UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

RICHNER COMMUNICATIONS, INC.; AIM MEDIA INDIANA OPERATING, LLC; AIM MEDIA MIDWEST OPERATING, LLC; AIM MEDIA TEXAS OPERATING, LLC; AMNEWS CORP. d/b/a THE NEW YORK AMSTERDAM NEWS; ARKANSAS DEMOCRAT-GAZETTE, INC.; CASA GRANDE VALLEY NEWSPAPERS INC.; CHERRYROAD MEDIA INC.; COMMUNITY IMPACT NEWSPAPER CO.; CONCORD PUBLISHING HOUSE, INC.; D-R MEDIA AND INVESTMENTS, LLC; D.A. PUBLISHING, LC; EAGLE URBAN MEDIA LLC; EL CREPUSCULO, INC.; H.S. GERE & SONS, INC.; IOWA INFORMATION INC.; LAKEWAY PUBLISHERS, INC.; THE NEW MEXICAN INC.; NEWSPAPERS OF MASSACHUSETTS, INC.; NEWSPAPERS OF NEW ENGLAND, INC.; NEWSPAPERS OF NEW HAMPSHIRE, INC.; NORTH COUNTRY THIS WEEK, INC.; THE OGDEN NEWSPAPERS, INC.; PATCHOGUE ADVANCE, INC.; RUST PUBLISHING ID, LC; RUST PUBLISHING MOKS, LC; RUST PUBLISHING NE, LC; RYE MEDIA PARTNERS LLC; SENTINEL MEDIA CO., INC.; SHAW FAMILY HOLDINGS, INC.; STRAUS MEDIA-MANHATTAN, LLC; STRAUS NEWSPAPERS, INC.; WEHCO NEWSPAPERS, INC.; WHITE MOUNTAIN PUBLISHING LLC; and WICK COMMUNICATIONS,

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. 26-cv-5320

COMPLAINT

JURY TRIAL DEMANDED

Plaintiffs Richner Communications, Inc.; AIM Media Indiana Operating, LLC; AIM Media Midwest Operating, LLC; AIM Media Texas Operating, LLC; AmNews Corp. d/b/a/ The New York Amsterdam News; Arkansas Democrat-Gazette, Inc.; Casa Grande Valley Newspapers Inc.; CherryRoad Media Inc.; Community Impact Newspaper Co.; Concord Publishing House, Inc.; D-R Media and Investments, LLC; D.A. Publishing, LC; Eagle Urban Media LLC; El Crepusculo, Inc.; H.S. Gere & Sons, Inc.; Iowa Information Inc.; Lakeway Publishers, Inc.; The New Mexican Inc.; Newspapers of Massachusetts, Inc.; Newspapers of New England, Inc.; Newspapers of New Hampshire, Inc.; North Country This Week, Inc.; The Ogden Newspapers, Inc.; Patchogue Advance, Inc.; Rust Publishing ID, LC; Rust Publishing MOKS, LC; Rust Publishing NE, LC; Rye Media Partners LLC; Sentinel Media Co., Inc.; Shaw Family Holdings, Inc.; Straus Media-Manhattan, LLC; Straus Newspapers, Inc.; WEHCO Newspapers, Inc.; White Mountain Publishing LLC; and Wick Communications (collectively the "Publishers"), by their attorneys Platkin LLP, 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. This lawsuit arises from Defendants' systematic and willful theft of hundreds of thousands of copyrighted articles belonging to the Publishers, who collectively own and operate nearly 400 local and regional newspaper outlets across the country, all of whom have spent decades—and in some cases over a century—investing in the journalists, editors, and infrastructure required to produce the trusted, original reporting on which their communities depend. Without

permission and without any compensation to the Publishers, Defendants scraped, copied, and ingested that content to build and commercialize their generative artificial intelligence ("GenAI") products, including ChatGPT and Microsoft Copilot. Those products have generated hundreds of billions of dollars (and counting) in market value for Defendants. Not a cent of it has gone to the Publishers whose work made it possible.

2. Using automated systems, Defendants systematically and secretly crawled the Publishers' websites—including content behind paywalls and other access restrictions—and copied the Publishers' articles, stories, and other original works onto their own servers without authorization. As part of that process, Defendants' systems stripped from the Publishers' works all copyright management information ("CMI") embedded in and associated with those works, such as author credits, publication names, copyright notices, and terms of use information, that establish ownership and signal that a work is protected. That CMI-stripping, an instrumental part of Defendants' ingestion pipeline, helped sever the link between the copied content and its rightful owners and authorizations. The scraped, stripped content was then used to train Defendants' large language models ("LLMs"), which have "memorized" that material and likely reproduced it, verbatim or near-verbatim, in response to user prompts for years. And because Defendants' models must be continuously updated with new material to remain current and commercially viable, these processes have been repeated over and over and over again.

3. That Defendants' conduct was willful is beyond dispute. OpenAI's founder, Sam Altman, acknowledged as much in testimony before the British House of Lords, conceding that it would be "impossible to train today's leading AI models without using copyrighted materials." Defendants made deliberate engineering choices to copy the Publishers' content and strip its CMI, knowing it would obscure the origins of the works they were taking and impair the Publishers'

ability to detect and prove the theft. And even as litigation from other publishers mounted, and as courts began to recognize the validity of these claims, Defendants pressed forward undeterred. Defendants' data scraping processes, their products' storage and reproduction of "memorized" content, and therefore their violations of the Copyright Act and the Digital Millennium Copyright Act, continue to this day.

4. On the backs of the Publishers, Defendants built some of the most valuable businesses in human history. OpenAI, once styled as a nonprofit, now commands a valuation worth nearly one trillion dollars. Microsoft's deployment of its Copilot product has added hundreds of billions of dollars to its market capitalization. These are not the fruits of Defendants' ingenuity alone. The Publishers' journalism was essential to the Defendants' explosive growth, and unless Defendants are held accountable for stealing, stripping, and misusing the Publishers' content, the AI boom Defendants orchestrated and benefit from will be a death knell for local journalism—which remains the most trusted news sources in America.

5. The Publishers are generally independently-owned newspaper companies. Many are family-owned small businesses. They are the lifeblood of the communities they serve. They send reporters to city council meetings and school board hearings. They investigate corruption and hold local officials accountable. They are the outlets that cover the latest high school football game, the new restaurant opening downtown, or the storm bearing down on the coast. They publish obituaries, job listings, and apartment notices. They convey to their readers everyday stories of local civic life that national outlets do not cover.

6. The Publishers have spent billions of dollars to sustain this work. Defendants helped themselves to all of it—without providing a cent of compensation.

7. The U.S. Constitution has, since the nation's founding, charged Congress with protecting authors' and publishers' exclusive rights in their work. Congress has exercised that authority to implement robust protections against copyright theft, including through enacting the Copyright Act and the Digital Millennium Copyright Act, and authorizing substantial penalties for willful violations of both. Defendants invoke those protections vigorously for their own products, shielding their code, their models, and their systems behind licenses, paywalls, and legal threats. In bringing this action, the Publishers seek to hold Defendants to the same standard they insist upon for themselves.

8. Novel as the technology at issue may be, this is not a case of first impression. The Publishers, along with other news publishers, authors, and other copyright holders across the country have brought these same claims against these same Defendants, and those cases have survived motions to dismiss largely intact. Defendants have chosen to continue their unlawful conduct rather than rectify it. This lawsuit seeks to hold Defendants fully accountable for every violation—past, present, and ongoing.

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.*, including the Digital Millenium Copyright Act amendments.

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 sale of Microsoft and OpenAI's Generative Pre-training Transformer ("GPT")-based products like ChatGPT, ChatGPT Enterprise, Copilot,

Azure OpenAI Service, Microsoft 365 Copilot, 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, and both Microsoft and OpenAI enjoy a substantial base of monthly active users of Bing Chat and ChatGPT in this District. OpenAI has licensed its GPT models to New York residents and companies headquartered in this District. For example, in 2023, OpenAI struck deals to license its GPT models to the Associated Press and Morgan Stanley, both companies headquartered in this District. Moreover, several of the Publishers' outlets operate within this District, including *The Riverdale Press*, *The New York Amsterdam News*, *Our Town*, *The West Side Spirit*, *The Chelsea News*, *The Warwick Advertiser*, *The Chronicle*, and *The Photo News*.

12. Defendants consented to or waived challenges to personal jurisdiction and venue in this District in several cases raising similar claims, including *The New York Times Company v. Microsoft Corporation*, 23-cv-11195; *Raw Story Media, Inc. v. OpenAI, Inc.*, No. 24-cv-01514; *The Intercept Media, Inc. v. OpenAI, Inc.*, No. 24-cv-01515; *Daily News v. Microsoft Corporation*,

No. 24-cv-03285; and *Authors Guild v. OpenAI, Inc.*, 23-cv-08292 (all cases filed in the Southern District of New York and transferred to Multidistrict Litigation No. 1:25-md-03143 in the Southern District).

