UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

CALIFORNIA NEWSPAPERS PARTNERSHIP; PRAIRIE MOUNTAIN PUBLISHING COMPANY LLP; MNG-BH ACQUISITION LLC; HARTFORD COURANT COMPANY, LLC; THE DAILY PRESS LLC; THE MORNING CALL, LLC; VIRGINIAN-PILOT MEDIA COMPANIES, LLC; LOS ANGELES DAILY NEWS PUBLISHING COMPANY; AND THE SAN DIEGO UNION-TRIBUNE, LLC Plaintiffs, v. MICROSOFT CORPORATION; OPENAI, INC.; OPENAI LP; OPENAI GP, LLC; OPENAI, LLC; OPENAI OPCO, LLC; OPENAI GLOBAL, LLC; OAI CORPORATION, LLC; OPENAI HOLDINGS, LLC; OPENAI FOUNDATION; and OPENAI GROUP PBC. Defendants.	Civil Action No. _____ **COMPLAINT** **JURY TRIAL DEMANDED**

Plaintiffs California Newspapers Partnership, Prairie Mountain Publishing Company LLP; MNG-BH Acquisition LLC; Hartford Courant Company, LLC; The Daily Press LLC; The Morning Call, LLC; Virginian-Pilot Media Companies, LLC; Los Angeles Daily News Publishing Company; and The San Diego Union-Tribune, LLC (collectively the "Publishers"), by their attorneys Rothwell, Figg, Ernst & Manbeck, P.C., for their complaint against Defendants Microsoft Corporation ("Microsoft") and OpenAI, Inc.; OpenAI LP; OpenAI GP, LLC; OpenAI, LLC; OpenAI OpCo, LLC; OpenAI Global, LLC; OAI Corporation, LLC; OpenAI Holdings, LLC, OpenAI Foundation and OpenAI Group PBC (collectively "OpenAI" and, with Microsoft, "Defendants"), allege as follows:

NATURE OF THE ACTION

1. In this lawsuit, the publishers of nine regional newspapers join the long list of publishers and authors who have filed lawsuits against OpenAI, Microsoft, and other AI companies. Most of these lawsuits have been consolidated in this Court, and have survived motions to dismiss largely intact.

2. In the course of these lawsuits, a number of questions have been resolved. For example, there is no longer any question that AI models feed on copyrighted content. The AI companies' large language models ("LLMs"), including Defendants', were trained on copyrighted content scraped from the internet, regardless of paywalls or other restrictions, and heedless of the rights of publishers and owners. As other publishers have noted in their lawsuits against these same Defendants, the founder of OpenAI, Sam Altman, readily acknowledged in testimony before the British House of Lords that his companies rely on copyrighted material:

> Because copyright today covers virtually every sort of human expression—including blog posts, photographs, forum posts, scraps of software code, and government documents—it would be impossible to train today's leading AI models without using copyrighted materials. **Limiting training data to public domain books and drawings created more than a century ago might yield an interesting experiment, but would not provide AI systems that meet the needs of today's citizens.**[1]

3. Also settled is the fact that the Defendants decided to use copyrighted works without regard for the rights of the authors and publishers who created the works Defendants stole, and whose livelihoods depend on respect for their efforts. And not only did OpenAI and Microsoft steal the copyrighted works they used to train their models, they knowingly and intentionally did

[1] OpenAI, *House of Lords Communications and Digital Select Committee inquiry: Large language models* (Dec. 5, 2023), https://committees.parliament.uk/writtenevidence/126981/pdf. (emphasis added).

so using systems that stripped from the publishers' and authors' works any indication that the works were covered by valid copyrights.

4. There is no question that the Defendants' models have "memorized" the pilfered copies of the authors' and publishers' copyrighted works. And in order to remain current, the Defendants cannot rely just on the content they stole in the past – they have to update their models regularly with new material so they can provide their users with the latest information.

5. Defendants' violations of the law were willful. As Mr. Altman's testimony confirms, Defendants knew that their models were trained on copyrighted materials. Defendants also knew, or were willfully blind to the fact, that their models have "memorized" the copyrighted materials, and will include verbatim or near-verbatim versions of the copyrighted materials when prompted by users. And these violations are ongoing, again knowingly, intentionally, and without regard to the rights of those who research, investigate, write, and publish the information that Defendants' AI systems filch on a regular basis.

6. The U.S. Constitution, in Article I, § 8, grants to Congress the authority "To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries." In its exercise of this power, Congress has recognized the importance of erecting barriers to the theft of copyrighted materials, and has authorized stiff penalties for violation of the laws that protect copyright owners.

7. In this lawsuit, the Publishers seek damages in excess of $10 billion. Upon information and belief, Defendants' models have copied hundreds of thousands of articles and other materials from Publishers' newspapers (the articles and other materials are, collectively, the "Publishers' Works"). By statute, the Publishers are entitled to "not more than $150,000" for each willful violation of the Publishers' copyrights, and up to $25,000 for each work from which

Defendants removed the Publishers' copyright management information. On balance, a ten-figure damages award would be more than just; not only would it be consistent with Congress' intention that infringers compensate copyright owners for the theft of copyrighted material, but as explained below, such an award it would be a pittance compared to the enormous revenues and other gains Defendants and their investors have already realized based largely on their disregard of Publishers' rights and the rights of other copyright owners.

8. The following sections provide a detailed recitation of the Publishers' grounds for seeking the maximum award that Congress has authorized. These and other facts to be developed during the lawsuit will demonstrate that Defendants' repeated and ongoing violations of plaintiff's rights, even in the wake of three years of litigation brought by other publishers, were committed knowingly, willfully, and without regard to the legitimate rights of those from whom they stole.

I. JURISDICTION AND VENUE

9. 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. This Court has jurisdiction over related state claims under 28 U.S.C. § 1367.

10. Jurisdiction over Microsoft and OpenAI is proper because they have purposely availed themselves of the privilege of conducting business in New York. A substantial portion of Microsoft and OpenAI's widespread infringement and other unlawful conduct alleged herein occurred in New York, including the distribution and sales of Microsoft and OpenAI's Generative Pre-training Transformer ("GPT")-based products like ChatGPT, ChatGPT Enterprise, Copilot, Azure OpenAI Service, Microsoft 365 Copilot, Copilot Search, and related application programming interface (API) tools within New York to New York residents. Furthermore, both Microsoft and the OpenAI Defendants maintain offices and employ personnel in New York who,

upon information and belief, were involved in the creation, maintenance, or monetization of Microsoft and OpenAI's widespread infringement and other unlawful conduct alleged herein.

11. Venue is proper under 28 U.S.C. § 1400(a) because Defendants or their agents reside or may be found in this District, through the infringing and unlawful activities—as well as Defendants' sales and monetization of such activity—that occurred in this District. Venue is also proper under 28 U.S.C. § 1391(b)(2) because a substantial part of the events giving rise to the Publishers' claims occurred in this District, including the marketing, sales, and licensing of Defendants' GenAI products built on the infringement of the Publishers' intellectual property within this District. Upon information and belief, OpenAI has sold subscriptions for ChatGPT Plus to New York residents, Microsoft has sold subscriptions for Copilot Pro to New York residents, and both Microsoft and OpenAI enjoy a substantial base of monthly active users of Copilot and ChatGPT in New York. OpenAI has licensed its GPT models to New York residents and companies headquartered in New York. For example, OpenAI has entered into deals to license its GPT models to the Associated Press ("AP") and Morgan Stanley, both companies headquartered in New York.

II. THE PARTIES

12. Plaintiff California Newspapers Partnership ("CNP") is a Delaware general partnership with its principal place of business at 4 N. 2nd Street. San Jose, California 95113-1308. CNP publishes digital and print products, including, among other newspapers, *The San Bernardino Sun*, which is available on its mobile application, on its website (www.sbsun.com), and as a printed newspaper. CNP owns over 4,500 registered copyrights for its newspaper issues of *The San Bernardino Sun*, including those set forth in Exhibit A ("San Bernardino Sun Works").

13. Plaintiff Prairie Mountain Publishing Company LLP ("Prairie Mountain") is Delaware partnership with its principal place of business at 2500 55th St., Boulder, Colorado.

Prairie Mountain is the publisher of the *Daily Camera*, a local newspaper covering the city of Boulder, Colorado and its environs, which is available on its mobile application, on its website (www.dailycamera.com), and as a printed newspaper. Prairie Mountain owns over 2,750 registered copyrights for its newspaper issues of the *Daily Camera*, including those set forth in Exhibit B ("Daily Camera Works").

14. Plaintiff MNG-BH Acquisition LLC ("MNG-BH") is a Delaware limited liability company with its principal place of business at 4 N. 2nd Street. San Jose, California 95113-1308. MNG-BH is the publisher of, among other newspapers, the *Boston Herald*, which is available on its mobile application, on its website (www.bostonherald.com), and as a printed newspaper. MNG-BH owns over 7,000 registered copyrights for its newspaper issues of the *Boston Herald*, including those set forth in Exhibit C ("Boston Herald Works").

15. Plaintiff Hartford Courant Company LLC ("Courant") is a Delaware limited liability company with its principal place of business at 265 Broad Street, Hartford, Connecticut. Courant is the publisher of the *Hartford Courant*, which is available on its mobile application, on its website (www.courant.com), and as a printed newspaper. Courant owns over 10,000 registered copyrights for its newspaper issues of the *Hartford Courant*, including those set forth in Exhibit D ("Hartford Courant Works").

16. Plaintiff The Daily Press LLC ("Daily Press") is a Delaware limited liability company with its principal place of business at 1000 Albian Avenue, Schaumburg, Illinois. The Daily Press is the publisher of the *Daily Press*, which serves the Tidewater region of Virginia and is available on its mobile application, on its website (www.dailypress com), and as a printed newspaper. The Daily Press owns over 3,000 registered copyrights for its newspaper issues of the *Daily Press*, including those set forth in Exhibit E ("Daily Press Works").

17. Plaintiff The Morning Call, LLC ("Morning Call") is a Delaware limited liability company with its principal place of business at 101 North 6th Street, Allentown, Pennsylvania 18101. Morning Call publishes *The Morning Call* newspaper, which covers the Lehigh Valley region in Pennsylvania and is available on its mobile application, on its website (www.mcall.com), and as a printed newspaper. The Morning Call owns over 10,000 registered copyrights for its newspaper issues of *The Morning Call*, including those set forth in Exhibit F ("Morning Call Works").

18. Plaintiff Virginian-Pilot Media Companies ("Virginian-Pilot") is a Virginia limited liability company with its principal place of business at 150 Brambleton Avenue, Norfolk, Virginia. Virginian-Pilot publishes *The Virginian-Pilot* newspaper, which serves the region around Norfolk, Virginia, southeastern Virginia, and northeastern North Carolina and is available on its mobile application, on its website (www.pilotonline.com), and as a printed newspaper. The Virginian-Pilot owns over 11,000 registered copyrights for its newspaper issues of *The Virginian-Pilot*, including those set forth in Exhibit G ("Virginian-Pilot Works").

19. Plaintiff Los Angeles Daily News Publishing Company ("L.A. Daily News") is a Delaware corporation with its principal place of business at 500 West Temple Street, Los Angeles, California. The L.A. Daily News publishes the *Los Angeles Daily News*, which is the second-largest circulating paid daily newspaper in Los Angeles and is available on its mobile application, on its website (www.dailynews.com), and as a printed newspaper. The L.A. Daily News owns over 3,500 registered copyrights for its newspaper issues of the *Los Angeles Daily News*, including those set forth in Exhibit H ("L.A. Daily News Works").

20. Plaintiff The San Diego Union-Tribune, LLC ("Union-Tribune") is a Delaware limited liability company with its principal place of business at 600 B St., San Diego, California.

The Union Tribune publishes a San Diego newspaper of the same name, the *San Diego Union-Tribune*, which is available on its mobile application, on its website (www.sandiegouniontribune.com), and as a printed newspaper. The Union-Tribune owns over 8,000 registered copyrights for its newspaper issues of the *San Diego Union-Tribune*, including those set forth in Exhibit I ("Union-Tribune Works").

