UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

RAW STORY MEDIA, INC., ALTERNET MEDIA, INC.,

Plaintiffs,

v.

OPENAI, INC., OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, OPENAI HOLDINGS, LLC,

Defendants.

Civil Action No. _____

COMPLAINT

JURY TRIAL DEMANDED

1. Plaintiffs Raw Story Media, Inc. and AlterNet Media, Inc., through their attorneys Loevy & Loevy, for their Complaint against the OpenAI Defendants, allege the following:

2. The Copyright Clause of the U.S. Constitution empowers Congress to protect works of human creativity. The resulting legal protections encourage people to devote effort and resources to all manner of creative enterprises by providing confidence that creators' works will be shielded from unauthorized encroachment.

3. In recognition that emerging technologies could be used to evade statutory protections, Congress passed the Digital Millennium Copyright Act in 1998. The DMCA prohibits the removal of author, title, copyright, and terms of use information from protected works where there is reason to know that it would induce, enable, facilitate, or conceal a copyright infringement. Unlike copyright infringement claims, which require copyright owners to incur significant and often prohibitive registration costs as a prerequisite to enforcing their copyrights, a DMCA claim does not require registration.

4. Generative artificial intelligence (AI) systems and large language models (LLMs) are trained using works created by humans. AI systems and LLMs ingest massive amounts of human creativity and use it to mimic how humans write and speak. These training sets have included hundreds of thousands, if not millions, of works of journalism.

5. Defendants are the companies primarily responsible for the creation and development of the highly lucrative ChatGPT AI products. According to the award-winning website Copyleaks, nearly 60% of the responses provided by Defendants' GPT-3.5 product in a study conducted by Copyleaks contained some form of plagiarized content, and over 45% contained text that was identical to pre-existing content.

6. When they populated their training sets with works of journalism, Defendants had a choice: they could train ChatGPT using works of journalism with the copyright management information protected by the DMCA intact, or they could strip it away. Defendants chose the latter, and in the process, trained ChatGPT not to acknowledge or respect copyright, not to notify ChatGPT users when the responses they received were protected by journalists' copyrights, and not to provide attribution when using the works of human journalists.

7. Plaintiffs Raw Story and AlterNet are news organizations, and bring this lawsuit seeking actual damages and Defendants' profits, or statutory damages of no less than $2500 per violation.

PARTIES

8. For over two decades, Raw Story has published award-winning investigative journalism, breaking news, and bold opinion columns. Raw Story publishes to more than ten million readers each month and has more than 1,000,000 daily readers. It is the largest independent

progressive political news website in America and was named the best news/political blog in America by *Editor & Publisher* in 2022 and 2023.

9. Among other important work, Raw Story has received *Editor & Publisher* (EPPY), Society of Professional Journalists, Fair Media Council, and ION awards for its reporting on white nationalism, the January 6 riots, South Dakota governor Kristi Noem's use of a state airplane for non-official purposes, and inappropriate Congressional stock trading. Raw Story reporters produce timely, illuminating work at great risk and cost to enrich understanding of critical issues and undermine threats to civil society.

10. As one example of the risks taken by Raw Story reporters in bringing the news to the public—risks never faced by AI bots—members of a neo-Nazi group showed up at the home of a Raw Story reporter who covers extremism and white supremacy in America. *See* Washington Post, *A reporter investigated neo-Nazis. Then they came to his house in masks.* (Feb. 20, 2024), https://www.washingtonpost.com/style/media/2024/02/20/raw-story-neo-nazi-journalist-house/.

11. Raw Story is a Massachusetts corporation with its headquarters in Miami Beach, Florida.

12. AlterNet is a three-time Webby award-winning publisher with a focus on civil rights, social justice, culture, health, and the environment. For 25 years, AlterNet's reporters have chased political news, and its opinion writers have probed the intersection of politics, science, and religion.

13. AlterNet is a Delaware corporation with its headquarters in Miami Beach, Florida.

14. Together, Plaintiffs have published more than 400,000 breaking news features, investigative news articles and opinion columns as a result of their considerable investments of time and resources.

15. Defendants are the inter-related organizations primarily responsible for the creation, training, marketing, and sale of ChatGPT AI products.

16. OpenAI Inc. is a Delaware nonprofit corporation with a principal place of business in San Francisco, CA.

17. OpenAI OpCo LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. OpenAI OpCo LLC is the sole member of OpenAI, LLC. Previously, OpenAI OpCo was known as OpenAI LP.

18. OpenAI GP, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. It is the general partner of OpenAI OpCo and controls OpenAI OpCo.

19. OpenAI, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. It owns some of the services or products operated by OpenAI.