THE PARTIES

13. Plaintiff Richner Communications, Inc. ("RCI") is incorporated and headquartered in New York. Its principal place of business is in Garden City, New York.

14. Plaintiff AIM Media Indiana Operating, LLC is incorporated in Delaware and headquartered in Indiana. Its principal place of business is in Columbus, Indiana.

15. Plaintiff AIM Media Midwest Operating, LLC is incorporated in Delaware and headquartered in Ohio. Its principal place of business is in Lima, Ohio.

16. Plaintiff AIM Media Texas Operating, LLC is incorporated in Delaware and headquartered in Texas. Its principal place of business is in McAllen, Texas.

17. Plaintiff AmNews Corp. d/b/a The New York Amsterdam News is incorporated and headquartered in New York. Its principal place of business is in New York, New York.

18. Plaintiff Arkansas Democrat-Gazette, Inc. is incorporated and headquartered in Arkansas. Its principal place of business is in Little Rock, Arkansas.

19. Plaintiff Casa Grande Valley Newspapers Inc. is incorporated and headquartered in Arizona. Its principal place of business is in Casa Grande, Arizona.

20. Plaintiff CherryRoad Media Inc. is incorporated in Florida and headquartered in New Jersey. Its principal place of business is in Parsippany, New Jersey.

21. Plaintiff Community Impact Newspaper Co. is incorporated and headquartered in Texas. Its principal place of business is in Pflugerville, Texas.

22. Plaintiff Concord Publishing House, Inc. is incorporated and headquartered in Missouri. Its principal place of business is Cape Girardeau, Missouri.

23. Plaintiff D-R Media and Investments, LLC is incorporated and headquartered in Florida. Its principal place of business is in Venice, Florida.

24. Plaintiff DA Publishing LC is incorporated and headquartered in Missouri. Its principal place of business is Cape Girardeau, Missouri.

25. Plaintiff Eagle Urban Media LLC is incorporated and headquartered in New York. Its principal place of business is in Brooklyn, New York.

26. Plaintiff El Crepusculo, Inc. is incorporated and headquartered in New Mexico. Its principal place of business is in Taos, New Mexico.

27. Plaintiff H.S. Gere & Sons, Inc. is incorporated in Delaware and headquartered in Massachusetts. Its principal place of business is in Northampton, Massachusetts.

28. Plaintiff Iowa Information Inc. is incorporated and headquartered in Iowa. Its principal place of business is in Sheldon, Iowa.

29. Plaintiff Lakeway Publishers, Inc. is incorporated and headquartered in Tennessee. Its principal place of business is in Morristown, Tennessee.

30. Plaintiff The New Mexican Inc. is incorporated and headquartered in New Mexico. Its principal place of business is in Santa Fe, New Mexico.

31. Plaintiff Newspapers of Massachusetts, Inc. is incorporated in Delaware and headquartered in Massachusetts. Its principal place of business is in Greenfield, Massachusetts.

32. Plaintiff Newspapers of New England, Inc. is incorporated in Delaware and headquartered in New Hampshire. Its principal place of business is in Concord and West Lebanon, New Hampshire.

33. Plaintiff Newspapers of New Hampshire, Inc. is incorporated in Delaware and headquartered in New Hampshire. Its principal place of business is in Concord and West Lebanon, New Hampshire.

34. Plaintiff North Country This Week, Inc. is incorporated and headquartered in New York. Its principal place of business is in Potsdam, New York.

35. Plaintiff The Ogden Newspapers, Inc. is incorporated and headquartered in West Virginia. Its principal place of business is in Wheeling, West Virginia.

36. Plaintiff Patchogue Advance, Inc. is incorporated and headquartered in New York. Its principal place of business is in Patchogue, New York.

37. Plaintiff Rust Publishing ID, LC is incorporated and headquartered in Missouri. Its principal place of business is in Cape Girardeau, Missouri.

38. Plaintiff Rust Publishing MOKS, LC is incorporated and headquartered in Missouri. Its principal place of business is in Cape Girardeau, Missouri.

39. Plaintiff Rust Publishing NE, LC is incorporated and headquartered in Missouri. Its principal place of business is in Cape Girardeau, Missouri.

40. Plaintiff Rye Media Partners LLC is incorporated and headquartered in New York. Its principal place of business is in Rye, New York.

41. Plaintiff Sentinel Media Co., Inc. is incorporated and headquartered in New York. Its principal place of business is in Rome, New York.

42. Plaintiff Shaw Family Holdings, Inc. is incorporated in Delaware and headquartered in Illinois. Its principal place of business is in Crystal Lake, Illinois.

43. Plaintiff Straus Media-Manhattan, LLC is incorporated and headquartered in New York. Its principal place of business is in Chester, New York.

44. Plaintiff Straus Newspapers, Inc. is incorporated and headquartered in New York. Its principal place of business is in Chester, New York.

45. Plaintiff WEHCO Newspapers, Inc. is incorporated and headquartered in Arkansas. Its principal place of business is in Little Rock, Arkansas.

46. Plaintiff White Mountain Publishing LLC is incorporated and headquartered in Arizona. Its principal place of business is in Show Low, Arizona.

47. Plaintiff Wick Communications is incorporated and headquartered in Arizona. Its principal place of business is in Sierra Vista, Arizona.

48. Defendant 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.

49. Microsoft has described its relationship with the OpenAI Defendants as a "partnership."[1] This partnership has included contributing and operating the cloud computing services used to copy the Publishers' works and train the OpenAI Defendants' GenAI models. It has also included, upon 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 to users of its

[1] *See* Microsoft Corp., *The Next Chapter of the Microsoft–OpenAI Partnership*, Microsoft Blogs (Oct. 28, 2025), https://blogs.microsoft.com/blog/2025/10/28/the-next-chapter-of-the-microsoft-openai-partnership. All hyperlinks last visited June 23, 2026.

products and online services. And that has proven lucrative: Microsoft's 2025 end-of-year investor report touted over $75 billion in revenue from its Azure cloud computing platform.[2]

50. The OpenAI Defendants consist of a web of interrelated entities.

51. Defendant OpenAI, Inc. is a Delaware nonprofit corporation with a principal place of business located at 3180 18th Street, 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 was directly involved in perpetrating the mass infringement and other unlawful conduct alleged here.

52. Defendant OpenAI LP is a Delaware limited partnership with its principal place of business located at 3180 18th Street, 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.

53. Defendant OpenAI GP, LLC is a Delaware limited liability company with a principal place of business located at 3180 18th Street, 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.

54. 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

[2] Microsoft Corp, 2025 Annual Report, https://www.microsoft.com/investor/reports/ar25/index.html.

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. and Microsoft Corporation through OpenAI Global, LLC and OpenAI OpCo, LLC.

55. 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.

56. 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. Microsoft Corporation has a minority stake in OpenAI Global LLC, and OpenAI, Inc. has a majority stake in OpenAI Global, LLC, indirectly through OpenAI Holdings, LLC and OAI Corporation, LLC. OpenAI Global, LLC was and is involved in the unlawful conduct alleged herein through its ownership, control, and direction of OpenAI, LLC.

57. 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.

58. 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.

59. On or around October 28, 2025, OpenAI announced an "updated structure."[3] After this announcement, the non-profit entity formerly known as OpenAI, Inc. is "now the OpenAI Foundation," and the "for-profit" business established in 2019 is "now a public benefit corporation, called OpenAI Group PBC."[4] OpenAI's website notes that the "OpenAI Foundation continues to control the OpenAI Group."[5]

60. Upon information and belief, the OpenAI Foundation is a Delaware foundation with a principal office in San Francisco, California and OpenAI Group PBC is a Delaware public benefit corporation also with a principal place of business in San Francisco, California.

FACTUAL ALLEGATIONS

A. The Publishers Are Trusted Producers of Regional and Local News.

61. The Publishers are among the most trusted sources of news in the communities they serve.[6] As producers of local news, the Publishers occupy a unique and essential role in American civic life. Unlike national outlets, they cover school board meetings, municipal elections, community events, and other local issues that directly shape people's daily lives. Readers rely on them because they are close to the ground, accountable to their neighbors, and embedded in their community. Because of this constant presence in their communities, the Publishers have consistently topped surveys as the nation's most trusted news sources.[7]

---

[3] OpenAI, *Our structure*, https://openai.com/our-structure.
[4] *Id.*
[5] *Id.*
[6] *See* Knight Found. & Gallup, *State of Public Trust in Local News* (2019), https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/440/original/State_of_Public_Trust_in_Local_Media_final_.pdf.
[7] *See* John Gramlich, Q&*A: What Pew Research Center's new survey says about local news in the U.S.*, Pew Research Ctr. (Mar. 16, 2019), https://www.pewresearch.org/short-reads/2019/03/26/qa-what-pew-research-centers-new-survey-says-about-local-news-in-the-u-s.