21. Microsoft Corporation is a Washington corporation with a principal place of business and headquarters in Redmond, Washington. Microsoft has invested at least $13 billion in OpenAI Global, LLC in exchange for which Microsoft will receive 75% of that company's profits until its investment is repaid, after which Microsoft will own a 49% stake in that company.

22. Microsoft has described its relationship with the OpenAI Defendants as a "partnership." This partnership has included contributing and operating the cloud computing services used to copy the San Bernardino Sun Works, the Daily Camera Works, the Boston Herald, the Hartford Courant Works, the Daily Press Works, the Morning Call Works, the Virginian-Pilot Works, the L.A. Daily News Works, and the Union-Tribune Works (collectively the "Publishers' Works") and train the OpenAI Defendants' GenAI models. It has also included, on information and belief, substantial technical collaboration on the creation of those models. Microsoft possesses copies of, or obtains preferential access to, the OpenAI Defendants' latest GenAI models that have been trained on and embody unauthorized copies of the Publishers' Works. Microsoft uses these models to provide infringing content and, at times, misinformation to users of its products and online services. During a quarterly earnings call in October 2023, Microsoft noted that "more than 18,000 organizations now use Azure OpenAI Service, including new-to-Azure customers."

23. The OpenAI Defendants consist of a web of interrelated Delaware entities.

24. Defendant OpenAI, Inc. is a Delaware nonprofit corporation with a principal place of business located at 3180 18th Street in San Francisco, California. OpenAI, Inc. was formed in December 2015. At least until October 28, 2025, OpenAI, Inc. indirectly owned and controlled all other OpenAI entities and has been directly involved in perpetrating the mass infringement and other unlawful conduct alleged here.

25. Defendant OpenAI LP is a Delaware limited partnership with its principal place of business located at 3180 18th Street in San Francisco, California. OpenAI LP was formed in 2019. OpenAI LP is a wholly owned subsidiary of OpenAI, Inc. that is operated for profit and is controlled by OpenAI, Inc. OpenAI LP was directly involved in perpetrating the mass infringement and commercial exploitation of the Publishers' Works alleged here.

26. Defendant OpenAI GP, LLC is a Delaware limited liability company with a principal place of business located at 3180 18th Street in San Francisco, California. OpenAI GP, LLC is the general partner of OpenAI LP, and it manages and operates the day-to-day business and affairs of OpenAI LP. OpenAI GP, LLC is wholly owned and controlled by OpenAI, Inc. OpenAI, Inc. uses OpenAI GP, LLC to control OpenAI LP and OpenAI Global, LLC. OpenAI GP, LLC was involved in perpetrating the mass infringement and unlawful exploitation of the Publishers' Works alleged here through its direction and control of OpenAI LP and OpenAI Global, LLC.

27. Defendant OpenAI, LLC is a Delaware limited liability company with a principal place of business located at 3180 18th Street, San Francisco, California. OpenAI, LLC was formed in September 2020. OpenAI, LLC owns, sells, licenses, and monetizes a number of OpenAI's offerings, including ChatGPT, ChatGPT Enterprise, and OpenAI's API tools, all of which were built on OpenAI's mass infringement and unlawful exploitation of the Publishers' Works. Upon

information and belief, OpenAI, LLC is owned and controlled by both OpenAI, Inc. through OpenAI Global, LLC and OpenAI OpCo, LLC.

28. Defendant OpenAI OpCo, LLC is a Delaware limited liability company with a principal place of business located at 3180 18th Street, San Francisco, California. OpenAI OpCo, LLC is a wholly owned subsidiary of OpenAI, Inc. and has facilitated and directed OpenAI's mass infringement and unlawful exploitation of the Publishers' Works through its management and direction of OpenAI, LLC.

29. Defendant OpenAI Global, LLC is a Delaware limited liability company formed in December 2022. OpenAI Global, LLC has a principal place of business located at 3180 18th Street, San Francisco, California. OpenAI, Inc. has a majority stake in OpenAI Global, LLC, indirectly through OpenAI Holdings, LLC and OAI Corporation, LLC. OpenAI Global, LLC was and is involved in unlawful conduct alleged herein through its ownership, control, and direction of OpenAI, LLC.

30. Defendant OAI Corporation, LLC is a Delaware limited liability company with a principal place of business located at 3180 18th Street, San Francisco, California. OAI Corporation, LLC's sole member is OpenAI Holdings, LLC. OAI Corporation, LLC was and is involved in the unlawful conduct alleged herein through its ownership, control, and direction of OpenAI Global, LLC and OpenAI, LLC.

31. Defendant OpenAI Holdings, LLC is a Delaware limited liability company, whose sole members are OpenAI, Inc. and Aestas, LLC, whose sole member, in turn, is Aestas Management Company, LLC. Aestas Management Company, LLC is a Delaware shell company formed for the purpose of executing a $495 million capital raise for OpenAI.

32. On or about October 28, 2025, OpenAI published the following statements on its website:[2]

- OpenAI has completed its recapitalization, simplifying its corporate structure. The nonprofit remains in control of the for-profit, and now has a direct path to major resources before AGI arrives.

  The nonprofit, now called the OpenAI Foundation, holds equity in the for-profit currently valued at approximately $130 billion, making it one of the best resourced philanthropic organizations ever. The recapitalization also grants the Foundation additional ownership as OpenAI's for-profit reaches a valuation milestone.

- With our updated structure, announced on October 28, 2025:

  The nonprofit is now the OpenAI Foundation.

  The for-profit is now a public benefit corporation, called OpenAI Group PBC, which—unlike a conventional corporation—is required to advance its stated mission and consider the broader interests of all stakeholders, ensuring the company's mission and commercial success advance together.

  The OpenAI Foundation continues to control the OpenAI Group. It now holds conventional equity in OpenAI Group – with all stockholders participating proportionally in any increase in value of the OpenAI Group – aligning long-term incentives around impact and growth.

  OpenAI Foundation and OpenAI Group have the same mission.

33. Upon information and belief, OpenAI Group PBC is a Delaware public benefit corporation with its principal place of business in San Francisco, California, and OpenAI Foundation is a Delaware foundation with a principal office at 1455 3rd Street, San Francisco California 94158.[3]

[2] *See* Bret Taylor, *Built to benefit everyone*, OPENAI (October 28, 2025), https://openai.com/index/built-to-benefit-everyone/ and *Our Structure*, OPENAI, https://openai.com/our-structure/ (last visited November 13, 2025).
[3] *Id.*

III. FACTUAL ALLEGATIONS

A. The Publishers

34. Each of the Publishers offers either a regional newspaper – that is, a newspaper that serves either medium-sized American city and its surrounding area – or one of several newspapers in a major metropolitan area. In the case of the regional publications, these newspapers are often the only means the served communities have to access information on local events, politics, sports, obituaries, the arts, and other topics of interest. In the case of the metropolitan newspapers, the Publishers offer the cities additional information on local events, and in many cases, alternative perspectives on local, national, and international issues.

35. The Publishers expend significant time and effort investigating and reporting local stories, and rely mainly on ad and subscription revenue to further their enterprises. Defendants' actions threaten the Publishers' continued efforts to provide American communities with quality, in-depth local journalism. By designing, training, and operating AI models that pilfer, copy, memorize, and replicate the Publishers' Works without compensation to the Publishers, Defendants deprive the Publishers of visits to their sites, decrease Publishers' ad and subscription revenue, and threaten to diminish (or already have diminished) the overall value of the Publishers' enterprises.

36. The fact that the Publishers' newspapers cover regional markets, or may be considered alternative newspapers in a metro area, does not diminish their importance to millions of people. Even smaller newspapers serve a wide audience. The Virginian-Pilot, for example, serves the Norfolk, Virginia, region, which has a population of about 1.8 million people.[4] Among other things, the Pilot serves one of the most densely populated military areas in the country. Over 100,000 people (military and civilian) are attached to Naval Station Norfolk, and another 15,000

[4] *See* https://censusreporter.org/profiles/31000US47260-virginia-beach-chesapeake-norfolk-va-nc-metro-area/.

people are attached to Naval Air Station Oceana in Virginia Beach. According to recent data, the Pilot's daily print and online publications reach about 219,000 people, and about 243,000 people on Sunday. In addition, pilotonline.com has about 3.9 million page views, and 1.5 million unique users, per month.[5]

37. To preserve the vitality of their print and on-line publications, the Publishers go to great lengths to protect their content. Each of the Publishers' print- and on-line editions includes copyright management information as defined in 17 U.S.C. § 1202(c). The Publishers have also registered copyrights. A chart containing a list of the Publishers' registered copyrights from 1978 to date is attached hereto as Exhibits A-I.

38. Beyond the Publishers' exclusive rights of reproduction, adaptation, publication, performance, and display under the copyright laws, the Publishers use paywalls to protect their content. Their online editions also specify terms of service or terms of use that govern access to the Publishers' material, and that restrict the use of the content provided on their websites.[6]

39. The terms of use include specific prohibitions relevant to this lawsuit. For example, several terms of use provide: "All Content is provided for informational purposes only. Subject to these Terms of Use, you may use the Content solely for your personal, non-commercial use, provided you do not remove any trademark, copyright, or other proprietary notice from the Content."[7] The terms of use also provide, "Except as expressly permitted by these Terms of Use or written permission from MediaNews Group, you agree that you will not, and will not permit any third party to: . . . (iii) use any robots, spiders, crawlers, data mining or extraction technology, or other automated scripts or means to collect or scrape information from or otherwise interact

[5] *See* https://www.virginiamedia.com/thevirginianpilot/.
[6] *See, e.g., General Terms of Use,* MEDIANEWS GROUP https://www.medianewsgroup.com/terms-of-use/ (last visited November 19, 2025).
[7] *See id.,* § 2.2 (emphasis added).

with the Technology or copy our Content without written permission; (iv) without our express prior written consent, use or attempt to use any Content or information available on the Platform, whether by manual input or automated means, for purposes of retrieval augmented generation, grounding, training, or development of large language or machine learning models, algorithms, or artificial intelligence (AI) systems, or to generate substitute content or develop any products, services, or technology[.]" [8]

40. The Publishers require that any third party that wishes to use their content obtain a license to do so. These licensing agreements allow the Publishers to control how third parties receive and display their content. The Publishers license their content only under narrowly tailored terms that provide explicit guardrails regarding how and to what extent third parties can use the content. By these means, the Publishers have reasonably sought to retain control over how third parties access and use the Publishers' content.

41. In sum, any party that accesses the content of the websites is forbidden from using the material for a commercial purpose, and specifically from removing copyright and trademark information from the content, "scraping" the content from the websites, and using the content to train AI systems. Defendants have knowingly and intentionally violated each of these provisions, and continue to do so.

B. Defendants

1. A Joint Enterprise Based on Mass Copyright Infringement

42. OpenAI was formed in December 2015 as a "non-profit artificial intelligence research company." OpenAI started with $1 billion in seed money from its founders, a group of wealthy technology entrepreneurs and investors, and companies like Amazon Web Services and

[8] *See id.,* § 3.1 (emphasis added).

InfoSys. This group included Elon Musk, the CEO of Tesla and X Corp. (formerly known as Twitter); Reid Hoffman, the co-founder of LinkedIn; Sam Altman, the former president of Y Combinator; and Greg Brockman, the former Chief Technology Officer of Stripe.

