from its Bing search engine team's work on data acquisition.

109. Microsoft failed to exercise its supervision and control to prevent and/or stop OpenAI's infringement.

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

111. 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 OpenAI OpCo LLC's infringing activities.

112. OpenAI Inc. and OpenAI GP LLC failed to exercise its supervision and control to prevent and/or stop OpenAI OpCo LLC.

COUNT 3

UCL — Unfair Competition

Cal. Bus. & Prof. Code §§ 17200 et seq.

(Against All Defendants on Behalf of the Non-Book Infringement Class and the Unregistered Class)

113. Plaintiffs incorporate by reference the preceding factual allegations.

114. Defendants' conduct constitutes unlawful business practices under the UCL by violating the Copyright Act, the DMCA, and the CDAFA, as alleged herein. Defendants' conduct is also unfair in that it offends established public policy and is immoral, unethical, oppressive, and unscrupulous, causing substantial injury to Plaintiffs and the Class that is not outweighed by any countervailing benefits. Plaintiffs and Class members have suffered injury in fact and lost money or property as a result of Defendants' conduct, including but not limited to the deprivation of licensing revenue, diminution in the value of their copyrighted works, and loss of control over the use and dissemination of their intellectual property.

115. Defendants engaged in unfair business practices by, among other things, acquiring Plaintiffs' works through unlawful means, including torrenting vast amounts of pirated copyrighted works from shadow libraries; removing copyright identification information from Plaintiffs' works; and using the copies of those works to acquire additional training data from Microsoft.

116. Defendants also misrepresented their adherence to ethics and respect for rights with respect to their AI operations, including but not limited to their procurement and use of data.

117. The unfair business practices described herein violate California Business and Professions Code § 17200 *et seq*. (the "UCL") and are unfair, unlawful, and fraudulent.

118. Microsoft directly contributed to these unfair business practices by providing substantial financial resources, cloud infrastructure, and strategic support to OpenAI, enabling the development of ChatGPT using unlawfully obtained copyrighted text data and material. Microsoft has further engaged in unfair practices by integrating OpenAI's infringing models into its own commercial products and services, thereby deriving its own profits from the exploitation of Plaintiffs' stolen copyrighted text and material.

119. The unfair business practices described herein violate the UCL because they are unfair, immoral, unethical, oppressive, unscrupulous, or injurious to consumers. Defendants unfairly profit from and take credit for developing a commercial product based on unattributed reproductions of those stolen writings and ideas.

120. The unlawful business practices described herein violate the UCL because consumers are likely to be deceived by them. Defendants knowingly and secretively acquired, copied, and trained ChatGPT using unauthorized and infringing copies of Plaintiffs' copyrighted text. Defendants deceptively marketed their product in a manner that fails to attribute the success of their product to copyrighted material on which it is based.

COUNT 4

Violation of the California Comprehensive Computer Data Access and Fraud Act (CDAFA)

Cal. Penal Code § 502

(Against Defendant OpenAI on Behalf of the Unregistered Class)

121. Plaintiffs incorporate by reference the preceding factual allegations.

122. Defendant OpenAI, without permission, knowingly accessed and used Plaintiffs' copyrighted works by circumventing technological barriers and downloading digital copies from shadow libraries such as LibGen, in violation of Cal. Penal Code § 502(c)(1), (2), and (7). As a direct and proximate result, Plaintiffs suffered damage and loss, including but not limited to the costs incurred in investigating the unauthorized access and the diminution in value of their intellectual property.

123. OpenAI's unauthorized access and use of Plaintiffs' copyrighted works by torrenting

digital copies of those works from shadow libraries such as LibGen constitute violations of the California Comprehensive Computer Data Access and Fraud Act (CDAFA), Cal. Penal Code § 502.

124. OpenAI knowingly and without permission accessed and used data from Plaintiffs' copyrighted works to train its AI models, thereby causing harm to Plaintiffs.

125. OpenAI's knowing and unauthorized access to Plaintiffs' copyrighted works was a substantial factor in causing Plaintiffs' harm, including, but not limited to, the loss of control over their copyrighted material and the unauthorized use of their intellectual property.