62. Plaintiff RCI is the publisher of several print and digital outlets focused on the New York metropolitan area. It employs approximately 150 employees. Its flagship publication is the *Long Island Herald*, which publishes 24 editions serving approximately 80 Long Island communities. Its publications also include *The Riverdale Press* and *The Jewish Star*. Its oldest outlet began publishing in 1883. *The Riverdale Press* won a Pulitzer Prize for Editorial Writing in 1998.

63. Plaintiff AIM Media Indiana Operating, LLC is the publisher of several print and digital outlets serving central and south-central Indiana. It employs 114 employees and has an approximate total circulation of 20,000. Its flagship publication is *The Republic*, which serves Columbus and Bartholomew County, Indiana, and began publishing in 1872. Its publications, including *The Republic*, the *Franklin Daily Journal*, and the *Greenfield Daily Reporter* have won numerous awards from the Hoosier State Press Association.

64. Plaintiff AIM Media Midwest Operating, LLC is the publisher of several print and digital outlets serving western and central Ohio. It employs approximately 118 employees and has an approximate total circulation of 30,000. Its oldest outlets, *The Delaware Gazette* and *The Times-Gazette*, began publishing in 1818. Its flagship publication is *The Lima News*, which serves Lima and Allen County, Ohio and has won numerous honors from the Ohio SPJ Awards and the Ohio Associated Press Media Editor Awards. The *Xenia Daily Gazette*, an AIM Media Midwest Operating, LLC publication, was the winner of a Pulitzer Prize in 1975.

65. Plaintiff AIM Media Texas Operating, LLC is the publisher of several print and digital outlets serving the Rio Grande Valley of south Texas and Ector County in the Permian Basin of west Texas. It employs approximately 100 employees and has an approximate total circulation of 18,000. Its flagship publication is *The Monitor*, which serves McAllen and Hidalgo County,

Texas. Its oldest outlet, *The Brownsville Herald*, began publishing in 1892. Its publications have won numerous journalism awards, including a 1988 Pulitzer Prize for Photography awarded to the *Odessa American*, which was also named Newspaper of the Year in 2025 by the Freedom of Information Foundation of Texas. Reporters employed by the *Odessa American*, as well as numerous other publications owned by AIM Media Texas Operating, LLC—including *The Monitor*, the *Valley Morning Star*, and *The Brownsville Herald*—received multiple Associated Press Managing Editor Awards in 2025 and 2026.

66. Plaintiff AmNews Corp. d/b/a The New York Amsterdam News, is the publisher of *The New York Amsterdam News*, serving the five boroughs of New York City and Westchester County, with a particular focus on New York City's Black and Brown community. It employs approximately 18 full-time employees and works with approximately 46 regular freelancers, and it has an approximate total monthly print and digital circulation of 106,000. *The New York Amsterdam News* began publishing in 1909. The publication, and its employees, have received numerous awards and recognition from the industry, including the 2024 *Editor & Publisher*'s Publisher of the Year award, which was the first time in the award's 125-year history that the honor was given to a Black woman publisher.

67. Plaintiff Arkansas Democrat-Gazette, Inc. publishes print and digital news. It serves the State of Arkansas. It employs 265 employees and has an approximate total circulation of over 30,000; it attracts over 700,000 monthly users in online traffic across its platforms. The *Arkansas Democrat-Gazette* won the Arkansas Press Association's top General Excellence award for larger daily newspapers in 2025, 2024, 2021, 2019, 2016, and 2008. The *Arkansas Democrat-Gazette* owns registered copyrighted works, including those set forth in Exhibit A.

68. Plaintiff Casa Grande Valley Newspapers Inc. is the publisher of print and digital outlets serving western Pinal County, Arizona. It employs approximately 74 employees and has an approximate total circulation of 36,851. Its publications include the *Casa Grande Dispatch* and the *Maricopa Monitor*. The *Casa Grande Dispatch* traces its roots to 1892. The publications have won several statewide and national awards, including recognition as the Arizona Newspapers Association's 2020 Non-Daily Newspaper of the Year.

69. Plaintiff CherryRoad Media Inc. is the publisher of approximately 102 print and digital outlets serving communities across nineteen states. It employs approximately 300 employees and has an approximate total circulation of 60,000. Its oldest outlet, the *Chronicle Express*, has a history dating back to 1824.

70. Plaintiff Community Impact Newspaper Co. publishes 37 local print and digital outlets serving the Austin, Houston, Dallas, and San Antonio regions; it published 40 outlets during the relevant timeframe in this Complaint. It currently employs approximately 165 people and has an approximate total circulation of 2 million.

71. Plaintiff Concord Publishing House, Inc. is the publisher of numerous local newspapers serving communities in Missouri, Illinois, Arkansas, and Tennessee. It employs approximately 87 people and has an approximate total circulation of over 19,000. Each of its newspapers has existed for over a century, while the oldest, the *Osceola Times*, was founded in 1870. Concord Publishing House, Inc.'s newspapers regularly receive top state press association awards for their class. For example, its publication, the *Southeast Missourian*, was recognized by Suburban Newspapers of America as "Best Newspaper" in the nation for circulation 50,000 and below in 2006. Local Media Association recognized the *Southeast Missourian* as second place in "Best Newspaper Contest" for 2014. In addition to thousands of first-place awards in journalism

contests, Concord Publishing House, Inc.'s newspapers regularly receive service awards for their content, which deal with critical issues for their communities. Concord Publishing House, Inc. owns registered copyrighted works, including those set forth in Exhibit B.

72. Plaintiff D.A. Publishing, LC publishes the *Standard Democrat*, located in Sikeston, Missouri, which primarily serves Scott, Mississippi, as well as New Madrid and Stoddard counties in Missouri. The *Standard Democrat*, founded in 1903, issues both print and online content. It employs 16 people and has a total circulation of approximately 1,032. It has been the recipient of top state press association awards for its class.

73. Plaintiff D-R Media and Investments, LLC is the publisher of several print and digital outlets serving the Florida counties of Highlands, Lake, Polk, and Sumter. It employs 51 employees and has an approximate total circulation of 70,500. Its publications include the *Highlands News-Sun*, the *Lakeland Sun*, the *Clermont Sun*, the *Winter Haven Sun*, the *Four Corners Sun*, the *Triangle Sun*, and the *Sumter Sun Times*. Its publications have won several Florida Press Association awards across a range of categories. The *Highland News Sun*, which currently publishes daily, has a history that dates back to 1919.

74. Plaintiff Eagle Urban Media LLC is the publisher of several print and digital outlets serving Brooklyn and Queens, New York. It employs approximately 30 employees and has an approximate total circulation of 140,000. Its flagship publication is the *Brooklyn Daily Eagle*. It owns registered copyrighted works.

75. Plaintiff El Crepusculo, Inc. is the publisher of several print and digital outlets serving the New Mexico counties of Taos, Colfax, and the cities within. It employs 22 employees and has an approximate total circulation of 7,500. Its flagship publication, the *Taos News*, began publishing in 1959. It been recognized as "Best Weekly Newspaper in the United States" by the

National Newspaper Association of America in 12 of the last 20 years and received the General Excellence Award for Best Weekly Newspaper in New Mexico, 2001-2006 and 2008-2024 from the New Mexico Press Association for its circulation category.

76. Plaintiff H.S. Gere & Sons, Inc. is the publisher of several print and digital outlets serving the Pioneer Valley region of western Massachusetts. It employs approximately 35 employees and has an approximate total paid circulation of 8,500. Its publications include the *Daily Hampshire Gazette*, the *Amherst Bulletin*, and the *Valley Advocate*. The *Daily Hampshire Gazette* traces its history to 1786. It has been named Newspaper of the Year in its circulation category by the New England Newspaper & Press Association several times. H.S. Gere & Sons, Inc. owns registered copyrighted works, including those set forth in Exhibit C.

77. Plaintiff Iowa Information Inc. is the publisher of 13 print and digital outlets serving western Iowa. Its oldest outlet began publishing in 1872. It employs more than 85 employees and has an approximate total circulation of 18,000. Its flagship publication is the *N'West Iowa Review*, based in Sheldon, Iowa, which covers a 45-mile radius of the surrounding region and is widely regarded by its peers as one of the best weekly newspapers in the country.

78. Plaintiff Lakeway Publishers, Inc. is the publisher of several print and digital outlets serving east and middle Tennessee and Virginia. It employs 82 employees and has an approximate total circulation of 27,633. Its flagship publication is the *Citizen Tribune*, which serves Morristown, Tennessee, and its surrounding counties. Lakeway's oldest outlet began publishing in 1966, and over the company's sixty-year history, its publications have won numerous state and national press association awards.