43. Despite accepting very large investments from enormously wealthy companies and individuals at its founding, OpenAI originally maintained that its research and work would be entirely unmotivated by profit. In a December 11, 2015, press release, Brockman and co-founder Ilya Sutskever (now OpenAI's President and Chief Scientist, respectively) wrote: "Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact."[9] In accordance with that mission, OpenAI promised that its work and intellectual property would be open and available to the public, that its "[r]esearchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code" and that its "patents (if any) will be shared with the world."[10]

44. Despite its early promises of altruism, OpenAI quickly became a multi-billion dollar for-profit business built in large part on the unlicensed exploitation of copyrighted works belonging to Publishers and others. Just three years after its founding, OpenAI shed its exclusively nonprofit status. It created OpenAI, LP in March 2019, a for-profit company dedicated to conducting the lion's share of OpenAI's operations—including product development—and to raising capital from investors seeking a return. OpenAI's corporate structure grew into an intricate web of for-profit holding, operating, and shell companies that manage OpenAI's day-to-day operations and grant OpenAI's investors (most prominently, Microsoft) authority and influence

[9] Greg Brockman & Ilya Sutskever, *Introducing OpenAI,* OPENAI (Dec. 11, 2015), https://openai.com/blog/introducing-openai.
[10] *Id.*

over OpenAI's operations, all while raising billions in capital from investors. The result: OpenAI today is a commercial enterprise valued at $500 billion.[11]

45. With the transition to for-profit status came another change: OpenAI also ended its commitment to openness. OpenAI released the first two iterations of its flagship generative artificial intelligence model ("GenAI model"), GPT-1 and GPT-2, on an open-source basis in 2018 and 2019, respectively. But OpenAI changed course in 2020, starting with the release of GPT-3 shortly after OpenAI LP and other for-profit OpenAI entities were formed and took control of product design and development.

46. GPT-3.5 and GPT-4 are both orders of magnitude more powerful than the two previous generations, yet Defendants have kept their design and training entirely a secret. For previous generations, OpenAI had voluminous reports detailing the contents of the training set, design, and hardware of the LLMs. Not so for GPT-3.5 or GPT-4. For GPT-4, for example, the "technical report" that OpenAI released said: "this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."[12]

47. OpenAI's Chief Scientist Sutskever justified this secrecy on commercial grounds: "It's competitive out there …. And there are many companies who want to do the same thing, so from a competitive side, you can see this as maturation of the field."[13] But its effect was clearly to conceal the identity of the data OpenAI copied to train its latest models from rightsholders like the Publishers.

---

[11] *See* https://www.cnbc.com/2025/10/28/open-ai-for-profit-microsoft.html (last visited November 13, 2025).
[12] OPENAI, GPT-4 TECHNICAL REPORT (2023), https://cdn.openai.com/papers/gpt-4.pdf.
[13] James Vincent, *OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: 'We Were Wrong'*, THE VERGE (Mar. 15, 2023), https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closedresearch-ilya-sutskever-interview.

48. OpenAI became a household name upon the release of ChatGPT in November 2022. ChatGPT is a text-generating chatbot that, given user-generated prompts, can mimic humanlike natural language responses. ChatGPT was an instant viral sensation, reaching one million users within a month of its release and gaining over 100 million users within three months. As of Summer 2025, ChatGPT receives an average of 330 million prompts daily from users within the U.S.

49. OpenAI, through OpenAI OpCo, LLC and at the direction of OpenAI, Inc., OpenAI LP, and other OpenAI entities, offers a suite of services powered by its LLMs, targeted to both ordinary consumers and businesses. OpenAI's business-focused offerings include ChatGPT Enterprise and ChatGPT API tools designed to enable developers to incorporate ChatGPT into bespoke applications. OpenAI also licenses its technology to corporate clients for licensing fees. Some of these offerings can be accessed without charge, but OpenAI charges a fee for the use of its more powerful or specialized offerings.

50. Upon information and belief, and insofar as relevant to this lawsuit, all of OpenAI's models have been developed using similar methods, and display similar features. The OpenAI models at issue in this lawsuit include GPT-1, GPT-2, GPT-3, GPT-3.5, GPT-3.5 Turbo, GPT-4, GPT-4 Turbo, GPT-4o, GPT-4o mini, GPT-4o mini Search Preview, GPT-4o Search Preview o1,o1-pro, o1-mini, o1 Preview, GPT-4.5 o3, o3-pro, o3-deep-research, o3-mini, o4-mini, o4-mini-deep-research, GPT-4.1, GPT-4.1 mini, GPT-OSS, gpt-oss-120b, gpt-oss-20b, GPT-5, GPT-5 mini, GPT-5 nano, GPT-5 pro, GPT 5.1, ChatGPT-4o, and GPT-4 Chat. Upon information and below, OpenAI plans to introduce new models that use similar methods and provide similar output, including the Publishers' content.

51. In addition, OpenAI offers the following GPT products that leverage its LLMs: application programming interface ("API") platform, ChatGPT (Free, Pro, Plus), ChatGPT Enterprise, ChatGPT Business (formerly ChatGPT Team), ChatGPT Education, ChatGPT Plugins (including original Browse with Bing plugin), Custom GPTs / GPTs (including Remove Paywall, News Summarizer, etc.), ChatGPT Search, ChatGPT Canvas, ChatGPT Agent (formerly Operator), ChatGPT Atlas (web browser), and ChatGPT apps. Upon information and belief, OpenAI plans to introduce new products that use similar methods and provide similar output (including the Publishers' content).

52. OpenAI's models and products have been immensely valuable for OpenAI. Over 80% of Fortune 500 companies are using ChatGPT.[14] OpenAI's founder Sam Altman anticipates annual revenues of $100 billion by 2027,[15] and as noted above, OpenAI's recent recapitalization valued OpenAI in the neighborhood of $500 billion.

53. OpenAI's commercial success is built in large part on its large-scale copyright infringement. One of the central features driving the use and sales of ChatGPT and its associated products is the LLM's ability to produce natural language text in a variety of styles. To achieve this result, OpenAI made numerous reproductions of copyrighted works, including the Publishers' Works, in the course of "training" the LLM.

54. Upon information and belief, all of the OpenAI Defendants have been either directly involved in or have directed, controlled, and profited from OpenAI's widespread infringement and commercial exploitation of the Publishers' Works. OpenAI, Inc., alongside

[14] OpenAI, *Introducing ChatGPT Enterprise*, OPENAI (Aug. 28, 2023), https://openai.com/blog/introducing-chatgpt-enterprise.

[15] *See* Anthony Ha, *Sam Altman says 'enough' to questions about OpenAI's revenue*, TECHCRUNCH (Nov. 2, 2025) https://techcrunch.com/2025/11/02/sam-altman-says-enough-to-questions-about-openais-revenue/ (last visited November 13, 2025).

Microsoft, controlled and directed the widespread reproduction, distribution, and commercial use of the Publishers' Works perpetrated by OpenAI LP and OpenAI Global, LLC, through a series of holding and shell companies that include OpenAI Holdings, LLC, OpenAI GP, LLC, and OAI Corporation, LLC., OpenAI LP and OpenAI Global, LLC were directly involved in the design, development, and commercialization of OpenAI's GPT-based products, and directly engaged in the widespread reproduction, distribution, and commercial use of the Publishers' Works. OpenAI LP and OpenAI Global, LLC also controlled and directed OpenAI, LLC and OpenAI OpCo, LLC, which were involved in distributing, selling, and licensing OpenAI's GPT-based products, and thus monetized the reproduction, distribution, and commercial use of the Publishers' Works. As noted above, as of October 28, 2025, these entities are now under the organizational umbrellas of OpenAI Foundation and Open AI Group PBC.

55. Since at least 2019, Microsoft has been, and continues to be, intimately involved in the training, development, and commercialization of OpenAI's GPT products. In an interview with the Wall Street Journal at the 2023 World Economic Forum, Microsoft CEO Satya Nadella said that the "ChatGPT and GPT family of models … is something that we've been partnered with OpenAI deeply now for multiple years." Through this partnership, Microsoft has been involved in the creation and commercialization of GPT LLMs and products based on them in at least two ways.

56. Microsoft created and operated bespoke computing systems to execute the mass copyright infringement detailed herein. These systems were used to create multiple reproductions of the Publishers' intellectual property for the purpose of creating the GPT models that exploit and, in many cases, retain large portions of the copyrightable expression contained in those works.

57. Microsoft is the sole cloud computing provider for OpenAI. Microsoft and OpenAI collaborated to design the supercomputing systems powered by Microsoft's cloud computer platform Azure, which were used to train all OpenAI's GPT models after GPT-1. In a July 2023 keynote speech at the Microsoft Inspire conference, Mr. Nadella said: "We built the infrastructure to train their models. They're innovating on the algorithms and the training of these frontier models."

58. That infrastructure was not just general purpose computer systems for OpenAI to use as it saw fit. Microsoft specifically designed it for the purpose of using essentially the whole internet—curated to disproportionately feature the Publishers' Works—to train the most capable LLM in history. In a February 2023 interview, Mr. Nadella said:

> But beneath what OpenAI is putting out as large models, remember, the heavy lifting was done by the [Microsoft] Azure team to build the computer infrastructure. Because these workloads are so different than anything that's come before. So we needed to completely rethink even the datacenter up to the infrastructure that first gave us even a shot to build the models. And now we're translating the models into products.[16]

59. Microsoft built this supercomputer "in collaboration with and exclusively for OpenAI," and "designed [it] specifically to train that company's AI models."[17] Even by supercomputing standards, it was unusually complex. According to Microsoft, it operated as "a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server."[18] This system ranked in the top five most powerful publicly known supercomputing systems in the world.

---

[16] *First on CNBC: CNBC Transcript: Microsoft CEO Satya Nadella Speaks with CNBC's Jon Fortt on "Power Lunch" Today*, CNBC (Feb. 7, 2023), https://www.cnbc.com/2023/02/07/first-on-cnbc-cnbc-transcriptmicrosoft-ceo-satya-nadella-speaks-with-cnbcs-jon-fortt-on-power-lunch-today.html.

[17] Jennifer Langston, *Microsoft Announces New Supercomputer, Lays Out Vision for Future AI Work*, MICROSOFT (May 19, 2020), https://news.microsoft.com/source/features/ai/openai-azure-supercomputer/.

[18] *Id.*

60. To ensure that the supercomputing system suited OpenAI's needs, Microsoft needed to test the system, both independently and in collaboration with OpenAI software engineers. According to Mr. Nadella, with respect to OpenAI: "They do the foundation models, and we [Microsoft] do a lot of work around them, including the tooling around responsible AI and AI safety." Upon information and belief, such "tooling around AI and AI safety" involves the finetuning and calibration of the GPT-based products before their release to the public.[19]

61. Second, in collaboration with OpenAI, Microsoft has also commercialized OpenAI's GPT-based technology, and combined it with its own Bing search index. In February 2023, Microsoft unveiled Bing Chat (now Copilot), a generative AI chatbot feature on its search engine powered by GPT-4. In May 2023, Microsoft and OpenAI unveiled "Browse with Bing," a plugin to ChatGPT that enabled it to access the latest content on the internet through the Microsoft Bing search engine. Copilot and Browse with Bing combine GPT-4's ability to mimic human expression—including the Publishers' expression—with the ability to generate natural language summaries of search result contents, including hits on the Publishers' Works, that obviate the need to visit the Publishers' websites. These "synthetic" search results purport to answer user queries directly and may include extensive paraphrases and direct quotes of the Publishers' reporting. Such copying maintains engagement with Defendants' own sites and applications instead of referring users to the Publishers' websites in the same way as organic listings of search results.

62. In an interview, Mr. Nadella acknowledged Microsoft's intimate involvement in OpenAI's operations and, therefore, its copyright infringement:

> [W]e were very confident in our own ability. We have all the IP rights and all the capability. If OpenAI disappeared tomorrow, I don't want any customer of ours to be worried about it quite honestly, because we have all of the rights to continue the

[19] SÉBASTIEN BUBECK ET AL., SPARKS OF ARTIFICIAL GENERAL INTELLIGENCE: EARLY EXPERIMENTS WITH GPT-4 (2023), https://arxiv.org/pdf/2303.12712.pdf.