126. As a direct and proximate result of OpenAI's actions, Plaintiffs have suffered damages, including, but not limited to, the loss of control over their copyrighted material and the unauthorized use of their intellectual property, as well as the amount spent to investigate or verify whether Plaintiffs' data was or was not altered, damaged, or deleted by OpenAI. Plaintiffs have engaged in and continue to engage in protracted efforts to determine how Defendants acquired their copyrighted data. On information and belief, and based on how a user typically obtains a torrent file, it is plausible that OpenAI redistributed that data through seeding and leeching, making it available to data pirates worldwide and thus furthering such piracy beyond their own downloading efforts.

127. Plaintiffs are entitled to compensatory damages, injunctive relief, and other equitable remedies as provided by Cal. Penal Code § 502(e).

COUNT 5

Violation of the Digital Millennium Copyright Act (DMCA)

U.S.C. § 1201

(Against all Defendants on Behalf of the Non-Book Infringement Class and the Unregistered Class)

128. Plaintiffs incorporate by reference the preceding factual allegations.

129. Defendants circumvented technological measures that effectively control access to Plaintiffs' copyrighted works, including but not limited to digital rights management systems and password protections in violation of 17 U.S.C. § 1201(a).

130. Defendants circumvented technological measures that control access to Plaintiffs' copyrighted works, including, but not limited to, by torrenting these and vast amounts of other

copyrighted material from shadow libraries such as LibGen and by ignoring robots.txt files (the filename used for implementing the Robots Exclusion Protocol, which is designed to indicate which websites crawlers are allowed to visit) and other access-related security measures and/or by using data obtained by similarly bypassing security measures.

131. Defendants' conduct in bypassing these technological measures was done without authorization and for the purpose of infringing Plaintiffs' copyrights.

132. Upon information and belief, based on how a user typically obtains a torrent file, it is plausible that Defendants also distributed these works, including, but not limited to, through torrenting and seeding.

133. As a result of these violations of the DMCA, Plaintiffs have suffered and will continue to suffer irreparable harm and are entitled to injunctive relief, statutory damages, and other remedies as provided by 17 U.S.C. § 1203.

COUNT 6

CMI-Stripping: Violation of the Digital Millennium Copyright Act (DMCA)

U.S.C. § 1201(b)(1)

(Against All Defendants on Behalf of the Non-Book Infringement Class and the Unregistered Class)

134. Plaintiffs incorporate by reference the preceding factual allegations.

135. OpenAI repeatedly and intentionally removed copyright management information ("CMI") from copyrighted works, including but not limited to the Selected Infringed Works, that OpenAI copied and used to train ChatGPT.

136. OpenAI source code specifically references an approach that "helps to eliminate copyright info from [the] state of [the document]" with respect to the LibGen dataset.[6]

137. Indeed, OpenAI removed CMI from Selected Infringed Works in part to enable and to facilitate infringement. Removal of CMI made it easier to use these Works as training data and because their use in training constitutes an infringement, the CMI removal "facilitated" that infringement.

[6] OPCO_AG_SRC_CODE00000941.

138. OpenAI also removed CMI from texts including, but not limited to, the Selected Infringed Works contained in OpenAI's training datasets to conceal OpenAI's infringement of copyrighted material, including but not limited to the Selected Infringed Works, from ChatGPT users and the public.

139. OpenAI sought to conceal its infringement to minimize risks that ChatGPT users and the public might learn or perceive that it had engaged in mass IP piracy, copyright infringement, and other unlawful activity in developing ChatGPT. OpenAI knew that ChatGPT could generate verbatim text from copyrighted material used to train ChatGPT. Open AI also knew that ChatGPT was prone to memorizing and generating outputs of CMI unless it was removed from the copyrighted works used to train ChatGPT.

140. Due to, among other things, the CMI that OpenAI removed from copyrighted works and OpenAI's knowledge that LibGen contained copyrighted articles and other textual works, OpenAI knew or had reasonable grounds to know that its removal of CMI from ChatGPT's training data would induce, enable, facilitate, or conceal its own copyright infringement or the copyright infringement of others. Among other things, OpenAI knew or had reasonable grounds to know its removal of CMI would reduce the chances that Plaintiffs and Class members would discover OpenAI had copied texts including but not limited to the Selected Infringed Works and/or used them to train ChatGPT.