79. Plaintiff The New Mexican Inc. is the publisher of the *Santa Fe New Mexican*, serving primarily the New Mexico counties of Santa Fe, Los Alamos, and Sandoval, as well as

northern New Mexico. It currently has a net paid distribution of 15,500 and employs 150 employees. The *Santa Fe New Mexican* began publishing in 1849 and has collected countless journalism awards during its 176-year history as the state's first newspaper. Its pages have chronicled the history of New Mexico, making it the state's most-respected news source. The *Santa Fe New Mexican* routinely has been named the Best Newspaper in New Mexico in the annual New Mexico Press Association contest. It was recognized nationally on *Editor & Publisher* magazine's "Media That Matters" list for 2025. In 2026, the Local Media Association recognized the *Santa Fe New Mexican* with two national awards, one for Audience Growth Strategy and another for Public Service related to a months-long investigation into unburied bodies found along the southern New Mexico border. It owns registered copyrighted works, including those set forth in Exhibit D.

80. Plaintiff Newspapers of Massachusetts, Inc. is the publisher of the *Greenfield Recorder*, which traces its roots to 1792 and serves Franklin County and the North Quabbin region in Massachusetts and the *Athol Daily News*, which covers the North Quabbin region. It employs approximately 20 employees and has an approximate total paid circulation of 6,600. Both outlets publish print and digital content.

81. Plaintiff Newspapers of New England, Inc. is the parent company of Plaintiffs H.S. Gere & Sons, Inc., Newspapers of Massachusetts, Inc., and Newspapers of New Hampshire, Inc.

82. Plaintiff Newspapers of New Hampshire, Inc. is the publisher of several print and digital outlets, including the *Concord Monitor*, the *Valley News*, and the *Monadnock Ledger-Transcript*. Its publications serve the Capital Region of New Hampshire, 46 towns along the New Hampshire–Vermont border along the Connecticut River, and the Peterborough, New Hampshire region. It employs approximately 120 employees and has an approximate total paid circulation of

16,000; it operates a commercial press facility. Its oldest outlet, the *Monadnock Ledger-Transcript*, traces its roots to 1849; the *Concord Monitor* began publication in 1864. Its publications have won numerous awards, including a 2008 Pulitzer Prize and repeated recognition as Newspaper of the Year by the New England Newspaper & Press Association. It owns registered copyrighted works, including those set forth in Exhibit E.

83. Plaintiff North Country This Week, Inc. is the publisher of *North Country Now* and *North Country This Week*, which serves Potsdam, Massena, Ogdensburg, and Canton in the North Country region of northern New York. It employs 20 employees and has an approximate total circulation of 18,000. *North Country This Week* began publishing in 1984. Its publications have won dozens of Better Newspaper Contest awards from the New York Press Association.

84. Plaintiff The Ogden Newspapers, Inc. and its affiliates ("Ogden") publish numerous print and digital outlets serving communities in 17 states across the nation. Ogden employs approximately 1,400 employees and has an approximate total print circulation of 300,000. Ogden was founded in 1890, and Ogden's newspapers have been continually and privately family-owned since that time; several of its newspapers predate the company itself. Ogden's publications have won numerous state press association awards for reporting, photography, and commentary.

85. Plaintiff Patchogue Advance, Inc. publishes several outlets that serve the Long Island area, including the *Islip Bulletin*, the *Long Island Advance*, the *Suffolk County News*, the *Tri Hamlet News*, and the *Tide of Moriches and Manorville*. It employs 13 people and has a total circulation of approximately 13,400. Its publications have won many awards from the New York Press Association.

86. Plaintiff Rust Publishing ID, LC is the publisher of the *Mountain Home News*, serving Elmore County, Idaho. It employs 6 employees and has an approximate circulation of 1,170. The *Mountain Home News* began publishing in 1888.

87. Plaintiff Rust Publishing MOKS, LC is the publisher of the *Nevada Daily Mail*, serving Vernon County, Missouri, and the *Fort Scott Tribune*, serving Bourbon County, Kansas. It employs 15 employees and has an approximate total circulation of 2,137. The *Nevada Daily Mail*, located in Nevada, Missouri, was founded in 1865. The *Fort Scott Tribune*, located in Fort Scott, Kansas, was founded in 1884. The publications have been regular recipients of the top state press association awards for their class.

88. Plaintiff Rust Publishing NE, LC is the publisher of the *McCook Gazette*, serving southwestern Nebraska and northwestern Kansas. It employs 23 employees and has an approximate circulation of 2,564. The *McCook Gazette* traces its history to 1911. Its publications have been regular recipients of the top state press association awards for their class.

89. Plaintiff Rye Media Partners LLC is the publisher of *The Rye Record*, serving Rye, Harrison, and Purchase, New York. It employs three employees and has an approximate circulation of 10,500. *The Rye Record* began publishing in 1996. In addition to its flagship newspaper, which maintains online content, Rye Media Partners LLC produces two newsletters and a podcast. Its publications have won numerous awards from the New York Press Association, including recognition as Non-Daily Newspaper of the Year for 2025, the 2025 Stuart C. Dorman Award for Editorial Excellence, and a total of 23 awards in the 2025 NYPA Better Newspaper Contest.

90. Plaintiff Sentinel Media Co., Inc. is the publisher of the *Daily Sentinel* and the *Boonville Herald*, serving central New York, including the cities of Rome and Utica. It employs

35 employees and has an approximate total circulation of 9,500. The *Daily Sentinel* traces its roots to 1864.

91. Plaintiff Shaw Family Holdings, Inc.is the publisher of print and digital media such as the *Northwest Herald*, serving northern Illinois and northwest Indiana. It employs 186 employees and has an approximate total circulation of 71,000 print and digital subscribers. Its oldest outlet traces its roots to 1851, and its publications have won numerous Illinois Press Association awards.

92. Plaintiff Straus Media-Manhattan, LLC is the publisher of several local news outlets in Manhattan. It has served the community since 1970. It has 7 employees and an approximate total circulation of 50,000. Its publications include *Our Town*, *Our Town Downtown*, *The West Side Spirit*, and the *Chelsea News*.

93. Plaintiff Straus Newspapers, Inc. is the publisher of several local news outlets in Orange County New York, Pike County, Pennsylvania, Sussex County New Jersey, and Passaic County New Jersey. It has 28 employees and an approximate total circulation of 80,000. Its publications include the *Warwick Advertiser*, whose history dates back to 1866. Straus Newspapers, Inc.'s publications have been acknowledged for their coverage, including six recent first-place awards in the New York Press Association Better Newspaper Contest.

94. Plaintiff WEHCO Newspapers, Inc. is the publisher of numerous daily and community newspapers serving communities throughout Arkansas, as well as Texarkana in both Arkansas and Texas, and Chattanooga, Tennessee. It employs approximately 602 employees across its publications and has approximately 67,609 total paid subscribers across its outlets. It attracts over 2.3 million monthly users in online traffic across its platforms. Staff across WEHCO Newspapers, Inc.'s newspapers regularly receive recognition from the Arkansas Press Association,

the Arkansas Press Women Communications Content and the Society of Professional Journalists Awards. For example, the *Northwest Arkansas Democrat-Gazette* won the Arkansas Press Association's top General Excellence award for larger daily newspapers in 2023 and 2018. WEHCO Newspapers, Inc. is the corporate parent of Plaintiff Arkansas Democrat-Gazette, Inc., which as discussed above, owns registered copyrighted works, including those set forth in Exhibit A.

95. Plaintiff White Mountain Publishing LLC is the publisher of print and digital outlets serving southern Navajo and Apache Counties and northern Gila County, Arizona. It employs approximately 24 employees and has an approximate total circulation of 10,477. Its publications include the *White Mountain Independent* and the *Payson Roundup*. The *White Mountain Independent* traces its roots to 1885. The publications have won several statewide and national awards, including recognition of the *White Mountain Independent* as the Arizona Newspapers Association's 2014 Non-Daily Newspaper of the Year and recognition of the *Payson Roundup* as the Association's Non-Daily Newspaper of the Year twelve times.

96. Plaintiff Wick Communications is the publisher of several print and digital outlets serving Arizona, Washington, and Oregon. It employs 135 employees and has an approximate total circulation of 22,000 print and digital subscribers. Its flagship publications are the *Arizona Daily Sun*, in Flagstaff, Arizona, which traces its roots to 1883, and the *Wenatchee World*, in Wenatchee, Washington, which was founded in 1905. Its publications are frequent winners in state press association contests, including a First Place award for General Excellence given to the *Wenatchee World* by the Washington Newspaper Publishers Association and repeated honors for the *Nogales International* in the Arizona Newspapers Association's Better Newspapers Contest.

97. Together, the Publishers operate nearly 400 news outlets across 33 states around the nation. They expend significant time, money, and editorial effort in delivering quality reporting to their communities. The Publishers also dedicate significant resources to protecting their content by developing and posting terms of service for their online news sites, as well as by implementing paywalls and access restrictions. Certain Publishers also pay to register their copyrights with the U.S. Copyright Office.

98. The Publishers sustain these costly operations through various combinations of advertising sales, reader subscriptions, and content licensing. As set forth below, Defendants have robbed the Publishers of all three. By covertly scraping, copying, adapting, and using the Publishers' content to train and deploy their GenAI products, Defendants have diverted traffic from the Publishers' news sites, thereby causing a reduction in the Publishers' advertising and subscription revenue while denying the Publishers the licensing fees the Defendants should have been required to pay to use the Publishers' protected works.