> innovation. Not just to serve the product, but we can go and just do what we were doing in partnership ourselves. We have the people, we have the compute, we have the data, we have everything.[20]

63. Through their collaboration in both the creation and the commercialization of the GPT models, Defendants have profited from the massive copyright infringement, commercial exploitation, and misappropriation of the Publishers' intellectual property. As Mr. Nadella put it, "[OpenAI] bet on us, we bet on them."[21] He continued, describing the effect of Microsoft's $13 billion investment:

> And that gives us significant rights as I said. And also this thing, it's not hands off, right? We are in there. We are below them, above them, around them. We do the kernel optimizations, we build tools, we build the infrastructure. So that's why I think a lot of the industrial analysts are saying, 'Oh wow, it's really a joint project between Microsoft and OpenAI.' The reality is we are, as I said, very self-sufficient in all of this.[22]

2. How GenAI Models Work

64. Microsoft and OpenAI created and distributed reproductions of the Publishers' Works in several independent ways while training their LLMs and operating the products that incorporate them.

65. At the collection stage (also known as the pre-training stage), Defendants collect and store a vast amount of content scraped from the internet, including content scraped from Publishers' websites that includes Publishers' Works. OpenAI then creates datasets from that content which is later used to train the LLMs.

---

[20] *Satya Nadella on Hiring the Most Powerful Man in AI When OpenAI threw Sam Altman overboard, Microsoft's CEO saw an opportunity,* NEW YORK MAGAZINE (Apr. 17, 2024), https://nymag.com/intelligencer/2023/11/on-with-kara-swisher-satya-nadella-on-hiring-sam-altman.html.

[21] Steven Levy, *Microsoft's Satya Nadella is Betting Everything on AI,* WIRED (June 13, 2023), https://www.wired.com/story/microsofts-satya-nadella-is-betting-everything-on-ai/.

[22] *Satya Nadella on Hiring the Most Powerful Man in AI When OpenAI threw Sam Altman overboard, Microsoft's CEO saw an opportunity, supra* n. 20.

66. At the training stage, Defendants train their LLMs on the collected content by a process that feeds the data through its LLMs to generate output. Appending the output of an LLM to its input and feeding it back into the model produces sentences and paragraphs word by word. This is how ChatGPT products generate responses to user queries, or "prompts."

67. LLMs encode the information from the training corpus that they use to make these predictions as numbers called "parameters." For example, there are approximately 1.76 trillion parameters in the GPT-4 LLM.

68. The process of setting the values for an LLM's parameters during the training process involves storing copies of the training articles in computer memory, providing a portion of the article to the model, and adjusting the parameters of the model so that the model accurately predicts the next word in the article.

69. After being trained on a general corpus, models may be further subject to "fine-tuning" by, for example, performing additional rounds of training using specific types of works to better mimic their content or style, or providing the models with human feedback to reinforce desired or suppress undesired behaviors in order to improve the model's ability to follow instructions.

70. Models trained in this way are known to exhibit a behavior called "memorization."[23] That is, given the right prompt, LLMs will repeat large portions of materials they were trained on. This phenomenon shows that LLM parameters encode retrievable copies of many of those training works.

71. Once trained, LLMs may be provided with information specific to a use case or subject matter in order to "ground" their outputs. This is accomplished through a process called

[23] Gerrit J.J. Van Den Burg & Christopher K.I. Williams, *On Memorization In Probabilistic Deep Generative Models* (2021), https://proceedings.neurips.cc/paper/2021/file/eae15aabaa768ae4a5993a8a4f4fa6e4- Paper.pdf.

retrieval augmented generation or "RAG." For example, an LLM may be asked to generate a text output based on specific external data, such as a document, provided as context. Using this method, OpenAI's synthetic search applications: (1) receive an input, such as a question; (2) retrieve relevant documents related to the input prior to generating a response; (3) combine the original input with the retrieved documents in order to provide context; and (4) provide the combined data to an LLM, which generates a natural-language response. As shown below, search results generated in this way may extensively copy or closely paraphrase works that the models themselves may not have memorized.

C. Defendants' Unauthorized Use and Copying of the Publishers' Works

72. Microsoft and OpenAI created and distributed reproductions of the Publishers' Works in several, independent ways in the course of training their LLMs and operating the products that incorporate them.

1. Unauthorized Reproduction of the Publishers' Works During GPT Model Training

73. OpenAI's GPT models are a family of LLMs, the first of which was introduced in 2018, followed by GPT-2 in 2019, GPT-3 in 2020, GPT-3.5 in 2022, GPT-4 in 2023, GPT-4o and OpenAI o1 in 2024, and by 2025 including all of the LLM models listed in paragraph 50. The "chat" style LLMs were developed in two stages. First, a transformer model was pre-trained on a very large amount of data. Second, the model was "fine-tuned" on a much smaller, supervised dataset in order to help the model solve specific tasks. As noted above, OpenAI has introduced numerous additional models, and continues to develop new and/or derivative models using these methods.

74. The pre-training step involved collecting and storing text content to create training datasets and processing that content through the GPT models. While OpenAI has not released its

training data, OpenAI has published general information about its pre-training process for the GPT models.[24]

75. GPT-2 includes 1.5 billion parameters, which was a 10X scale up of GPT.[25] The training dataset for GPT-2 includes an internal corpus OpenAI built called "WebText," which includes "the text contents of 45 million links posted by users of the 'Reddit' social network."[26] The contents of the WebText dataset were created as a "new web scrape which emphasizes document quality."[27]

76. The WebText dataset contains a large amount of content scraped from the Publishers' websites. GPT-3 includes 175 billion parameters and was trained on the datasets listed in the table below.[28]

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

77. OpenAI developed the WebText2 dataset to prioritize high value content. Like the original WebText, it is composed of popular outbound links from Reddit. As shown in the table above, the WebText2 corpus was weighted 22% in the training mix for GPT-3 despite constituting less than 4% of the total tokens in the training mix. Like the original WebText, OpenAI describes

[24] OpenAI, *Better Language Models and Their Implications*, OPENAI (Feb. 14, 2019), https://openai.com/research/better-language-models.
[25] *Id.*
[26] *GPT-2 Model Card*, GITHUB (Nov. 2019), https://github.com/openai/gpt-2/blob/master/model_card.md.
[27] Radford et al., *Language Models Are Unsupervised Multitask Learners 3* (2018), https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.
[28] Brown et al., *Language Models Are Few-Shot Learners 9* (2020), https://arxiv.org/pdf/2005.14165.pdf.

WebText2 as a "high-quality" dataset that is "an expanded version of the WebText dataset … collected by scraping links over a longer period of time."[29]

78. The most highly weighted dataset in GPT -3, Common Crawl, is a "copy of the Internet" made available by an eponymous 501(c)(3) organization run by wealthy venture capital investors.[30] According to publicly available tax statement, in 2023 OpenAI made a $250,000 contribution to Common Crawl.[31] For example, in C4, a filtered English-language subset of a 2019 snapshot of Common Crawl, the Publishers' websites account for approximately 40.8 million tokens (basic units of text), broken down as follows:[32]

Publisher URL	**C4 Database (Tokens)**
https://www.sandiegouniontribune.com/	8.9M
https://www.mcall.com/	7.8M
https://www.courant.com/	7.0M
https://www.dailycamera.com/	2.8M
https://www.dailynews.com/	2.8M
https://www.dailypress.com/	2.8M
https://www.sbsun.com/	2.4M
https://www.bostonherald.com/	1.8M
https://www.pilotonline.com/	220k
Total	**40.82M**

79. Critically, OpenAI admits that "datasets we view as higher-quality are sampled more frequently" during training.[33] Accordingly, by OpenAI's own admission, high-quality

[29] *Id.* at 8.
[30] COMMON CRAWL, https://commoncrawl.org/ (last visited November 25, 2025).
[31] https://projects.propublica.org/nonprofits/organizations/261635908/202403189349101980/full
[32] Dodge et al., *Documenting Large Webtext Corpora: A Case Study On The Colossal Clean Crawled Corpus* (2021), https://arxiv.org/abs/2104.08758.
[33] Brown et al., *supra* n. 28.

content, including the Publishers' Works, was more important and valuable for training the GPT models as compared to content taken from other, lower-quality sources.

80. While OpenAI has not publicly released much information about GPT-4, experts suspect that GPT-4 includes 1.8 trillion parameters, which is over 10X larger than GPT-3, and was trained on approximately 13 trillion tokens.[34] The training set for GPT-3, GPT-3.5, and GPT-4 was comprised of 45 terabytes of data—the equivalent of a Microsoft Word document that is over 3.7 billion pages long.[35]

81. On information and belief, the Defendants have used, and continue to use, the WebText, WebText2, and other training datasets to train the GPT models. For example, ChatGPT's "knowledge cutoff date" – the last date for which its ChatGPT products will offer information to users – has shifted from as early as September 2021 to as recently as May 31, 2024, which demonstrates that the Defendants are continuing to create and use unauthorized copies of the Publishers' Works contained in the training datasets and elsewhere on the internet.

82. Defendants repeatedly copied the Publishers' Works, without any license or other compensation to the Publishers. As part of training the GPT models, Microsoft and OpenAI collaborated to develop a complex, bespoke supercomputing system to house and reproduce copies of the training dataset, including copies of the Publishers' Works. Hundreds of thousands of the Publishers' Works were copied and ingested—multiple times—for the purpose of "training" Defendants' GPT models.

83. Upon information and belief, OpenAI continues to create unauthorized copies of Publishers' Works, and is currently or will imminently commence making additional copies of

[34] Maximilian Schreiner, *GPT-4 Architecture, Datasets, Costs and More Leaked*, THE DECODER (July 11, 2023), https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.
[35] Kindra Cooper, *OpenAI GPT-3: Everything You Need to Know [Updated]*, SPRINGBOARD (Sept. 27, 2023), https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/.

websites to train and/or fine-tune new models for use in yet-to-be-released products, and for use in creating RAG-generated content.

84. Upon information and belief, Microsoft and OpenAI acted jointly in the large-scale copying of the Publishers' Works involved in generating the GPT models programmed to accurately mimic the Publishers' Works and writers. Microsoft and OpenAI collaborated in designing the GPT models, selecting the training datasets, and supervising the training process. As Mr. Nadella stated:

> So, there are a lot of, I call it, product design choices one gets to make when you think about AI and AI safety. Then, let's come at it the other way. You have to take real care of the pretrained data because models are trained on pretrained data. What's the quality, the provenance of that pretrained data? That's a place where we've done a lot of work.[36]

85. To the extent that Microsoft did not select the works used to train the GPT models, it acted in self-described "partnership" with OpenAI respecting that selection, knew or was willfully blind to the identity of the selected works by virtue of its knowledge of the nature and identity of the training corpora and selection criteria employed by OpenAI, and/or had the right and ability to prevent OpenAI from using any particular work for training by virtue of its physical control of the supercomputer it developed for that purpose and its legal and financial influence over the OpenAI Defendants.

86. Upon information and belief, Microsoft and OpenAI continue to create unauthorized copies of the Publishers' Works in the form of synthetic search results returned by their Copilot and Browse with Bing products. Microsoft actively gathers copies of the Publishers'

[36] Nilay Patel, *Microsoft Thinks AI Can Beat Google at Search — CEO Satya Nadella Explains Why*, THE VERGE (Feb. 7, 2023), https://www.theverge.com/23589994/microsoft-ceo-satya-nadella-bing-chatgpt-googlesearch-ai.

Works used to generate such results in the process of crawling the web to create the index for its Bing search engine.

87. Upon information and belief, Microsoft and OpenAI continue to create unauthorized copies of the Publishers' Works, and are currently or will imminently commence making additional copies of websites to train and/or fine-tune new models for use in yet-to-be-released products, and for use in creating RAG-generated content.

88. Defendants' large-scale commercial exploitation of the Publishers' Works is not licensed, nor have Defendants received permission from the Publishers to copy and use their works to build their GenAI tools.

2. Unauthorized Reproductions and Derivatives of the Publishers' Works Embodied in the GPT Models and Unauthorized Public Display of the Publishers' Works in GPT Product Outputs

89. As further evidence of being trained using unauthorized copies of the Publishers' Works, the GPT LLMs themselves have "memorized" copies of many of those same works encoded into their parameters. As shown below, the current GPT-4 LLM will output near-verbatim copies of significant portions of the Publishers' Works when prompted to do so, thus demonstrating that these articles were used to train their GPT LLMs and that these LLMs have memorized Publishers' Works.