141. Defendants also distributed these works, including by trading training data between OpenAI and Microsoft.

COUNT 7

Conversion

(Against Defendant OpenAI on Behalf of the Unregistered Class)

142. Plaintiffs incorporate by reference the preceding factual allegations.

143. OpenAI's unauthorized acquisition and use of Plaintiffs' copyrighted works by unlawfully scraping and/or crawling and/or illegally torrenting them from shadow libraries such as LibGen constitute conversion under California law.

144. OpenAI took Plaintiffs' copyrighted works without permission. In doing so, OpenAI wrongfully exercised dominion and control over Plaintiffs' property, depriving them of their rights to

use and control their works. OpenAI's unauthorized acquisition and use of Plaintiffs' copyrighted works deprived Plaintiffs of the exclusive right to control, license, and exploit their works, including the right to determine the terms and conditions of use. By unlawfully appropriating and using the works in a manner inconsistent with Plaintiffs' rights, OpenAI exercised wrongful dominion and control over specific, identifiable digital files embodying Plaintiffs' intellectual property, resulting in actual and substantial interference with Plaintiffs' property interests.

145. As a result of OpenAI's conversion, Plaintiffs suffered damages, including but not limited to the loss of control over their copyrighted works and the unauthorized use of their intellectual property.

146. Plaintiffs are entitled to compensatory damages, punitive damages, and other equitable relief as provided by California law.

COUNT 8

Unjust Enrichment / Quasi-Contract

(Against All Defendants on Behalf of the Unregistered Class)

147. Plaintiffs incorporate by reference the preceding factual allegations.

148. In the alternative to Plaintiffs' statutory and tort claims, Plaintiffs allege that Defendants have been unjustly enriched by their unauthorized acquisition, use, and exploitation of Plaintiffs' copyrighted works, from which Defendants derived substantial commercial benefits. It would be inequitable for Defendants to retain the profits and advantages obtained through such conduct without compensating Plaintiffs for the value of their intellectual property.

149. OpenAI has been unjustly enriched by its unauthorized access, copying, and use of Plaintiffs' copyrighted material to train its GenAI models, deriving significant commercial benefits and profits from this use.

150. Microsoft directly benefited from this unjust enrichment by integrating OpenAI's models into its own products and services, leveraging Plaintiffs' unlawfully obtained intellectual property to enhance its offerings and increase its profits.

151. Defendants' enrichment came at the expense of Plaintiffs, who have not been compensated for the acquisition and use of their copyrighted material.

152. It would be inequitable for Defendants to retain the benefits derived from their unauthorized acquisition and use of Plaintiffs' copyrighted material without providing compensation to Plaintiffs.

153. Plaintiffs are entitled to restitution and non-restitutionary disgorgement of all profits obtained by OpenAI and Microsoft as a result of their unjust enrichment, as well as other equitable relief as provided by law.

COUNT 9

Breach of Contract as a Third-Party Beneficiary

(Against All Defendants on Behalf of the Unregistered Class)

154. Plaintiffs incorporate by reference the preceding factual allegations.

155. Many of the websites from which Defendants obtained data for training GenAI models have terms and conditions that prohibit the unauthorized copying and use of their content. These terms and conditions are intended to protect the rights of content creators, including Plaintiffs, who publish their works on these platforms.

156. These terms and conditions are designed to protect the intellectual property rights of content creators and intellectual property owners from unauthorized copying.

157. Plaintiffs, as content creators and copyright holders whose works are hosted on these platforms, are intended third-party beneficiaries of these contractual provisions. Defendants breached these contracts by knowingly and willfully copying and using content in violation of the express terms, thereby depriving Plaintiffs of the protections and benefits intended by the contracts

158. Put another way, because the terms and conditions are designed to protect Plaintiffs' intellectual property, they are third-party beneficiaries of these contracts between the websites and their users.

159. Defendants, including Microsoft, breached these contracts by facilitating and benefiting from the unauthorized copying and use of content from these websites in violation of their terms and conditions. Microsoft contributed to these contract breaches by providing OpenAI with the necessary financial resources, infrastructure, and operational support to copy and use the protected content unlawfully.

160. As a direct and proximate result of Defendants' breaches, Plaintiffs suffered damages, including but not limited to the loss of control over their copyrighted works and the unauthorized use of their intellectual property.