B. OpenAI and Microsoft Collaborated to Develop Transformational AI Technology.

1. OpenAI's Early History.

99. OpenAI was incorporated in December 2015 as a nonprofit artificial intelligence research organization. Its founders, including Elon Musk, Reid Hoffman, Sam Altman, and Greg Brockman, launched the organization with a $1 billion investment toward the stated mission of "advancing digital intelligence" to "benefit humanity" while remaining "unconstrained by a need to generate a financial return."[8] OpenAI made this commitment publicly and repeatedly, representing to regulators, the public, and the research community that it would share its work openly and transparently with the world.

[8] *See* OpenAI, *Introducing OpenAI* (Dec. 11, 2015), https://openai.com/index/introducing-openai.

100. Those representations proved short-lived. Within three years, OpenAI began unwinding its nonprofit structure and constructing a fully commercial enterprise in its place. Today, OpenAI generates $2 billion in monthly revenue and was most recently valued at $852 billion.[9] In June 2026, OpenAI filed confidentially with the U.S. Securities and Exchange Commission for an initial public offering that could rank among the largest in the history of public markets.[10]

101. In 2022, OpenAI released ChatGPT, its flagship product. ChatGPT is an LLM-powered GenAI assistant chatbot. Users input prompts into the chatbot, and it delivers responses in plain language. GenAI systems like ChatGPT are able to produce plain language content because it learns from existing data—in this case, enormous quantities of human-authored written material.

102. ChatGPT became one of the fastest-adopted consumer products in history, reaching one million users within a month of its release and 100 million users within three months. Its growth has only continued since. Now, ChatGPT has over "900 million weekly active users, and over 50 million subscribers."[11] More than one year ago, OpenAI CEO Sam Altman described ChatGPT's reach as encompassing something on the order of 10% of the world's population.[12]

103. OpenAI offers a robust and expanding suite of LLM-powered services to its massive user base. Currently, OpenAI offers free and paid access to ChatGPT for individual consumers. The free version of ChatGPT provides users with limited access to GPT-5.5, OpenAI's

---

[9] OpenAI, *OpenAI raises $122 billion to accelerate the next phase of AI* (Mar. 31, 2026), https://openai.com/index/accelerating-the-next-phase-ai.
[10] Cade Metz, *OpenAI Files to Go Public as A.I. Companies Rush to Wall St.*, N.Y. Times (Jun. 8, 2026), https://www.nytimes.com/2026/06/08/technology/openai-ipo.html.
[11] OpenAI, *OpenAI raises $122 billion to accelerate the next phase of AI* (Mar. 31, 2026), https://openai.com/index/accelerating-the-next-phase-ai.
[12] Beatrice Nolan, *Sam Altman Says '10% of the World Now Uses Our Systems a Lot' as Studio Ghibli-Style AI Images Help Boost OpenAI Signups*, Fortune (Apr. 14, 2025), https://fortune.com/2025/04/14/sam-altman-openai-user-base-doubled-few-weeks-10-of-world-uses-system.

latest model. Consumers can then pay a monthly fee, starting at $8-per-month and increasing to $100-per-month, for increasing access to GPT-5.5 and other features.

104. OpenAI has used similar subscription models throughout the relevant time period. For example, OpenAI previously made available a version of ChatGPT powered by GPT-3.5 to users for free, and OpenAI also offered a premium service, powered by OpenAI's then "most capable model" GPT-4, to consumers for $20 per month.

105. OpenAI also offers business-focused subscriptions. These offerings include ChatGPT Enterprise and ChatGPT API tools designed to enable developers to incorporate ChatGPT into bespoke applications. OpenAI offers these tools to businesses via different payment options. For example, businesses may pay as they go to obtain OpenAI's "Business Codex," pay a monthly rate of $20 per user to obtain Business ChatGPT & Codex, or use OpenAI's custom pricing model to obtain ChatGPT Enterprise.

106. Upon information and belief, OpenAI plans to introduce new products that use similar methods and provide similar output (including the Publishers' content).

107. OpenAI's models and products have been immensely valuable for OpenAI. As of 2026, public reporting estimated that over 92% of Fortune 500 companies are using ChatGPT.[13] As noted above, OpenAI generates $2 billion in revenue per month and has a post-money valuation of $852 billion.[14] Some analysts suggest OpenAI's IPO valuation could exceed $1 trillion, a number that exceeds the 2025 gross domestic product of all but 21 countries.[15]

---

[13] Sameer Khan, *50 ChatGPT Statistics Every Business Leader Should Know in 2026*, AI Business Weekly (Feb. 28, 2026), https://aibusinessweekly.net/p/chatgpt-statistics.

[14] *See* OpenAI, *OpenAI raises $122 billion to accelerate the next phase of AI* (Mar. 31, 2026), https://openai.com/index/accelerating-the-next-phase-ai.

[15] Douglas A. McIntyre, *OpenAI's $1 Trillion IPO*, Yahoo! Finance (Oct. 30, 2025), https://finance.yahoo.com/news/openai-1-trillion-ipo-141519224.html.

108. Upon information and belief, all the OpenAI Defendants have been either directly involved in or have directed, controlled, and benefitted from OpenAI's widespread infringement and commercial exploitation of the Publishers' protected works. OpenAI, Inc., alongside Microsoft, controlled and directed the widespread reproduction, distribution, and commercial use of the Publishers' material 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 stripping, copying, reproduction, distribution, and commercial use of the Publishers' protected 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 stripping, copying, 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 OpenAI Group PBC.

2. Microsoft's Role as Co-Architect.

109. Microsoft has been an indispensable partner in virtually every aspect of OpenAI's commercial enterprise. In 2019, Microsoft made an initial $1 billion investment in OpenAI, the first installment of what would become a series of investments totaling approximately $13 billion. More than making passive investments, however, Microsoft became deeply intertwined with OpenAI as Microsoft became the primary architect of the physical and computational infrastructure on which OpenAI's models were trained and on which the Publishers' copyrighted works were reproduced and misused at a massive scale. As Microsoft CEO Satya Nadella has said, Microsoft works on every aspect of the joint enterprise—performing optimizations, building

tooling, and constructing infrastructure—and affirming the public view that the venture is a "joint project between Microsoft and OpenAI."[16]

110. First, Microsoft and OpenAI jointly developed Azure-based AI supercomputing infrastructure, including a dedicated supercomputer that Microsoft built and operated for OpenAI's exclusive use. The system was designed to allow OpenAI to train its AI models on essentially the entire public internet. The system provided an integrated architecture comprising more than 285,000 Central Processing Unit ("CPU") cores, 10,000 Graphics Processing Units ("GPUs"), and 400 gigabits per second of network connectivity per GPU server, placing it among the five most powerful publicly known supercomputing systems in the world.[17] According to Nadella, Microsoft redesigned its data center infrastructure to be able to process OpenAI's massive training workloads, which were unlike any that had previously existed, and in order to provide the infrastructure that made OpenAI's frontier models achievable.[18]

111. In short, Microsoft "build[s] the infrastructure to train [OpenAI's] models. They're innovating on the algorithms and the training of these frontier models."[19]

112. Second, Microsoft has helped bring OpenAI's products to market by integrating them into Microsoft's own products. In February 2023, Microsoft launched Bing Chat, now rebranded as Copilot, a GenAI product powered by OpenAI's latest GPT model and integrated into Microsoft's search engine. Shortly thereafter, Microsoft and OpenAI jointly released Browse with

---

[16] *Satya Nadella on Hiring the Most Powerful Man in AI: When OpenAI threw Sam Altman overboard, Microsoft's CEO saw an opportunity*, On With Kara Swisher, N.Y. Mag. (Nov. 21, 2023) https://nymag.com/intelligencer/2023/11/on-with-kara-swisher-satya-nadella-on-hiring-sam-altman.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] *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-transcript-microsoft-ceo-satya-nadella-speaks-with-cnbcs-jon-fortt-on-power-lunch-today.html.

[19] Microsoft, *Full Keynote: Satya Nadella at Microsoft Inspire 2023*, at 24:00, YouTube (Jul. 18, 2023), https://www.youtube.com/watch?v=RhwVMt_XCUE.

Bing, a plug-in on the Bing browser allowing OpenAI's LLM to retrieve and summarize live web content.

113. Upon information and belief, like ChatGPT, these Microsoft products also provide responses to user input by taking content from the Publishers' websites and delivering them to the user without necessarily directing the user to that site or including any attribution in its response.

114. As described in further detail below, by substituting their own product outputs for conventional search results, Defendants capture user engagement that would otherwise accrue to the Publishers, depriving them of the user traffic and associated revenue their content would otherwise generate.

C. Defendants Improperly Used the Publishers' Protected Works in Developing and Deploying their GenAI Products.

115. Defendants' massively lucrative enterprise relied on rampant copyright infringement. Microsoft and OpenAI created and distributed reproductions of the Publishers' works while using those works to train their LLMs and deploying the GenAI products that incorporate them.