90. Such memorized examples constitute unauthorized copies or derivative works of the Publishers' Works used to train the model. Defendants directly engaged in the unauthorized reproduction and publication of the Publishers' Works as part of generative output provided by their products built on the GPT models. Defendants' commercial applications built using GPT models, listed above, include, *inter alia*, ChatGPT (including its associated offerings, ChatGPT Plus, ChatGPT Enterprise) and OpenAI's API Platform. These products display the Publishers' Works in generative output in at least two ways: (1) by showing "memorized" copies or derivatives

of the Publishers' Works retrieved from the models themselves, and (2) by showing synthetic search results that are substantially similar to the Publishers' Works generated from copies that OpenAI has obtained from a repository containing Publishers' Works (e.g., a search index) or from the internet in real-time.

91. For example, ChatGPT displays copies or derivatives of the Publishers' Works memorized by the underlying GPT models in response to user prompts. Upon information and belief, the underlying GPT models for ChatGPT were trained on these and many more of the Publishers' Works and are able to generate such expansive summaries and verbatim text.

92. In addition, synthetic search products built on the GPT LLMs have outputted the contents of search results, including the Publishers' Works, that may not have been included in the LLMs' training set through RAG. As noted above, RAG includes receiving a prompt from a user, using the prompt to search for the Publishers' Works from the internet, providing the prompt together with a copy of the works as additional context for the LLM, and having the LLM use the works to create natural-language substitutes that serve the same informative purpose as the original.

93. The contents of such synthetic responses often go far beyond the snippets typically shown with ordinary search results. Even when synthetic search responses include links to source materials, users have less need to navigate to those sources because their expressive content is already included in the narrative result. Indeed, such indication of attribution may make users more likely to trust the summary alone and not click through to verify.

94. In this way, synthetic search results divert important traffic away from copyright holders like the Publishers. A user who has already read the latest news, even—or especially—with attribution to one of the Publishers, has less reason to visit the original source.

95. Below are illustrative and non-exhaustive examples of synthetic search results that include misappropriated Publishers' Works:

96. From the Hartford Courant:[37]

Model output:	**Text from Hartford Courant:**
top aide to U.S. Attorney John Durham in his Russia investigation, has quietly resigned from the U.S. Justice Department probe - at least partly out of concern that the investigative team is being pressed for political reasons to produce a report before its work is done, colleagues said. Dannehy, a highly respected prosecutor who has worked with or for Durham for decades, informed colleagues in the New Haven U.S. attorney's office of her resignation from the Department of Justice by email Thursday evening. The short email was a brief farewell message and said nothing about political pressure, her work for Durham or what the Durham team has produced	top aide to U.S. Attorney John H. Durham in his Russia investigation, has quietly resigned - at least partly out of concern that the investigative team is being pressed for political reasons to produce a report before its work is done, colleagues said. Dannehy, a highly regarded prosecutor who has worked with or for Durham for decades, informed colleagues in the U.S. Attorney's office in New Haven of her resignation from the Department of Justice by email Thursday evening. The short email was a brief farewell message and said nothing about political pressure, her work for Durham or what the Durham team has produced, according to people who received it.

97. From the Los Angeles Daily News:[38]

Model output:	**Text from Los Angeles Daily News:**
to turn a half block of black into a sea of gray. The morning temperature of the asphalt in the sun: 93 degrees. After the seal:a cool 70 - on what would turn out to be the first major heat wave of the year. "It's awesome. It's very cool - both literally and figuratively," exclaimed Councilman Bob	to turn a half block of black into a sea of gray. The morning temperature of the black asphalt in the middle of a nearby intersection read 93 degrees. The new light gray surface on Jordan Avenue read a cool 70 - on what would turn out to be the first heat wave of the year.

[37] *See* https://www.courant.com/2020/09/11/nora-dannehy-connecticut-prosecutor-who-was-top-aide-to-john-durhams-trump-russia-investigation-resigns-amid-concern-about-pressure-from-attorney-general-william-barr/ (last visited November 25, 2025).

[38] *See* https://www.dailynews.com/2017/05/20/cool-pavement-to-cut-urban-street-heat-gets-first-california-tryout-in-canoga-park/ (last visited November 25, 2025).

Blumenfield, whose district includes Canoga Park. "We are trying to control 'the heat island effect' " - or hotter temperatures caused by urban sprawl. "The downside: we won't be able to fry eggs on the streets," he joked.	"It's awesome. It's very cool - both literally and figuratively," exclaimed Councilman Bob Blumenfield, whose Los Angeles district includes Canoga Park, squinting into the laser handheld thermometer. "We are trying to control 'the heat island effect' " – or hotter temperatures caused by urban sprawl. "The downside: we won't be able to fry eggs on the streets."

98. From the Daily Camera:[39]

Model output:	**Text from Daily Camera:**
fan, Buglione has spent the better part of the past two decades doing everything she can to be a part of the sport she loves, whether it be as a player, as an official or even if it meant coaching an eight-man team in Colorado's smallest classification. But in the male-dominated world of high school football, Buglione is indeed a trailblazer. She is believed to be the first female head football coach in Colorado history after being hired by Nederland High School in June 2017.	fan, Buglione has spent the past two decades doing everything she could to be a part of the sport she loves, whether it was as a player, as an official or even if it meant "carrying someone's bag." But when Buglione, who moved to Firestone from Oregon in May, agreed to take the reins of the football program at Nederland High School in early June, she is believed to have become the first female head coach in Colorado history.

99. From the Boston Herald:[40]

Model output:	**Text from Boston Herald:**
morning when they encountered the RV blaring a recorded warning that a bomb would detonate in 15 minutes, Metro Nashville Police Chief John Drake said. Police evacuated nearby buildings and called in	when they encountered the RV blaring a recorded warning that a bomb would detonate in 15 minutes, Metro Nashville Police Chief John Drake said. Police evacuated nearby buildings and called in the bomb

[39] *See* https://www.dailycamera.com/2017/07/01/in-beth-buglione-nederland-hires-colorados-first-female-high-school-football-coach/ (last visited, November 23, 2024)

[40] *See* https://www.bostonherald.com/2020/12/25/nashville-rocked-by-christmas-day-explosion/ (last visited November 23, 2025).

the bomb squad. The RV exploded shortly	squad. The RV exploded shortly afterward, Drake said at a news conference.

100. From the San Diego Union:[41]

Model output:	**Text from San Diego Union Tribune:**
stopping - under legislation proposed by two lawmakers who say it would make the roads safer. The two-tiered approach to the rules of the road - one for cyclists and one for cars - is unlikely to ease growing tensions over sharing California's roadways. Bike advocates have won such victories in the Statehouse as requiring drivers to yield a three-foot radius of manoeuvring room to cyclists or face fines.	stopping - under legislation proposed by two lawmakers who say it would make the roads safer. The two-tiered approach to the rules of the road - one for cyclists and one for cars - is unlikely to ease growing tensions over sharing California's roadways. Environmentally conscious bike advocates have won such victories in the statehouse as requiring drivers give them three feet of elbow room or face fines.

101. From the Daily Press:[42]

Model output:	**Text from Daily Press (VA):**
is tracking the outbreak of this respiratory illness and will continue to update this page as the state health department announces new cases.	is tracking the outbreak of this respiratory illness and will continue to update this map and other graphics as the state health department announces new cases.

102. From the San Bernardino Sun:[43]

Model output:	**Text from The Sun (San Bernardino):**

[41] *See* https://www.sandiegouniontribune.com/2017/02/22/new-bill-would-let-bicyclists-roll-through-stop-signs/ (last visited November 23, 2025).

[42] *See https://www.dailypress.com/news/health/vp-nw-coronavirus-virginia-tracking-cases-20200315-hwasfeeewbdbbcol2pcwtosqri-story.html* (last visited November 23, 2025)

[43] *See* https://www.sbsun.com/2011/08/20/dismissal-of-burum-charges-not-expected-to-affect-postmus/ (last visited November 23, 2025)

of the San Bernardino County Board of Supervisors and assessor, pleaded guilty in March 2011 to accepting a $100,000 bribe from Burum in exchange for his vote approving a $102 million settlement to Burum's investor group, Colonies Partners LP, in November 2006. The settlement ended a nearly five-year legal battle	of the San Bernardino County Board of Supervisors and former county Assessor, pleaded guilty in March to accepting a bribe from Burum in exchange for approving a $102 million legal settlement with Burum's company, Colonies Partners LP, in November 2006. The settlement ended a nearly five-year-old lawsuit

103. From Morning Call:[44]

Model output:	**Text from Morning Call:**
smiles, beautiful voices and lively African songs and dances. The program features well-loved children's songs, traditional spirituals and gospel favorites. Concerts are free and open to all. A freewill offering is taken at the performance to support African Children's Choir programs, such as education, care and relief and development programs	smiles, beautiful voices and lively African songs and dances. The organization is touring with its 50th choir since it began in 1985. *** The program features well-loved children's songs, traditional spirituals and Gospel favorites. The concerts are free and open to all with a free-will offering taken at the performance to support African Children's Choir programs, such as education, care and relief and development programs.

104. From the Virginian-Pilot:[45]

Model output:	**Text from Virginian-Pilot:**
and racketeering. Bob Heghmann, 70, filed a lawsuit Thursday in U.S. District Court, saying the national and Virginia Republican parties and some GOP	and racketeering. Bob Heghmann, 70, filed a lawsuit Thursday in U.S. District Court, saying the national and Virginia Republican parties and some GOP

[44] *See* https://www.mcall.com/2019/06/14/family-fun-african-childrens-choir-to-offer-pair-of-performances-in-valley/ (last visited November 23, 2024)

[45] *See* *https://www.pilotonline.com/2017/08/04/republican-donor-from-virginia-beach-sues-gop-accusing-the-party-of-fraud-over-failed-obamacare-repeal-2/* (last visited November 25, 2025)

leaders raised millions of dollars in campaign funds while knowing they weren't going to be able to overturn the law also known as Obamacare. The GOP "has been engaged in a pattern of Racketeering which	leaders raised millions of dollars in campaign funds while knowing they weren't going to be able to overturn the law also known as Obamacare. The GOP "has been engaged in a pattern of Racketeering which

105. In addition to the foregoing output of content from Publishers' Works, and unlike a traditional search result, the synthetic output does not include a prominent hyperlink that sends users to the Publishers' website. Rather, the output disguises the results as the work of the GPT system itself.

106. The foregoing examples, as well as the explanation of (i) the means by which OpenAI acquired the Publishers' Works for use in its models and products, and (ii) the means by which OpenAI has used the Publishers' Works in the development and operation of its models and Products, establish at least a reasonable inference that (x) OpenAI's models and products were trained on, have memorized, and have regurgitated the Publishers' Works, and (y) that all of the Publishers' Works are available and produced when prompted by user queries. Further information regarding the regurgitation and output of the Publishers' Works is solely in OpenAI's possession, custody, and control.

3. Willful Infringement

107. Defendants' unauthorized reproduction and display of the Publishers' Works is willful. Defendants were intimately involved in training, fine-tuning, and otherwise testing the GPT models. Defendants knew or should have known that these actions involved unauthorized copying of the Publishers' Works on a massive scale during training, resulted in the unauthorized encoding of huge numbers of such works in the models themselves, and would inevitably result in the unauthorized display of such works that the models had either memorized or would present to

users in the form of synthetic search results. In fact, in late 2023 before his ouster and subsequent reinstatement as OpenAI's CEO, Sam Altman reportedly clashed with OpenAI board member Helen Toner over a paper that Toner wrote criticizing the company over "safety and ethics issues related to the launches of ChatGPT and GPT-4, including regarding copyright issues."[46]

108. The Publishers put Defendants on notice that these uses of the Publishers' Works were not authorized by placing copyright notices and linking to their terms of service (which contain, among other things, terms and conditions for the use of their works) on every page of their websites whose contents Defendants copied and displayed.