20. OpenAI Global LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA.

21. OAI Corporation, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. Its sole member is OpenAI Holdings, LLC.

22. OpenAI Holdings, LLC is a Delaware limited liability company with a principal place of business in San Francisco, CA. Its sole members are OpenAI, Inc. and Aestas Corporation.

JURISDICTION AND VENUE

23. 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., as amended by the Digital Millennium Copyright Act.

24. Jurisdiction over Defendants is proper because they have purposefully availed themselves of New York to conduct their business. OpenAI maintains offices and employs staff in New York who, on information and belief, were engaged in training and/or marketing OpenAI's GenAI systems and LLMs, and thus in the removal of Plaintiffs' copyright management information as discussed in this Complaint and/or the sale of products to New York residents resulting from that removal. Defendants consented to personal jurisdiction in this Court in at least *Authors Guild v. OpenAI Inc.*, 23-cv-08292.

25. Venue is proper under 28 U.S.C. § 1400(a) because Defendants or their agents reside or may be found in this District.

26. Venue is also proper under 28 U.S.C. § 1391(b)(2) because a substantial part of the acts or omissions giving rise to Plaintiffs' claims occurred in this District. Specifically, OpenAI employs staff in New York who, on information and belief, were engaged in the activities alleged in this Complaint.

27. The OpenAI Defendants consented to venue in this Court in at least *Authors Guild v. OpenAI Inc.*, 23-cv-08292.

DEFENDANTS' DMCA VIOLATIONS

28. Defendants have kept secret the specific content used to train all versions of ChatGPT beginning with GPT-4. Plaintiffs' allegations are therefore based upon an extensive review of publicly available information regarding earlier versions of ChatGPT and consultations with a data scientist employed by Plaintiffs' counsel to analyze that information and provide insights into the manner in which AI is developed and functions.

29. Earlier versions of ChatGPT[1] (prior to GPT-4) were trained using at least the following training sets: WebText, WebText2, and Common Crawl. These training sets range from collections of links posted on the website Reddit to a scrape of most of the internet.

30. WebText and WebText2 were created by the OpenAI Defendants. Common Crawl originated elsewhere, but was adapted and utilized by Defendants for inclusion in ChatGPT training sets. Upon information and belief, the OpenAI Defendants have created their own Common Crawl datasets, as opposed to copying a dataset already created by someone else.

31. Plaintiffs' copyrighted works of journalism are published on the internet, and are conveyed to the public with author, title, and copyright information.

32. Plaintiffs' copyright-protected works are the result of significant investments by Plaintiffs in the human and other resources necessary to report on the news.

33. ChatGPT offers a product to its customers that provides responses to questions or other prompts. ChatGPT's ability to provide these responses is the key value proposition of its product, one which it is able to sell to its customers for enormous sums of money, soon likely to be in the billions of dollars.

34. At least some of the time, ChatGPT provides or has provided responses to users that regurgitate verbatim or nearly verbatim copyright-protected works of journalism without providing any author, title, or copyright information contained in those works.

35. At least some of the time, ChatGPT provides or has provided responses to users that mimic significant amounts of material from copyright-protected works of journalism without providing any author, title, or copyright information contained in those works. For example, if a

[1] Plaintiffs collectively refer to all versions of ChatGPT as "ChatGPT" unless a specific version is specified.

user asks ChatGPT about a current event or the results of a work of investigative journalism, ChatGPT will provide responses that mimic copyright-protected works of journalism that covered those events, not responses that are based on any journalism efforts by Defendants.

36. ChatGPT does not have any independent knowledge of the information provided in its responses. Rather, to service Defendants' paying customers, ChatGPT instead repackages, among other material, the copyrighted journalism work product developed by Plaintiffs and others at their expense.

37. Various sources have recreated approximations of the Common Crawl and WebText training sets based on publicly available information discussing the methodologies used to create them. Those sources have made these recreated data sets, or instructions on how to derive them, available on the internet. Thousands of Plaintiffs' works are contained in the recreated versions of these data sets without the author, title, and copyright information found in Plaintiffs' original publications.

38. If ChatGPT was trained on works of journalism that included the original author, title, and copyright information, ChatGPT would have learned to communicate that information when providing responses to users unless Defendants trained it otherwise.

39. When ChatGPT provides responses to users, it generally does not provide the author, title, and copyright information applicable to the works on which its responses are based. Upon information and belief, in the instances in which author or title information is included in a response, it is because other material used in a training set references the author or title in the text of such material (e.g., a Wikipedia article discussing the underlying works of journalism).

40. When providing responses, ChatGPT gives the impression that it is an all-knowing, "intelligent" source of the information being provided, when in reality, the responses are frequently based on copyrighted works of journalism that ChatGPT simply mimics.

41. Based on the publicly available information described above, thousands of Plaintiffs' copyrighted works were included in Defendants' training sets without the author, title, and copyright information that Plaintiffs conveyed in publishing them.