116. Microsoft and OpenAI improperly scraped, copied, reproduced, and used the Publishers' protected works without authorization, both in training their LLMs and in deploying their GenAI products.

117. As described in further detail below, OpenAI's secrecy effectively made it impossible for copyright holders to determine whether their works had been swept into OpenAI's training pipeline without permission or compensation. It was not until years later that the Publishers and other copyright holders would discover that they had.

118. As it pivoted from a nonprofit mission to profit-making, OpenAI also retreated from transparency—a move that has made it nearly impossible for copyright holders such as the

Publishers to uncover infringing activity. In its early years, OpenAI released its first two GPT models, GPT-1 and GPT-2, under open-source licenses and published substantial information about the resources it used to develop them. Beginning with its GPT-3 release in May 2020, however, OpenAI stopped disclosing meaningful information to the public about how its products were developed and what datasets it was using to train them. In fact, when it released GPT-4 in March 2023, OpenAI released a "technical report" in which it explicitly acknowledged that it was not disclosing "further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."[20] When OpenAI released GPT-5 in 2025, it did not even release an accompanying "technical report." Instead, it produced a "System Card" that focuses on safety evaluations and results but contains very little, if any, details on architectural specifications or training data sources.[21]

1. Defendants' LLM Development Process.

119. Defendants' LLM development process starts with a collection stage, during which Defendants assemble and store massive training datasets drawn from text scraped across the internet, including from news and media sites like those of the Publishers, reference sources, books, and other written works.

120. Defendants have used a series of such datasets across OpenAI's GPT model generations, including filtered versions of Common Crawl—a broad web archive encompassing billions of web pages dating back to 2008—as well as its own curated datasets, WebText and WebText2, which it assembled by curating high-quality sources of online content, such as news sites.

---

[20] OpenAI, *GPT-4 Technical Report* (Mar. 22, 2023), https://cdn.openai.com/papers/gpt-4.pdf.
[21] OpenAI, *GPT-5 System Card* (Aug. 7, 2025), https://openai.com/index/gpt-5-system-card; *see also* Eric Landau, *GPT-5: A Technical Breakdown*, Encord (Aug. 8, 2025), https://encord.com/blog/gpt-5-a-technical-breakdown ("OpenAI has not released detailed architecture specs or training data sources.").

121. The "training" process involves repeatedly feeding these massive datasets, which inevitably include copyrighted content, through the LLM in an iterative process. Text from the training data is broken into "tokens," which are encoded units of words and sub-words, and stored in computer memory. The LLM is then programmed to repeatedly attempt to predict masked or missing tokens within sequences of text, adjusting billions of internal numerical parameters with each iteration to improve its predictions. Repeated with great frequency, this process results in an LLM whose parameters encode the statistical patterns—and, in many cases, the specific content—of the works on which it was trained. This process is what enables ChatGPT to generate responses to user prompts.

122. After initial training, LLMs are typically refined through fine-tuning, which involves additional training rounds using targeted content or human feedback, further shaping the model's outputs and deepening its encoding of specific material.

123. Researchers, and now copyright owners and courts, have observed that LLMs trained this way exhibit a behavior called "memorization."[22] Given the right prompt, LLMs will repeat large portions of materials on which they were trained. This phenomenon shows that LLM parameters encode retrievable copies of many of those training works.

124. After an LLM is trained, OpenAI's models are enhanced through a technique called retrieval-augmented generation, or "RAG," which takes place at the point of response generation rather than during training. Rather than relying solely on knowledge encoded in the model's parameters, a RAG-enabled system queries external sources in real time, including by crawling or

[22] Gerrit J.J. van den Burg & Christopher K.I. Williams, *On Memorization in Probabilistic Deep Generative Models*, 35th Conference on Neural Information Processing Systems (Dec. 29, 2021), https://arxiv.org/pdf/2106.03216.

indexing the Publishers' websites, and incorporates the retrieved content directly into its generated responses.[23]

2. Defendants Copied and Reproduced the Publishers' Works Without Authorization while Training OpenAI's LLMs.

125. As described above, Defendants developed the LLMs that power OpenAI's GPT products by training and fine-tuning the models on massive datasets of human-made text material. Without permission or compensation, Defendants copied the Publishers' materials to populate certain of the datasets they used to train various versions of their LLMs.

126. According to OpenAI's own reports, GPT-2 included 1.5 billion parameters and was developed using a training dataset OpenAI built called WebText, which includes "the text contents of 45 million links posted by users of the 'Reddit' social network."[24] This includes the text contents of works owned by many of the Publishers.

127. An analysis performed by a technologist consultant employed by Plaintiffs' counsel found that an open-source approximation of the WebText dataset called OpenWebText contains millions of tokens (basic units of text) worth of content extracted from the Publishers' news sites, broken down as follows:[25]

Publisher	OpenWebText Tokens
AIM Media Indiana Operating, LLC	891,256
AIM Media Midwest Operating, LLC	158,658
AIM Media Texas Operating, LLC	100,550
AmNews Corp.	706,403
Arkansas Democrat-Gazette, Inc.	138,144
Casa Grande Valley Newspapers Inc.	4,576
CherryRoad Media Inc.	550,205

[23] As discussed further below, OpenAI deploys RAG functionality across several of its products, including ChatGPT Search, Deep Research, and—in collaboration with Microsoft—Copilot and Browse with Bing.
[24] *GPT-2 Model Card*, Github (Nov. 2019), https://github.com/openai/gpt-2/blob/master/model_card.md; Tom B. Brown, *et al.*, *Language Models are Few-Shot Learners*, at 9 (July 22, 2020), https://arxiv.org/pdf/2005.14165.pdf.
[25] Aaron Gokaslan & Vanya Cohen, *OpenWebText Corpus*, Skylion007 (2019), https://skylion007.github.io/OpenWebTextCorpus.

Community Impact Newspaper Co.	115,721
D.A. Publishing, LC	526
Eagle Urban Media LLC	116,229
El Crepusculo, Inc.	19,171
H.S. Gere & Sons, Inc.	106,853
Lakeway Publishers, Inc.	20,398
The New Mexican Inc.	14,848
Newspapers of New Hampshire, Inc.	142,467
North Country This Week, Inc.	13,944
Patchogue Advance, Inc.	2,557
Rust Publishing NE, LC	484
WEHCO Newspapers, Inc.	179,415
White Mountain Publishing LLC	96,772
Wick Communications	980

128. GPT-3, for its part, includes 175 billion parameters and was trained on a weighted combination of datasets, including the Common Crawl, WebText2, Books1, and Books2 datasets, as well as a dataset comprised of Wikipedia content.[26]

129. The most highly weighted dataset in GPT-3, Common Crawl, is a "copy of the Internet" made available by a 501(c)(3) organization of the same name that systematically scrapes the open internet, archiving petabytes (one petabyte is roughly one quadrillion bytes) of raw web pages, metadata, and text to make massive-scale data accessible to researchers, AI developers, and the public.[27]

130. A straightforward analysis of the Common Crawl dataset clearly shows that hundreds of thousands of the Publishers' copyrighted articles and tens of millions of tokens worth of the Publishers' content likely appeared in the Common Crawl training dataset Defendants used to train GPT-3.

[26] *Id.* at 9-10.
[27] *Commoncrawl Foundation*, ProPublica https://projects.propublica.org/nonprofits/organizations/261635908/202403189349101980/full; Common Crawl, *Frequently Asked Questions*, https://commoncrawl.org/faq.

131. An analysis performed by a technologist consultant employed by Plaintiffs' counsel found that in C4, a filtered English-language subset of a 2019 snapshot of Common Crawl, the Publishers' websites account for more than ***115 million tokens*** (basic units of text), broken down as follows:[28]

Publisher	C4 Tokens
AIM Media Indiana Operating, LLC	1,487
AIM Media Midwest Operating, LLC	15,119,562
AIM Media Texas Operating, LLC	123,466
AmNews Corp.	653,629
Arkansas Democrat-Gazette, Inc.	2,096,772
CherryRoad Media Inc.	15,854,582
Community Impact Newspaper Co.	1,306,989
Concord Publishing House, Inc.	1,976,747
D-R Media and Investments, LLC	150,164
D.A. Publishing, LC	194,866
Eagle Urban Media LLC	1,275,691
El Crepusculo, Inc.	211,479
H.S. Gere & Sons, Inc.	1,207,887
Iowa Information Inc.	299,932
Lakeway Publishers, Inc.	1,299,443
The New Mexican Inc.	334
Newspapers of Massachusetts, Inc.	736,794
Newspapers of New Hampshire, Inc.	2,346,200
North Country This Week, Inc.	507,630
Patchogue Advance, Inc.	230,357
Richner Communications, Inc.	2,963,190
Rust Publishing ID, LC	75,875
Rust Publishing MOKS, LC	161,274
Rust Publishing NE, LC	677,230
Rye Media Partners LLC	118,855
Shaw Family Holdings, Inc.	16,318
Sentinel Media Co., Inc.	246,823
Straus Media-Manhattan, LLC	378,846
Straus Newspapers, Inc.	1,204,814
The Ogden Newspapers, Inc.	71,039,148
WEHCO Newspapers, Inc.	6,346,220

[28] Jesse Dodge, *et al.*, *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus* (2021), https://arxiv.org/abs/2104.08758.