161. Plaintiffs are entitled to compensatory damages and other equitable relief as provided by law.

COUNT 10

Violation of the Computer Fraud and Abuse Act (CFAA)

18 U.S.C. § 1030

(Against All Defendants on Behalf of the Unregistered Class)

162. Plaintiffs incorporate by reference the preceding factual allegations.

163. OpenAI's unauthorized access and use of data from websites that prohibit such activities in their terms and conditions constitute violations of the Computer Fraud and Abuse Act. OpenAI, with Microsoft's knowledge and facilitation, knowingly and with intent to defraud accessed protected computers without authorization, or exceeded authorized access, and obtained information from these computers. Microsoft contributed to these actions by providing infrastructure, financial support, and resources necessary for OpenAI to engage in these unlawful activities.

164. Defendants knowingly and with intent to defraud accessed protected computers hosting Plaintiffs' copyrighted works without authorization, or exceeded authorized access, and obtained information from such computers in violation of 18 U.S.C. § 1030(a)(2)(C). As a result, Plaintiffs suffered loss as defined by 18 U.S.C. § 1030(e)(11), including the costs of investigating the unauthorized access and the impairment of the integrity and value of their intellectual property.

165. As a result of Defendants' violations of the CFAA, Plaintiffs suffered damages, including but not limited to the loss of control over their copyrighted works and the unauthorized use of their intellectual property.

166. Plaintiffs are entitled to compensatory damages, injunctive relief, and other equitable remedies as provided by 18 U.S.C. § 1030(g).

COUNT 11

Larceny/Receipt of Stolen Property

Cal. Penal Code § 496(a), (c)

(Against All Defendants on Behalf of the Unregistered Class)

167. Plaintiffs incorporate by reference the preceding factual allegations.

168. California Penal Code § 496(a) creates an action against any person who (1) receives any property that has been stolen or obtained in any manner constituting theft, knowing the property to be stolen or obtained, or (2) conceals, sells, withholds, or aids in concealing or withholding any property from the owner, knowing the property to be so stolen or illegally obtained.

169. Defendants knowingly received, concealed, and withheld property—specifically, digital files embodying Plaintiffs' copyrighted works—that had been obtained in a manner constituting theft, including unauthorized copying and distribution from shadow libraries, in violation of Cal. Penal Code § 496(a). Defendants knew or had reason to know that the property was obtained without the consent of the rightful owners and in violation of law

170. Under Cal. Penal Code § 7, "the word 'person' includes a corporation as well as a natural person." Thus, Defendants are persons under Cal. Penal Code § 496(a).

171. As discussed above, Defendants unlawfully acquired copyrighted material from the internet, including by torrenting works from shadow libraries and violating websites' terms and conditions, and used the material to develop GenAI models and products in order to generate massive profits. At no point did Defendants have consent to take/scrape this information and use it in connection with their GenAI models and products. Defendants meet the grounds for liability under Cal. Penal Code § 496(a) because each of them:

a. Knew that the taken copyrighted material was stolen or obtained without permission, and with such knowledge;

b. Concealed, withheld, or aided in concealing or withholding said data from their rightful owners by unlawfully manipulating (for example, removing CMI) and using the data to train their GenAI models.

172. Pursuant to Cal. Penal Code § 496(c), Plaintiffs, on behalf of themselves and the Class, seek actual damages, treble damages, costs of suit, and reasonable attorneys' fees.

COUNT 12

Sherman Act – Conspiracy to Restrain Trade

15 U.S.C. §§ 1 & 3

(Against All Defendants on Behalf of the Non-Book Infringement Class and the Unregistered Class)

173. Plaintiffs incorporate by reference the preceding factual allegations.

174. Defendants OpenAI and Microsoft entered into a partnership, including substantial financial investments and operational support, to develop and commercialize large language models.

175. All textual work has value for use as training data for LLMs. This is demonstrated by the numerous licensing deals announced between generative AI companies such as OpenAI and Microsoft with licensors of textual works. This includes licenses between AI companies and licensors of registered works such as the highly publicized licensing deal between Microsoft and HarperCollins, and licenses between AI companies and licensors of unregistered textual works such as between OpenAI and Reddit.