42. Based on the publicly available information described above, the OpenAI Defendants intentionally removed author, title, and copyright information from Plaintiffs' copyrighted works in creating ChatGPT training sets.

43. Defendants had reasonable grounds to know that the removal of author, title, and copyright information from copyright-protected works and their use in training ChatGPT would result in ChatGPT providing responses to ChatGPT users that incorporated or regurgitated material verbatim from copyrighted works in creating responses to users, without revealing that those works were subject to Plaintiffs' copyrights. This is at least because Defendants were aware that ChatGPT responses are the product of its training sets and that ChatGPT would not know any author, title, and copyright information that was not included in training sets.

44. Defendants had reason to know that users of ChatGPT would further distribute the results of ChatGPT responses. This is at least because Defendants promote ChatGPT as a tool that can be used by a user to generate content for a further audience.

45. Defendants had reason to know that users of ChatGPT would be less likely to distribute ChatGPT responses if they were made aware of the author, title, and copyright information applicable to the material used to generate those responses. This is at least because

Defendants were aware that at least some likely users of ChatGPT respect the copyrights of others or fear liability for copyright infringement.

46. Defendants had reason to know that ChatGPT would be less popular and would generate less revenue if users believed that ChatGPT responses violated third-party copyrights or if users were otherwise concerned about further distributing ChatGPT responses. This is at least because Defendants were aware that they derive revenue from user subscriptions, that at least some likely users of ChatGPT respect the copyrights of others or fear liability for copyright infringement, and that such users would not pay to use a product that might result in copyright liability or did not respect the copyrights of others.

Count I – Violation of 17 U.S.C. § 1202(b)(1) by OpenAI Defendants

47. The above paragraphs are incorporated by reference into this Count.

48. Plaintiffs are the owners of copyrighted works of journalism that contain author, title, and copyright information.

49. Upon information and belief, the OpenAI Defendants created copies of Plaintiffs' works of journalism with author information removed and included them in training sets used to train ChatGPT.

50. Upon information and belief, the OpenAI Defendants created copies of Plaintiffs' works of journalism with title information removed and included them in training sets used to train ChatGPT.

51. Upon information and belief, the OpenAI Defendants created copies of Plaintiffs' works of journalism with copyright information removed and included them in training sets used to train ChatGPT.

52. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright information would induce ChatGPT to provide responses to users that incorporated material from Plaintiffs' copyright-protected works or regurgitated copyright-protected works verbatim or nearly verbatim.

53. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright information would induce ChatGPT users to distribute or publish ChatGPT responses that utilized Plaintiffs' copyright-protected works of journalism that such users would not have distributed or published if they were aware of the author, title, and copyright information.

54. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright information would enable copyright infringement by ChatGPT and ChatGPT users.

55. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright information would facilitate copyright infringement by ChatGPT and ChatGPT users.

56. The OpenAI Defendants had reason to know that inclusion in their training sets of Plaintiffs' works of journalism without author, title, and copyright information would conceal copyright infringement by Defendants, ChatGPT, and ChatGPT users.

57. The OpenAI Defendants have acknowledged that use of copyright-protected works to train ChatGPT requires a license to that content and, in some instances, have entered licensing agreements with large copyright owners such as Associated Press and Axel Springer. They are also in licensing talks with other copyright owners in the news industry, but have offered no compensation to Plaintiffs.

58. The OpenAI Defendants created tools in late 2023 to allow copyright owners to block their work from being incorporated into training sets. This further corroborates that the OpenAI Defendants had reason to know that use of copyrighted material in their training sets is copyright infringement, which is enabled, facilitated, and concealed by the OpenAI Defendants' removal of author, title, copyright, and terms of use information from their training sets.

PRAYER FOR RELIEF

Plaintiffs seek the following relief:

(i) Either statutory damages or the total of Plaintiffs' damages and Defendants' profits, to be elected by Plaintiffs;

(ii) An injunction requiring Defendants to remove all copies of Plaintiffs' copyrighted works from which author, title, copyright, and terms of use information was removed from their training sets and any other repositories;

(iii) Attorney fees and costs.

JURY DEMAND

Plaintiff demands a jury trial.

RESPECTFULLY SUBMITTED,

/s/ Stephen Stich Match

Jonathan Loevy*
Michael Kanovitz*
Lauren Carbajal*
Stephen Stich Match (No. 5567854)
Matthew Topic*

LOEVY & LOEVY
311 North Aberdeen, 3rd Floor
Chicago, IL 60607
312-243-5900
jon@loevy.com
mike@loevy.com
carbajal@loevy.com
match@loevy.com
matt@loevy.com

*pro hac vice forthcoming

February 28, 2024