109. Upon information and belief, Defendants were aware of many examples of copyright infringement after ChatGPT, Browse with Bing, and Copilot (formerly Bing Chat) were released, some of which were widely publicized. These include multiple lawsuits dating back to 2023 and pending in this Court that allege such copyright infringement.

D. Defendants' Material Contributions to End-User Infringement

110. Should Defendants argue that the end-user is the direct infringer when the Defendants' GenAI products output unauthorized copies of the Publishers' Works, Defendants directly and materially aided in such infringement by providing end users with unauthorized copies of the Publishers' Works.

111. Defendants know or should have known about infringement by end-users for multiple reasons.

112. First, the Defendants knew or reasonably should have known that training the GPT models on the Publishers' Works would result in the GenAI products outputting material that infringes the Publishers' Works. The Defendants know that the GPT models have the propensity

[46] Andrew Imbrie, Owen J. Daniels & Helen Toner, *Decoding Intentions,* CENTER FOR SECURITY AND EMERGING TECHNOLOGY (Oct. 2023).

to "memorize" training materials such that the GPT models regurgitate those training materials in response to prompts.[47] Indeed, the propensity of LLMs to memorize training data is a well-known and well-documented behavior in the industry.[48]

113. Second, the Defendants knew or reasonably should have known that end-users use the GenAI products to elicit copyrighted content based on, *inter alia*, Defendants' own acknowledgment of the issue on its website[49] and the widely publicized reporting that users were using ChatGPT's Browse with Bing plug-in to circumvent paywalls.[50] Indeed, ChatGPT's circumvention of paywalls became a viral topic of many conversations online, including one post[51] on X that received 1.9M views and a Reddit thread[52] that gained 6.3K upvotes from Reddit users. Despite recognizing that the GPT models can reproduce copyrighted content and being aware that at least some of its users use the GPT-based products to do so, Defendants continued to use copyrighted material without authorization.

114. Well publicized reporting also describes use of the GPT models to create disinformation, misinformation, or simply poor replications of newspapers' copyrighted content

---

[47] *OpenAI and Journalism,* OPENAI, https://openai.com/blog/openai-and-journalism (last visited Nov. 24, 2024).
[48] Gerrit J.J. Van Den Burg & Christopher K.I. Williams, *On Memorization in Probabilistic Deep Generative Models* (2021), https://proceedings.neurips.cc/paper/2021/file/eae15aabaa768ae4a5993a8a4f4fa6e4- Paper.pdf.
[49] *How do I use ChatGPT Browse with Bing to search the web?* INTERNET ARCHIVE, https://web.archive.org/web/20230704050417/https://help.openai.com/en/articles/8077698-how-do-i-use-chatgpt-browse-with-bing-to-search-the-web (last visited Nov. 23, 2025).
[50] *See, e.g.*, Emily Dreibelbis, *'Browse With Bing' Disabled on ChatGPT Plus Because It Bypassed Paywalls*, PC MAG (July 5, 2023), https://www.pcmag.com/news/browse-with-bing-disabled-on-chatgpt-plus-because-it-bypassed-paywalls; Trevor Mogg, *ChatGPT's Bing browsing feature disabled over paywall access flaw*, DIGITAL TRENDS (July 4, 2023), https://www.digitaltrends.com/computing/chatgpts-bing-browsing-feature-disabled-over-paywall-flaw/; Cesar Cadenas, *ChatGPT pulls plug on Bing integration after people used it to bypass paywalls,* TECH RADAR (July 5, 2023), https://www.techradar.com/computing/artificial-intelligence/chatgpt-pulls-plug-on-bing-integration-after-people-used-it-to-bypass-paywalls.
[51] Arvind Narayanan (@random_walker), X, https://twitter.com/random_walker/status/1673090895929810945?lang=en (last visited Nov. 25, 2025).
[52] *HOLUPREDICTIONS*, REDDIT, https://www.reddit.com/r/ChatGPT/comments/14j8q1u/it_looks_like_you_can_use_chatgpt_to_bypass/?utm_source=embedv2&utm_medium=post_embed&utm_content=action_bar&embed_host_url=https://www.pcmag.com/news/browse-with-bing-disabled-on-chatgpt-plus-because-it-bypassed-paywalls (last visited Nov. 25, 2025).

on AI-generated "pink-slime" news sites.[53] The Defendants were aware of the risk of such use of the GPT models to create unauthorized copies and derivatives of newspaper content,[54] and upon information and belief, were aware or should have been aware of the actual use of the GPT models to replicate such material.

115. Indeed, as further evidence that OpenAI knew or reasonably should have known that end users use its GPT models to reproduce copyrighted content, OpenAI's Custom GPT Store contains numerous Custom GPTs specifically designed to circumvent the Publishers' paywalls despite OpenAI's representation that it "set up new systems to help review GPTs against [OpenAI's] usage policies" and that it "continue[s] to monitor and learn how people use GPTs".[55] For illustrative examples, OpenAI's store includes a "Bypass Paywall" Custom GPT, designed to "[a]ccess paywalled articles easily" and a "News Summarizer" Custom GPT that encourages users to "save on subscription costs" and "skip paywalls just using the link text or URL."

[53] Jack Brewster, *How I Built an AI-Powered, Self-Running Propaganda Machine for $105*, WALL STREET JOURNAL (Apr. 12, 2024), https://www.wsj.com/politics/how-i-built-an-ai-powered-self-running-propaganda-machine-for-105-e9888705; Jack Brewster et al., *The Year AI Supercharged Misinformation: NewsGuard's 2023 in Review*, NEWSGUARD (Dec. 27, 2023), https://www.newsguardtech.com/misinformation-monitor/december-2023/.
[54] Sébastien Bubeck et al., *Sparks of Artificial General Intelligence: Early Experiments With Gpt-4* (2023), https://arxiv.org/pdf/2303.12712.pdf.
[55] *Introducing GPTs,* OPENAI (Nov. 6, 2023), https://openai.com/blog/introducing-gpts.

Bypass Paywall (by Paywall.vip)

By Niklas D. Palladini

Access paywalled articles easily. Share your URL and get a Paywall.vip bypass link instantly.

Share an article URL - I'll give you the bypass

News Summarizer Ace

By MR MW KAHN

Articles, Videos and Podcasts: Concise briefs from any URL, paywall or not, with our GPT! Select word length and language preference effortlessly.

Productivity	300+
Category	Conversations

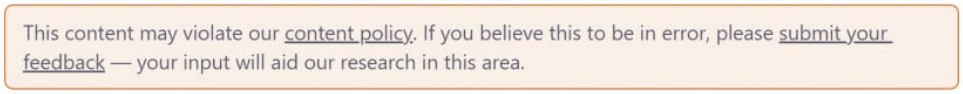
Conversation Starters

- Skip paywalls by just using the link text or URL.
- Bypass tedious copy-past or video watch or listen.
- Save on subscription costs.
- Our GPT guides your reading, watching, listening choices.

116. OpenAI approved and continues to monitor and support these Custom GPTs on its platform notwithstanding its representation that it "led the AI industry in providing a simple opt-out process for publishers … to prevent [its] tools from accessing their sites."[56]

117. Moreover, such Custom GPTs are easily accessible due to the shared nature of OpenAI's links. For example, one may locate Custom GPTs capable of summarizing news (even if paywalled) by searching on Google for: "site:chatgpt.com 'news summarizer'".

118. Third, on information and belief, Defendants have the ability to monitor users that infringe the rights of copyright owners such as the Publishers. For example, in at least some instances where ChatGPT detects that a user's query seeks to elicit output violating the OpenAI content policy, which requires that users "comply with all applicable laws,"[57] instead of providing the requested output, ChatGPT will sometimes provide a message to the user stating, "this content may violate our content policy."



119. Not only are Defendants capable of monitoring infringing outputs from their GPT-based products, Defendants have the ability to terminate user accounts that request and elicit copyrighted content from the Publishers and other rights holders. ChatGPT's Terms of Use provide[58]:

> We reserve the right to suspend or terminate your access to our Services or delete your account if we determine:

[56] *OpenAI and Journalism,* OPENAI (Jan. 8, 2024), https://openai.com/blog/openai-and-journalism.
[57] *Terms of Use,* OPENAI, https://openai.com/policies/terms-of-use/ (last visited November 26, 2025).
[58] *Id.*

- You breached these Terms or our Usage Policies.
- We must do so to comply with the law.
- Your use of our Services could cause risk or harm to OpenAI, our users, or anyone else.

120. Similarly, Microsoft's Terms of Use provide[59]:

> **CODE OF CONDUCT**
> **Don't infringe the rights of others**. Don't use Copilot to infringe on other people's legal rights, including their intellectual property and publicity rights.
>
> **OUR DECISIONS ABOUT COPILOT**
> We may choose to limit or stop offering or supporting Copilot or any feature within Copilot at any time and for any reason.
>
> **Unless prohibited by law, we may limit, suspend, or permanently revoke your access to or use of Copilot (and potentially all other Services) in our sole discretion, at any time and without notice**. Some of the reasons we might do this, for example, is if you breach these Terms or violate the Code of Conduct, if we suspect you're engaged in fraudulent or illegal activity, or if your Microsoft Account or the account you use to log in to Copilot is suspended or closed.

E. Defendants' Removal of Copyright Management Information from the Publishers' Works

121. The Publishers convey copyright management information ("CMI") with their copyrighted works. Each Publisher conveys authors' names, titles, and the Publishers' names with their works. For example, the following byline appears below the Virginian-Pilot article entitled title "Here's how Hampton Roads school board pay compares across the region":

[59] *Microsoft Copilot Terms of Use,* MICROSOFT, https://www.microsoft.com/en-us/microsoft-copilot/for-individuals/termsofuse (last visited Nov. 25, 2025).

By TREVOR METCALFE | trevor.metcalfe@pilotonline.com | Staff writer
PUBLISHED: November 23, 2025 at 2:01 PM EST

122. Each Publisher also conveys its terms and conditions and copyright notice in the webpage footer accompanying their works. For example, the following appears on the webpage that displays the article referenced in the preceding paragraph:

Subscriber Terms and Conditions Cookie Policy Cookie Preferences California Notice at Collection CA Notice of Financial Incentive Do Not Sell/Share My Personal Information
Copyright © 2025 The Virginian-Pilot

123. Defendants intentionally removed the Publishers' CMI from the Publishers' Works in the process of scraping the Publishers' Works from the Publishers' websites, storing the Publishers' Works in training datasets, using the Publishers' Works to train the GenAI products and/or in distributing unauthorized copies of the Publishers' Works through the operation of Defendants' GenAI products. The Defendants knew that by removing the Publishers' CMI, the CMI would not be retained within the GPT models and/or displayed when the GenAI products disseminate unauthorized copies of the Publishers' Works to end-users, and thereby would conceal the Defendants' own infringement as well as induce, enable, facilitate, or conceal end-users' infringement resulting from their operation of the Defendants' GenAI products.

124. The Defendants intentionally removed the Publishers' CMI from the Publishers' Works in one or more different ways. For example, in order to construct the datasets used to train their GenAI products, the Defendants used content extractors that, by design, removed the Publishers' CMI from the Publishers' Works. For example, OpenAI used the Dragnet[60] and

[60] Matthew E. Peters & Dan Lecocq, *Content Extraction Using Diverse Feature Sets*, WWW '13 COMPANION (May 2013).

Newspaper[61] content extractors[62] in creating the WebText dataset, which intentionally removed the Publishers' CMI from the Publishers' Works scraped from their website. Upon information and belief, Defendants used these two extraction methods, rather than one method, to create redundancies in case one method experienced a bug or did not work properly in a given case. Applying two methods rather than one would lead to a training set that is more consistent in the kind of content it contains, which is desirable from a training perspective. The abstract of the paper describing the Dragnet content extractor describes that copyright notices are removed as part of the process of extracting the text content of a website: "The goal of content extraction or boilerplate detection is to separate the main content from navigation chrome, advertising blocks, copyright notices and the like in web pages."[63] Likewise, upon information and belief, the Newspaper content extractor operates according to instructions to separate and extract the article text on the Publishers' webpages while removing the Publishers' CMI, including the Publishers' CMI located in the footer of the webpages, which includes the Publishers' terms and conditions and copyright notices.