176. The relevant product market is the market for registered copyright works and unregistered textual works for LLM training data, broadly defined as data that can be used in consideration for training, evaluation, validation, and actual training. The relevant geographic market is nationwide.

177. Defendants OpenAI and Microsoft are horizontal competitors for training data, and both compete in the markets for registered copyrighted works and unregistered text for training data.

178. The relevant product market is the market for copyrighted and unregistered textual works used as training data for large language models in the United States. Defendants OpenAI and Microsoft, as horizontal competitors in this market, entered into an agreement to share and exchange training data, including unlawfully acquired copyrighted works, with the purpose and effect of suppressing the price and availability of such data, foreclosing competition, and restraining trade. As a direct and proximate result, Plaintiffs and the Class suffered antitrust injury, including reduced

compensation for their works, diminished market opportunities, and suppression of innovation and output in the market for AI training data.

179. As a key part of that partnership, OpenAI and Microsoft reached an agreement and common understanding. While the arrangements were in part formalized in written documentation, the basic purpose and effect of the agreement was that OpenAI and Microsoft would cooperate to prevent the development of a fee and open market for training data.

180. In particular, OpenAI agreed to share training data for LLMs with Microsoft. On its part, instead of entering the market and buying copyrighted and unregistered works for its own training data, Microsoft received those same works from OpenAI, who had previously stolen them, and then provided them to Microsoft in exchange for data Microsoft had from its Bing search engine team's work on data acquisition. OpenAI, in turn, would receive training data derived from textual works scraped from the public internet for pennies on the dollar.

181. As a result of that agreement and understanding, OpenAI and Microsoft artificially and unreasonably restrained the market for training data for LLMs. Instead of Microsoft entering the market and purchasing copyrighted works for its own training data, it received those works from OpenAI in exchange for its own Bing data (on information and belief, comprised of both registered copyright works and unregistered works). This arrangement represents anticompetitive conduct between horizontal competitors in the market for training data for LLMs and as such limited competition in the AI training space, since Microsoft did not compete with OpenAI to acquire high-quality training data for LLMs. The agreement to "tone down the effort on training on the large models" and to be "super-careful not competing with [OpenAI]" further solidified the restraint of trade. As a direct and foreseeable result, OpenAI and Microsoft paid less for training data than they would have but for their agreement and common course of conduct.

182. As a direct and proximate result of Defendants' conspiracy to restrain trade, Plaintiffs and the Class have suffered antitrust injury, including but not limited to the loss of control over their registered copyrighted works and unregistered works, the unauthorized use of their intellectual property, and diminished market value of their works. The conspiracy affected Plaintiffs' ability to compete in the market, resulting in financial losses and other specific damages. By artificially

restraining the price of training data and limiting competition, Defendants' actions have caused significant harm to Plaintiffs and the market as a whole.

183. Because Defendants are horizontal competitors in the relevant markets for registered copyrighted works as training data, and unregistered textual works for training data, Defendants' agreement to restrain trade constitutes a *per se* violation of the Sherman Act.

184. While the conspiracy constitutes a *per se* violation of the Sherman Act, Defendants also exploited their collective market power in the relevant market, which is the market for training data for LLMs in the United States.

185. Through their conspiracy, Defendants exercised and maintained market power, and did in fact suppress the market value for copyrighted works as training data for LLMs.

186. The purpose and effect of this restraint of trade was to restrain competition. Prices decreased, output decreased and innovation was constrained.

187. The conduct was not part of a legitimate joint venture and was not ancillary to another legitimate agreement.

188. The conspiracy and the conduct of Defendants and their agents and co-conspirators in furtherance thereof did not have procompetitive effects and were not intended to have procompetitive effects.

189. In the alternative, any procompetitive effects that may have resulted from the conspiracy are substantially outweighed by the anticompetitive harm alleged herein, including, but not limiting to eliminating Class members' ability to control their works and suppressing the price of copyrighted works as training data for LLMs.

190. Defendants are also liable under a "quick look" analysis where one with even a rudimentary understanding of economics could conclude that the arrangements and agreements alleged would have an anticompetitive effect on Class members and the relevant market.