White Mountain Publishing LLC	2,827
Wick Communications	389,483

132. Attached as Exhibit F is a more granular breakdown of the number of tokens that appear in the C4 dataset for each URL domain owned by the Publishers that appeared in C4. To be clear, this is just one aspect of the breadth of the Publishers' protected material that Defendants pirated.

133. The number of parameters OpenAI uses to train its GPT models has increased exponentially with each new version release. On information and belief, the Defendants have used, and continue to use, WebText, WebText2, Common Crawl, and other training datasets containing the Publishers' protected content to train the GPT models.

134. Both OpenAI and Microsoft performed the infringing activity and collaborated in the unauthorized scraping, copying, storing, and reproduction of the Publishers' works that were required to train the GPT models. Both OpenAI and Microsoft caused hundreds of thousands of the Publishers' works to be copied and repeatedly run through the models.

135. As described above, Microsoft developed the underlying computing infrastructure for OpenAI so that the latter could scrape, copy, store, and reproduce the LLM training data at issue. Upon information and belief, Microsoft was intimately involved in every step of OpenAI's LLM training, from designing the models themselves to building the training datasets and supervising the models' training and fine-tuning. Regardless of whether Microsoft directly selected the Publishers' works to train OpenAI's models, it knew or was willfully blind to the source of the data and that it was enabling and executing training processes using copyrighted material. Microsoft was able to control and prevent the infringing activity but failed to do so. And

it reaped significant rewards as a result—both from its investment in OpenAI and in the incorporation of ChatGPT in its own products.

3. Defendants Reproduced and Publicly Displayed the Publishers' Copyrighted Works Without Authorization while Deploying its GPT Products.

136. Defendants' infringement has also been observed in the outputs of their GPT Products. Researchers, courts, and other copyright holders have taken notice that various models of ChatGPT have demonstrated their underlying LLMs' propensity to "memorize" content on which they were trained and reproduce copies of it near-verbatim in response to targeted user prompts.

137. This phenomenon, well-documented in academic literature and confirmed by memorialized outputs from OpenAI's products, constitutes direct evidence that the works that appear in Common Crawl and other training datasets were copied during the training process and that retrievable copies of their expressive content remain encoded within the models' parameters.

138. The Publishers and other news publishers have been among the most affected groups of copyright owners whose protected material is contained in the Defendants' training data. Investigative testing conducted in connection with substantially identical copyright infringement actions brought against the Defendants by The New York Times, the New York Daily News, the Chicago Tribune, the Denver Post, the Orlando Sentinel, the Sun Sentinel, the San Jose Mercury News, the Orange County Register, the St. Paul Pioneer Press, the Center for Investigative Reporting, The Intercept, Raw Story, AlterNet, and Ziff Davis ("News Plaintiffs"), among others, has demonstrated that GPT-based models will reproduce substantial portions of published news articles when prompted appropriately, in some instances generating hundreds of words of verbatim text from a single article in a single response.

139. Across these cases, these News Plaintiffs have supported their claims with side-by-side comparisons showing near-identical matches between AI-generated output and original published articles, establishing that news publishers' content from Defendants' training data has been encoded into OpenAI's models through the LLM training process.

140. Upon information and belief, since these News Plaintiffs conducted their investigative testing, OpenAI has disabled public access to the versions of ChatGPT that produced the demonstrably infringing, near-verbatim outputs they presented in their pleadings.

141. Upon information and belief, given that Defendants' training datasets, including WebText and Common Crawl, contained an enormous quantity of the Publishers' protected content, the previous, now-offline versions of ChatGPT that these News Plaintiffs tested would likewise generate near-verbatim copies of the Publishers' copyrighted works if prompted to do so.

142. Each instance of a memorized news article constitutes an unauthorized copy or derivative work belonging to that article's publisher that was used to train one or more of OpenAI's LLM models. Defendants therefore directly engaged in the unauthorized reproduction and publication of news articles in its training data to produce outputs from its GenAI products.

143. In addition to memorizing training data, Defendants' products can also generate infringing output through RAG functionality. As described above, RAG-enabled products such as ChatGPT Search and Deep Research retrieve content from the Publishers' websites in real time and incorporate that content, oftentimes verbatim or in close paraphrase, into synthesized responses delivered to users. Unlike memorization, which reflects content encoded during training, RAG-based infringement occurs after a prompt from the user, at which point the model actively copies and processes publishers' current content to produce its output.

144. As with their memorization testing, several news publishers that have brought substantially identical copyright lawsuits against the Defendants have demonstrated that previous LLM models would, when prompted accordingly, retrieve and regurgitate protected news content that is publicly available on the internet.

145. Upon information and belief, in response to claims brought by News Plaintiffs, OpenAI has disabled public access to the models of ChatGPT that produced the demonstrably infringing, near-verbatim outputs that the News Plaintiffs presented in their pleadings. But prior to that, Defendants' LLM models when prompted accordingly would retrieve and regurgitate protected news content from the Publishers' websites.

146. The presence of the Publishers' copyrighted works in the Defendants' training data, combined with the copious amount of evidence that such training data has been regurgitated by the Defendants' products strongly suggests that the Publishers' content can and has been reproduced and displayed publicly as a result of user prompts. Further information regarding the regurgitation and output of the Publishers' content is solely in OpenAI's possession, custody, and control.

4. Defendants' Infringement Was Willful.

147. Defendants' actions described above were performed willfully.

148. As two of the world's leading technology companies, Defendants possessed full knowledge of how their training models and products worked. All Defendants were involved in each step of the training, fine-tuning, testing, and commercialization of the GPT models.

149. OpenAI has expressly stated, "it would be impossible to train today's leading AI models without using copyrighted materials."[29] And Microsoft, through its statements and investment activity, has made clear that it is joined at the hip with OpenAI.

150. Defendants knew or should have known that the actions involved in scraping and copying of the Publisher's works on a massive scale was unauthorized. They also knew or should have known that the retention and encoding of these works within Defendants' models and systems was also unauthorized. Defendants also knew that these actions would 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.

151. In fact, reports about the corporate infighting at OpenAI shows that in late 2023, 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."[30] OpenAI ousted, but subsequently reinstated at Microsoft's behest, Altman as CEO.

152. 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.

153. Upon information and belief, Defendants were aware of many examples of copyright infringement after ChatGPT, Browse with Bing, and Copilot (formerly Bing Chat) were

[29] OpenAI, *Written Evidence before the United Kingdom House of Lords Communications and Digital Select Committee inquiry: Large language models* (Dec. 5, 2023), at 4, https://committees.parliament.uk/writtenevidence/126981/pdf.

[30] Andrew Imbrie, Owen J. Daniels & Helen Toner, *Decoding Intentions*, Ctr. for Sec. & Emerging Tech. (Oct. 2023), at 29, https://cset.georgetown.edu/wp-content/uploads/CSET-Decoding-Intentions.pdf.

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. OpenAI Removed Copyright Management Information from the Publishers' Works.

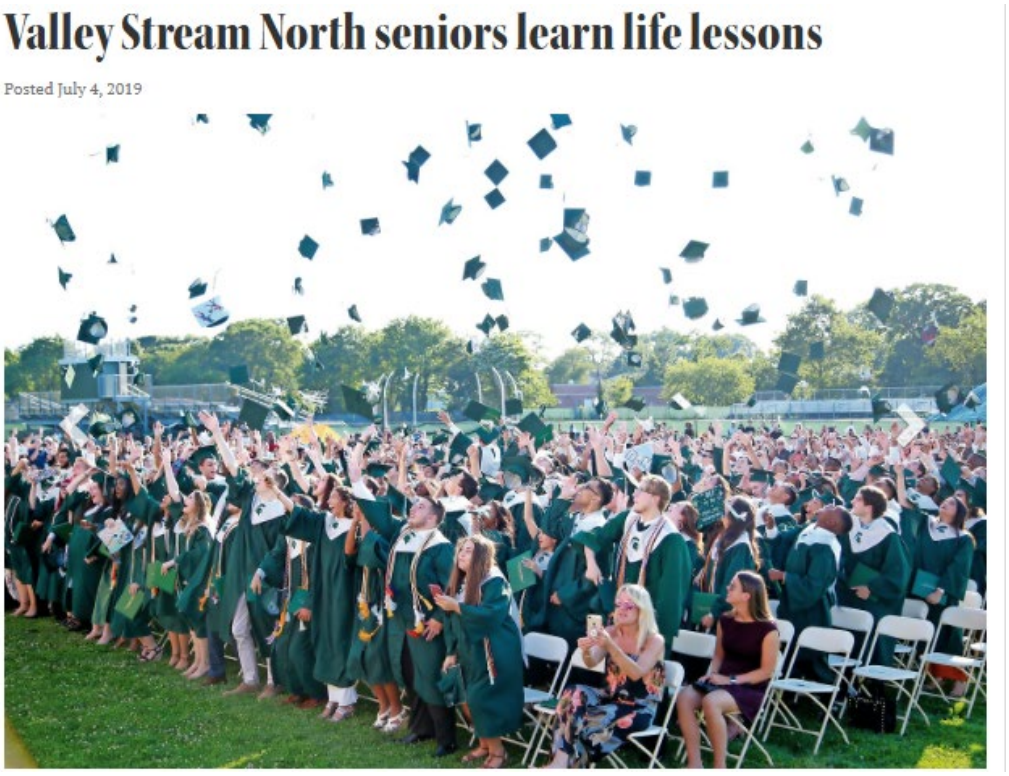
154. In 1998, Congress passed the Digital Millennium Copyright Act ("DMCA") in the wake of emerging technologies that could be used to evade existing copyright protections. Unlike copyright infringement claims, which require copyright owners to pay to register their works before bringing statutory claims, a DMCA claim does not require registration.