125. Dragnet's algorithms are designed to "separate the main article content" from other parts of the website, including "footers" and "copyright notices," and allow the extractor to make further copies only of the "main article content." Dragnet is also unable to extract author and title information from the header or byline, and extracts it only if it happens to be separately contained in the main article content. Put differently, copies of news articles made by Dragnet are designed *not* to contain author, title, copyright notices, and footers, and do not contain such information unless it happens to be contained in the main article content.

[61] *codelucas/newspaper,* GITHUB, https://github.com/codelucas/newspaper (last visited Nov. 24, 2025).
[62] *Language Models Are Unsupervised Multitask Learners 3*, *supra* n. 27.
[63] *Content Extraction Using Diverse Feature Sets*, *supra* n. 60.

126. Like Dragnet, the Newspaper algorithms are incapable of extracting copyright notices and footers. Further, a user of Newspaper has the choice to extract or not extract author and title information. On information and belief, the OpenAI Defendants chose not to extract author and title information because they desired consistency with the Dragnet extractions, and Dragnet is typically unable to extract author and title information.

127. In applying the Dragnet and Newspaper algorithms while assembling the WebText dataset, the OpenAI Defendants removed Plaintiff's author, title, copyright notice, and terms of use information, the latter of which is contained in the footers of Plaintiff's websites.

128. Upon information and belief, the OpenAI Defendants, when using Dragnet and Newspaper, first download and save the relevant webpage before extracting data from it. This is at least because, when they use Dragnet and Newspaper, they likely anticipate a possible future need to regenerate the dataset (e.g., if the dataset becomes corrupted), and it is cheaper to save a copy than it is to recrawl all the data.

129. Because, by the time of its scraping, Dragnet and Newspaper were publicly known to remove author, title, copyright notices, and footers, and given that OpenAI employs highly skilled data scientists who would know how Dragnet and Newspaper work, the OpenAI Defendants intentionally and knowingly removed this copyright management information while assembling WebText.

130. The absence of author, title, copyright notice, and terms of use information from the copies of Publishers' Works generated by applying the Dragnet and Newspaper codes—codes that OpenAI has admitted to have intentionally used when assembling WebText—further corroborates that Defendants intentionally removed author, title, copyright notice, and terms of use information from Publishers' Works.

131. Upon information and belief, OpenAI has continued to use the same or similar Dragnet and Newspaper text extraction methods when creating training sets for every version of ChatGPT since GPT-2. This is at least because OpenAI has admitted to using these methods for GPT-2 and has neither publicly disclaimed their use for later version of ChatGPT nor publicly claimed to have used any other text extraction methods for those later versions.

132. Moreover, the Defendants distributed the Publishers' Works as output of the GenAI products knowing that the CMI originally conveyed with the Publishers' Works was removed without the Publishers' permission. The Defendants did so knowing that such distribution would induce, enable, facilitate, or conceal the Defendants' infringement or the infringement by an end-user of the GenAI Products.

133. Moreover, Defendants removed the Publishers' CMI with the intent to allow end-users to claim as their own the Publishers' Works output from the GenAI products. For example, OpenAI's terms of use provide that end-users "own the Output" notwithstanding the fact that the output contains reproductions of the Publishers' Works.[64]

134. Accordingly, Defendants knew or should have known that removing the Publishers' CMI from the Publishers' Works and outputting the Publishers' Works without the CMI wrongfully implied that Defendants had permission to use the Publishers' Works, thus concealing their own infringement. Defendants also knew or should have known that removing the Publishers' CMI in this manner would induce, enable, conceal, or facilitate infringement by end-users of the GenAI products.

[64] *Terms of Use, supra* n. 57.

F. Profit to Defendants

135. Each Defendant has greatly benefited from its wrongful conduct in multiple ways – and most notably, the financial benefit to Defendants from their unlawful conduct has been monumental.

136. According to OpenAI, ChatGPT has 700-800 million weekly active users.[65] A subset of those users pay for ChatGPT Plus, for which OpenAI charges users $20 per month.[66] When announcing the release of ChatGPT Enterprise, a subscription-based high-capability GPT-4 application targeted at corporate clients, in August 2023, OpenAI claimed that teams in "over 80% of Fortune 500 companies" were using its products.[67]

137. As of March 2025, OpenAI was on pace to generate more than $12.7 billion in revenue in 2025—more than $1 billion in revenue per month.[68] As noted above, in 2025 it was reported that OpenAI is valued at $500 billion.[69]

138. The value of Microsoft's investments in OpenAI has substantially increased over time. Microsoft initially invested $1 billion in OpenAI in 2019, an investment that one publication has said may be "one of the shrewdest bets in tech history."[70] In 2021, OpenAI was valued at $14 billion; just two years later, in early 2023, it was valued at approximately $29 billion.[71] Microsoft

---

[65] Rebecca Bellan, "Sam Altman says ChatGPT has hit 800 million active users," TECHCRUNCH (Oct. 6, 2025); https://openai.com/index/how-people-are-using-chatgpt/ (last visited November 20, 2025).
[66] OpenAI, *Introducing ChatGPT Plus*, OPENAI (Feb. 1, 2023), https://openai.com/blog/chatgpt-plus.
[67] *Introducing ChatGPT Enterprise*, *supra* n. 14.
[68] *https://www.forbes.com/sites/martineparis/2025/03/31/top-ai-stars-of-sxsw-2025-from-waymo-uber-to-woolly-mouse/*
[69] *https://openai.com/index/march-funding-updates/*
[70] Hasan Chowdhury, *Microsoft's Investment into ChatGPT's Creator May Be the Smartest $1 Billion Ever Spent*, BUSINESS INSIDER (Jan. 6, 2023), https://www.businessinsider.com/microsoft-openai-investment-the-smartest-1-billion-ever-spent-2023-1.
[71] Phil Rosen, *ChatGPT's Creator OpenAI Has Doubled in Value Since 2021 as the Language Bot Goes Viral and Microsoft Pours in $10 Billion*, BUSINESS INSIDER (Jan. 24, 2023), https://markets.businessinsider.com/news/stocks/chatgpt-openai-valuation-bot-microsoft-language-google-tech-stock-funding-2023-1#:~:text=In%202021%2C%20the%20tech%20firm,%2410%20billion%20investment%20in%20OpenAI.

eventually increased its investment in OpenAI to a reported $13 billion. In February 2024 it was reported that OpenAI is valued at $80 billion or more.[72] As noted above, recent reports based on OpenAI's recent recapitalization place OpenAI's value at $500 billion.

139. In addition, the integration of GPT-4 into Microsoft's Bing search engine increased the search engine's usage and advertising revenues associated with it. Just a few weeks after Bing Chat (now Copilot) was launched, Bing reached 100 million daily users for the first time in its 14-year history.[73] A subset of those users pay for Copilot Pro, for which Microsoft charges $20 per month.[74] Similarly, page visits on Bing rose 15.8% in the first approximately six weeks after Bing Chat was unveiled.[75] According to recent reports, Microsoft's total revenue increased 17% to $61.86 billion during the first quarter of 2024 due in large part to its AI related products and services.[76]

140. Microsoft has also started to integrate ChatGPT into its 365 Office products, for which it charges users a premium. Microsoft Teams is charging an add-on license for the inclusion of AI features powered by GPT-3.5.[77] Microsoft is also charging $30 per user per month for Microsoft 365 Copilot, a tool powered by GPT-4 that is designed to assist with the creation of

[72] *OpenAI valued at $80 billion after deal, NYT reports,* REUTERS (Feb. 16, 2024), https://www.reuters.com/technology/openai-valued-80-billion-after-deal-nyt-reports-2024-02-16/.
[73] Tom Warren, *Microsoft Bing Hits 100 Million Active Users in Bid to Grab Share from Google*, THE VERGE (Mar. 9, 2023), https://www.theverge.com/2023/3/9/23631912/microsoft-bing-100-million-daily-active-users-milestone.
[74] As of October 1, 2025, Copilot Pro has been rolled into a Microsoft 365 Premium feature, which is still offered at $20/month. *See* Yusuf Mehdi, *Meet Microsoft 365 Premium: Your AI and productivity powerhouse*, MICROSOFT (Oct. 1, 2025), https://www.microsoft.com/en-us/microsoft-365/blog/2025/10/01/meet-microsoft-365-premium-your-ai-and-productivity-powerhouse.
[75] Akash Sriram & Chavi Mehta, *OpenAI Tech Gives Microsoft's Bing a Boost in Search Battle with Google*, REUTERS (Mar. 22, 2023), https://www.reuters.com/technology/openai-tech-gives-microsofts-bing-boost-search-battle-with-google-2023-03-22/.
[76] Blake Montgomery, *Microsoft's heavy bet on AI pays off as it beats expectations in latest quarter,* THE GUARDIAN (Apr. 25, 2024), https://www.theguardian.com/technology/2024/apr/25/microsoft-earnings.
[77] Tom Warren, *Microsoft Launches Teams Premium with Features Powered by OpenAI*, THE VERGE (Feb. 2, 2023), https://www.theverge.com/2023/2/2/23582610/microsoft-teams-premium-openai-gpt-features.

documents, emails, presentations, and more.[78] That $30 per user per month premium will nearly double the cost for businesses subscribed to Microsoft 365 E3, and will nearly triple the cost for those subscribed to Microsoft 365 Business Standard.[79]

141. None of this would have been possible had Defendants not used copyrighted content, including the Publishers' Works, to develop their models and products.

G. Harm to the Publishers

142. Defendants' unlawful conduct has caused, and will continue to cause, substantial harm to the Publishers. The Publishers have spent of millions of dollars and uncountable hours to gather news and information for the reports they provide to their readers. Those readers support the Publishers' businesses by purchasing and renewing subscriptions to Publishers' products, which include print newspapers, paywalled websites, mobile applications, and premium newsletters.

143. Defendants' illegal and unauthorized use of the Publishers' Works to train GenAI models has enabled Defendants to create products that provide news and information plagiarized and stolen from the Publishers, often without any reference to the Publishers' original work or reporting. Such activity fundamentally undermines the Publishers' business model, which is critically dependent on subscription revenues to fund journalism, because it results in substitutive products for which Defendants seek to charge their customers for access, siphoning off existing and potential customers through their unlawful and uncompensated use of the Publishers' own products.

---

[78] Tom Warren, *Microsoft Announces Copilot: The AI-Powered Future of Office Documents*, THE VERGE (Mar. 16, 2023), https://www.theverge.com/2023/3/16/23642833/microsoft-365-ai-copilot-word-outlook-teams; Tom Warren, *Microsoft Puts a Steep Price on Copilot, Its AI-Powered Future of Office Docum*ents, THE VERGE (July 18, 2023), https://www.theverge.com/2023/7/18/23798627/microsoft-365-copilot-price-commercial-enterprise.

[79] *Microsoft Announces Copilot: The AI-Powered Future of Office Documents*, *supra* n. 78.

144. If people are able to access the Publishers' Works through the Defendants' GenAI products without paying the Publishers or subscribing to their products, they are likely to do so and less likely to visit Publishers' websites or subscribe to Publishers' products.

145. Additionally, the Publishers' business includes licensing content to other media entities and publishers, but with clear guidelines as to how the Publishers' Works can be displayed and used. In these cases, the Publishers are rightfully paid for the use of their work product. The Defendants' illegal use of the Publishers' Works undermines these arrangements as well by providing the Publishers' Works directly to readers.

146. In these ways, Defendants' illegal use and distribution of the Publishers' Works damages the Publishers' ability to attract and retain paying subscribers while at the same time eroding the Publishers' ability to engage in and maintain licensing agreements with other publishers of news and information.

147. Defendants' practice of generating misinformation and then wrongfully attributing it to the Publishers damages the Publishers' brands, credibility and reputation, and undermines the Publishers' investment, goodwill and reputation.