X. DEMAND FOR JUDGMENT

Wherefore, Plaintiffs request that the Court enter judgment on their behalf and on behalf of the Class defined herein, by ordering and decreeing:

a. This Action may proceed as a class action, with Plaintiffs serving as Class

Representatives, and with Plaintiffs' counsel as Class Counsel;

b. A declaration that Defendants have infringed Plaintiff Catherine Denial's and members of the Non-Book Infringement Class's exclusive copyrights, including but not limited to those in the Selected Infringed Works, under the Copyright Act;

c. A declaration that such infringement is willful;

d. Judgment in favor of Plaintiffs and the Class and against Defendants;

e. An award of statutory and other damages under 17 U.S.C. § 504 for Defendants' willful infringement of Plaintiff Catherine Denial's and members of the Non-Book Infringement Class's exclusive copyrights;

f. Defendants have engaged in a trust, contract, combination, or conspiracy in violation of Sections 1 and 3 of the Sherman Act, and that Plaintiffs and Class members have been damaged and injured in their business and property as a result of this violation;

g. The alleged combinations and conspiracy are *per se* violations of the Sheman Act;

h. Reasonable attorneys' fees and costs as available under 17 U.S.C. § 505, Cal. Penal Code § 502 or other applicable statute;

i. Pre- and post-judgment interest on the damages awarded to Plaintiffs and the Class, and that such interest be awarded at the highest legal rate from and after the date this class action complaint is first served on Defendants;

j. Defendants are to be jointly and severally responsible financially for the costs and expenses of a Court-approved notice program through post and media designed to give immediate notification to the Class;

k. Nominal, treble, and punitive damages, as warranted;

l. Permanent injunctive relief, including but not limited to the return, destruction, and cessation of the use of any data illegally or unlawfully acquired, or of the products dependent upon the use thereof;

m. Restitution and non-restitutionary disgorgement of all profits obtained as a result of Defendants' unjust enrichment, as well as other equitable relief as provided by law;

n. Further relief for Plaintiffs and the Class as may be just and proper.

JURY TRIAL DEMANDED

Under Federal Rule of Civil Procedure 38(b), Plaintiffs demand a trial by jury of all the claims asserted in this Complaint so triable.

By: *Joseph R. Saveri*
JOSEPH SAVERI LAW FIRM, LLP
Joseph R. Saveri (SBN 130064)
Cadio Zirpoli (SBN 179108)
Christopher K.L. Young (SBN 318371)
Holden Benon (SBN 325847)
Aaron Cera (SBN 351163)
Alexander Y. Zeng (SBN 360220)
jsaveri@saverilawfirm.com
czirpoli@saverilawfirm.com
hbenon@saverilawfirm.com
cyoung@saverilawfirm.com
acera@saverilawfirm.com
azeng@saverilawfirm.com601 California Street, Suite 1505
San Francisco, CA 94108
Telephone: (415) 500-6800
Facsimile: (415) 395-9940

By: *Bryan L. Clobes*
CAFFERTY CLOBES MERIWETHER & SPRENGEL LLP
Bryan L. Clobes (*pro hac vice*)
Mohammed A. Rathur (*pro hac vice*)
bclobes@caffertyclobes.com
mrathur@caffertyclobes.com
135 South LaSalle Street, Suite 3210
Chicago, IL 60603
Tel: (312) 782-4880

By: *Joshua Michelangelo Stein*
BOIES SCHILLER FLEXNER LLP
Maxwell V. Pritt (SBN 253155)
Joshua Michelangelo Stein (SBN 298856)
Reed D. Forbush (SBN 347964)
mpritt@bsfllp.com
jstein@bsfllp.com44 Montgomery Street, 41st Floor
San Francisco, CA 94104
Telephone: (415) 293-6800

David Boies (*pro hac vice*)
dboies@bsfllp.com 333 Main Street
Armonk, NY 10504
Telephone: (914) 749-8200

Jesse Panuccio (*pro hac vice*)
jpanuccio@bsfllp.com
1401 New York Ave, NW
Washington, DC 20005
Telephone: (202) 237-2727

Evan Matthew Ezray *(pro hac vice)*
eezray@bsfllp.com
401 E. Las Olas Blvd., Suite 1200
Ft. Lauderdale, FL 33301
Telephone: (954) 356-0011

Counsel for Individual and Representative Plaintiff and the Proposed Class