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

On Behalf of the Publishers Against All Defendants

148. The Publishers incorporate by reference and reallege the preceding allegations as though fully set forth herein.

149. The Publishers are the owners of the registered copyrights listed in Exhibits A-I. It is and has been since at least the mid-2000's the business practice of the Publishers to publish in electronic format on the Publishers' respective websites every article that also appears in that newspaper's print edition. It has additionally been the business practice of some of the Publishers

to publish in electronic format on those Publishers' respective websites articles from older print editions of those newspapers. All of the copyright registrations containing the Publishers' Works that are asserted as infringed are reflected on Exhibits A-I. (The Publishers reserve the right to revise or supplement these exhibits if it becomes clear during discovery that additional registered works of the Publishers were also copied by Plaintiffs.) The electronic versions of the articles are substantially the same as their print-edition counterparts.

150. The electronic versions of the articles found in each of the Publishers' Works set forth in Exhibits A-I were copied to train Defendants' GPT models and, in many cases, have been distributed by and encoded within Defendants' GPT models. As the owners of the registered copyrights in the literary works that Defendants have copied, distributed, and encoded in Defendants' GPT models, the Publishers hold the exclusive rights to those works under 17 U.S.C. § 106.

151. By illegally building training datasets containing the Publishers' Works, including by scraping copies of the Publishers' Works from the Publishers' websites and reproducing these works from third-party datasets, Microsoft and the OpenAI Defendants have directly infringed the Publishers' exclusive rights in their copyrighted works.

152. By illegally storing, processing, and reproducing the training datasets containing the Publishers' Works to train the GPT models on Microsoft's supercomputing platform, Microsoft and the OpenAI Defendants have jointly directly infringed the Publishers' exclusive rights in their copyrighted works.

153. On information and belief, by storing, processing, and reproducing the GPT models trained on the Publishers' Works, which GPT models themselves have memorized, on Microsoft's

supercomputing platform, Microsoft and the OpenAI Defendants have jointly directly infringed the Publishers' exclusive rights in their copyrighted works.

154. By disseminating generative output containing copies and derivatives of the Publishers' Works through the ChatGPT offerings, the OpenAI Defendants have directly infringed the Publishers' exclusive rights in their copyrighted works.

155. By disseminating generative output containing copies and derivatives of the Publishers' Works through the Copilot (formerly known as Bing Chat) offerings, Microsoft has directly infringed the Publishers' exclusive rights in their copyrighted works.

156. On information and belief, Defendants' infringing conduct alleged herein was and continues to be willful and carried out with full knowledge of the Publishers' rights in their Works. As a direct result of their conduct, Defendants have wrongfully profited from copyrighted works that they do not own.

157. By and through the actions alleged above, Defendants have infringed and will continue to infringe the Publishers' copyrights.

158. As a direct and proximate result of Defendants' infringing conduct alleged herein, the Publishers have sustained and will continue to sustain substantial, immediate, and irreparable injury for which there is no adequate remedy at law. Unless Defendants' infringing conduct is enjoined by this Court, Defendants have demonstrated an intent to continue to infringe the Publishers' Works. The Publishers therefore are entitled to permanent injunctive relief restraining and enjoining Defendants' ongoing infringing conduct.

159. The Publishers are further entitled to recover statutory damages, actual damages, restitution of profits, attorneys' fees, and other remedies provided by law.

COUNT II: Vicarious Copyright Infringement

On Behalf of the Publishers Against Microsoft, OpenAI, Inc., OpenAI, GP, OpenAI LP, OAI Corporation, LLC, OpenAI Holdings, LLC, OpenAI Global, LLC, and OpenAI Global PBC

160. The Publishers incorporate by reference and reallege the preceding allegations as though fully set forth herein.

161. Microsoft controlled, directed, and profited from the infringement perpetrated by the OpenAI Defendants. Microsoft controls and directs the supercomputing platform used to store, process, and reproduce the training datasets containing the Publishers' Works, the GPT models, and OpenAI's ChatGPT offerings. Microsoft profited from the infringement perpetrated by the OpenAI defendants by incorporating the infringing GPT models trained on the Publishers' Works into its own product offerings, including Copilot (formerly known as Bing Chat).

162. Defendants OpenAI, Inc.; OpenAI, GP; OAI Corporation, LLC; OpenAI Holdings, LLC; and Microsoft controlled, directed, and profited from the infringement perpetrated by Defendants OpenAI LP; OpenAI Global, LLC; OpenAI OpCo, LLC; OpenAI, LLC; and OpenAI Global PBC including the reproduction and distribution of the Publishers' Works.

163. Defendants OpenAI Global, LLC and OpenAI LP directed, controlled, and profited from the infringement perpetrated by Defendants OpenAI OpCo, LLC and OpenAI, LLC, including the reproduction and distribution of the Publishers' Works.

164. Defendants OpenAI, Inc.; OpenAI LP; OAI Corporation, LLC; OpenAI Holdings, LLC; OpenAI Global, LLC; and Microsoft are vicariously liable for copyright infringement.

COUNT III: Contributory Copyright Infringement

On Behalf of the Publishers Against Microsoft

165. The Publishers incorporate by reference and reallege the preceding allegations as though fully set forth herein.

166. Microsoft materially contributed to and directly assisted in the direct infringement attributable to the OpenAI Defendants.

167. Microsoft provided the supercomputing infrastructure and directly assisted the OpenAI Defendants in: (i) building training datasets containing copies of the Publishers' Works; (ii) storing, processing, and reproducing the training datasets containing copies of the Publishers' Works used to train the GPT models; and (iii) providing the computing resources to host, operate, and commercialize the GPT models and GenAI products.

168. Microsoft knew or had reason to know of the direct infringement perpetrated by the OpenAI Defendants because Microsoft and OpenAI's partnership extends to the development, commercialization, and monetization of the OpenAI Defendants' GPT-based products. Microsoft was fully aware of the capabilities of OpenAI's GPT-based products.

COUNT IV: Contributory Copyright Infringement

On Behalf of the Publishers Against All Defendants

169. The Publishers incorporate by reference and reallege the preceding allegations as though fully set forth herein.

170. In the alternative, to the extent Defendants argue that an end-user may be liable as a direct infringer based on the output of the GPT-based products, Defendants materially contributed to and directly assisted with the direct infringement perpetrated by end-users of the GPT-based products by way of: (i) jointly-developing LLM models capable of distributing unlicensed copies of the Publishers' Works to end-users; (ii) building and training the GPT LLMs using the Publishers' Works; and (iii) deciding what content is actually outputted by the GenAI products, such as grounding output in the Publishers' Works through retrieval augmented generation, fine-tuning the models for desired outcomes, and/or selecting and weighting the parameters of the GPT LLMs.

171. On information and belief Defendants continue to maintain and support user accounts that are used to retrieve infringing output from Defendants' GPT-based products.

172. Defendants had either actual knowledge or constructive knowledge of the direct infringement by end-users or were willfully blind to the direct infringement of end-users because: (i) Defendants undertake extensive efforts in developing, testing, and troubleshooting their LLM models and GPT-based products; (ii) Defendants programmed their systems to flag infringing outputs and prompts seeking infringing output; (iii) Defendants have been repeatedly informed of instances where their GPT-based products output infringing content to users and the capability of their GPT-based models to produce infringing output has been the subject of public conversation; (iv) Defendants are aware that at least some users use their GPT-based products for the purpose of accessing copyrighted works; and (v) Defendants have publicly recognized and admitted that their GPT-based products are capable of distributing unlicensed copies of copyrighted works and derivatives thereof.

COUNT V: Digital Millennium Copyright Act – Removal of Copyright Management Information (17 U.S.C. § 1202)

On Behalf of the Publishers Against All Defendants

173. The Publishers incorporate by reference and reallege the preceding allegations as though fully set forth herein.

174. The Publishers included one or more forms of copyright-management information in each of the Publishers' Works, including: a copyright notice, authors' names, publisher's name, title and other identifying information, terms and conditions of use, and identifying numbers or symbols referring to the copyright-management information.

175. Without the Publishers' authority, Defendants copied the Publishers' Works and used them as training data for their GenAI models.

176. On information and belief, Defendants removed the Publishers' copyright-management information in building the training datasets containing copies of the Publishers' Works, including removing the Publishers' copyright-management information from the Publishers' Works scraped directly from the Publishers' websites and removing the Publishers' copyright-management information from the Publishers' Works reproduced from third-party datasets.

177. On information and belief, OpenAI removed Publishers' copyright-management information through generating synthetic search results, including removing Publishers' copyright-management information when scraping the Publishers' Works from Publishers' websites and generating copies or derivatives of the Publishers' Works as the output of ChatGPT offerings.

178. On information and belief, Microsoft and OpenAI removed the Publishers' copyright-management information through generating synthetic search results, including removing the Publishers' copyright-management information when scraping the Publishers' Works from the Publishers' websites and generating copies or derivatives of the Publishers' Works as output for the Browse with Bing and Copilot (formerly known as Bing Chat) offerings.

179. Microsoft and OpenAI removed the Publishers' copyright-management information in generating outputs from the GPT models containing copies or derivatives of the Publishers' Works.

180. By design, Defendants' GPT-based products do not preserve any copyright-management information, and the outputs of Defendants' GPT models removed any copyright notices, titles, and identifying information, despite the fact that those outputs were often verbatim reproductions of the Publishers' Works. Therefore, Defendants intentionally removed copyright-management information from the Publishers' Works in violation of 17 U.S.C. § 1202(b)(1).

181. Defendants' removal or alteration of the Publishers' copyright-management information has been done knowingly and with the intent to induce, enable, facilitate, or conceal Defendants' or end-users' infringement of the Publishers' copyrights.

182. Defendants knew or had reasonable grounds to know that their removal of copyright-management information would facilitate copyright infringement by concealing the fact that the GPT models are infringing copyrighted works and that outputs from the GPT models are infringing copies and derivative works.

183. The Publishers have been injured by Defendants' removal of copyright-management information. The Publishers are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law, including full costs and attorneys' fees.

PRAYER FOR RELIEF

WHEREFORE, the Publishers demand judgment against each Defendant as follows:

184. Awarding the Publishers statutory damages, compensatory damages, restitution, disgorgement, and any other relief that may be permitted by law or equity, in an amount well in excess of $10 billion;

185. Awarding the Publishers statutory damages, compensatory damages, restitution, disgorgement, and any other relief that may be permitted by law or equity in an amount no less than $25,000 for each and every work from which Defendants removed the Publishers' copyright management information;

186. Declaring that Defendants' past, present, and future use of the Publishers' Works is unlawful, and infringes Publishers' copyrights, trademarks, and other lawful rights and privileges;

187. Permanently enjoining Defendants from the unlawful, unfair, and infringing conduct alleged herein;

188. Ordering destruction under 17 U.S.C. § 503(b) of all GPT or other LLM models and training sets that incorporate the Publishers' Works;

189. An award of costs, expenses, and attorneys' fees as permitted by law; and

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

DEMAND FOR JURY TRIAL

The Publishers hereby demand a jury trial for all claims so triable.

Dated: November 26, 2025

By: /s/ *Steven Lieberman*

Steven Lieberman (SL8687)
Jennifer B. Maisel (5096995)
Robert Parker *(pro hac vice forthcoming)*
Jenny L. Colgate *(pro hac vice forthcoming)*
Kristen J. Logan *(pro hac vice forthcoming)*
Bryan B. Thompson (6004147)
Alexandra Hughes (*pro hac vice forthcoming*)
ROTHWELL, FIGG, ERNST & MANBECK, P.C.
901 New York Avenue, N.W., Suite 900 East
Washington, DC 20001
Telephone: (202) 783-6040
Facsimile: (202) 783-6031
slieberman@rothwellfigg.com
jmaisel@rothwellfigg.com
rparker@rothwellfigg.com
jcolgate@rothwellfigg.com
klogan@rothwellfigg.com
bthompson@rothwellfigg.com
ahughes@rothwellfigg.com

Attorneys for Plaintiffs