155. Among other things, the DMCA prohibits the knowing removal or alteration of CMI and the dissemination of works knowing that CMI has been removed or altered. *See* 17 U.S.C. § 1202(b). In so doing, the DMCA provided additional, broad protections to copyright owners such as the Publishers—particularly as digital technology was evolving.

156. The DMCA defines the term "copyright management information," or CMI, as encompassing certain "information conveyed in connection with copies . . . of a work . . . , including in digital form," such as "the title and other information identifying the work, including the information set forth on a notice of copyright," "[t]he name of, and other identifying information about, the author of a work," "[t]he name of, and other identifying information about, the copyright owner of the work, including the information set forth in a notice of copyright," "[t]erms and conditions for use of the work," "[i]dentifying numbers or symbols referring to such information or links to such information," and "[s]uch other information as the Register of Copyrights may prescribe by regulation, except that the Register of Copyrights may not require the provision of any information concerning the user of a copyrighted work." 17 U.S.C. § 1202(c).

157. In their works, the Publishers convey CMI, such as the Publishers' names, the titles, the authors' names, and terms and conditions for the use of the work.

158. For example, the following title and byline appears below RCI's article in the *Long Island Herald* entitled "Valley Stream North seniors learn life lessons":



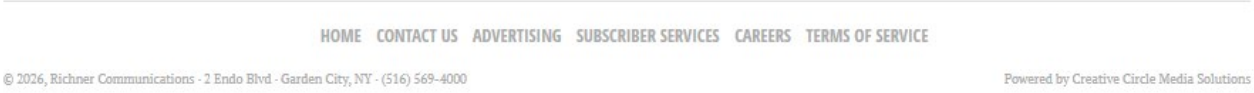
Valley Stream North seniors learn life lessons

Posted July 4, 2019

By Melissa Koenig

More than 200 North High School seniors learned life lessons on June 26 before they walked across the stage to receive their diplomas.

159. The Publishers also convey their terms and conditions and copyright notice on the pages containing articles and other works. For example, the following appears on RCI's liherald.com website containing the prior article and byline:



HOME CONTACT US ADVERTISING SUBSCRIBER SERVICES CAREERS TERMS OF SERVICE

© 2026, Richner Communications · 2 Endo Blvd · Garden City, NY · (516) 569-4000 Powered by Creative Circle Media Solutions

160. Each Publisher conveys relevantly similar CMI content on their websites. No Publisher has licensed or otherwise permitted Defendants to include any of the Publishers' works in their training datasets.

161. In compiling the datasets to train their products, Defendants not only scraped works from the Publishers, but also intentionally removed the Publishers' CMI from the Publishers' works that were collected.

162. OpenAI has stated that it eschewed training language models on "a single domain of text" in favor of "building as large and diverse a dataset as possible."[31]

163. To do so, OpenAI "scraped all outbound links from Reddit, a social media platform, which received at least 3 karma" (a form of endorsement by Reddit users) to create a dataset that contained the text subset of 45 million links. That dataset was called WebText.[32] As described above, an analysis performed by a technologist consultant employed by Plaintiffs' counsel found that an open-source approximation of the WebText dataset called OpenWebText contained thousands of records totaling millions of tokens of text from the Publishers' websites.

164. OpenAI explained, "[t]o extract the text from HTML responses we use a combination of the Dragnet and Newspaper content extractors."[33]

165. 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, including Common Crawl. After publicly describing these methodologies for GPT-2, OpenAI has never indicated a change in their text extraction methodology for later models. Nor has OpenAI ever claimed to use any other text extraction methods for later versions of ChatGPT.

[31] Alec Radford, *et al.*, *Language Models are Unsupervised Multitask Learners*, at 3, https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
[32] *Id.*
[33] *Id.* (citations omitted).

166. Upon information and belief, OpenAI used the Dragnet and Newspaper content extractors to intentionally remove CMI from the Publishers' works, including author, title, copyright notice, and terms of use information.

167. For example, in an early article describing Dragnet, its creators describe the characteristics of the "content extractor." Specifically, the article explains, "[t]he 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."[34] In a later update to "extend the original Dragnet algorithm," the design and purpose of Dragnet is described as to extract "chrome" containing certain CMI like copyright notices and footers, from the "main article content"—*i.e.*, the body of the work.[35]

168. Upon information and belief, Dragnet is designed in such a way to separate out CMI that exists in the form of author, title, copyright notices, and term of use information from main article text—the ultimate goal of the extraction. So, if CMI exists outside the main article text, it is not extracted by Dragnet.

169. Newspaper is another content extractor that OpenAI used. One key feature of this content extractor is "avoiding IP blocks" on website.[36] Upon information and belief, Newspaper is designed to separate and extract the article text on the Publishers' webpages while removing the Publishers' CMI, such as the Publishers' terms and conditions and copyright notices.

170. While users of the Newspaper content extractor can choose to extract author and title information (but not copyright notices and term of use information), upon information and

---

[34] Matthew E. Peters & Dan Lecocq, *Content Extraction Using Diverse Feature Sets* (May 2013), https://archives.iw3c2.org/www2013/companion/p89.pdf.
[35] Matthew Peters, *Benchmarking Python Content Extraction Algorithms: Dragnet, Readability, Goose, and Eatiht* (Jan. 29, 2015), https://matt-peters.github.io/benchmarking-python-content-extraction-algorithms-dragnet-readability-goose-and-eatiht.
[36] *codelucas/newspaper*, Github, https://github.com/codelucas/newspaper.

belief, OpenAI did not choose to do so. That is because Dragnet did not generally extract such information, and one purpose of using both datasets was to have consistent sets of data.

171. Upon information and belief, the reason that OpenAI chose to use Dragnet and Newspaper was because of these particular content extraction qualities. At the time OpenAI chose to use these content extractors, publicly available material on both extractors indicated that these methods remove author, title, copyright notices, and terms of use information from text. Upon information and belief, OpenAI intentionally and knowingly removed this copyright management information while assembling WebText.

172. Indeed, the use of both methods gave OpenAI backup methodologies in case a particular method did not yield optimal results for OpenAI's purposes. And having both methods would allow OpenAI to have consistency in training its models.

173. Upon information and belief, OpenAI would save the extractions from the Publishers' relevant webpage in order to have a backup of the content for future use.

174. Numerous examples, including from other similarly situated parties who have brought claims, exist within OpenAI's WebText training dataset where a copy of the full text of a certain article is retained, but without CMI such as author, title, copyright notice, and term of use information. Upon information and belief, the same is true for the Publishers' content that exists within WebText.

175. Upon information and belief, OpenAI employed the same or a similar CMI-stripping process when constructing the Common Crawl dataset it used to train its more advanced GPT models. Indeed, the C4 data set, which is a snapshot of the Common Crawl dataset that OpenAI used, contains the full text of online news articles by the Publishers devoid of the publication name, article title, subtitle, byline, date, copyright notice, and terms of use links that

appear on the Publishers' webpages. Attached as Exhibit G is an illustration of how the content from Publishers' articles appears in the C4 dataset using examples of articles from the *The Long Island Herald* and the *Arkansas Democrat-Gazette*. It follows from these examples that OpenAI removed CMI from the Publishers' works at the collection stage ahead of their use in the LLM training process.

176. The lack of CMI from copies generated and retained in OpenAI's training datasets like WebText and Common Crawl, combined with fact that the extraction methods that OpenAI used were known to OpenAI, demonstrates that OpenAI acted intentionally in removing author, title, copyright notice, and terms of use information from the Publishers' works. That harmed the Publishers, who rely on CMI information to protect their copyright interests in their works.

177. Because OpenAI knew that its models were trained on datasets that had removed CMI, it knew that its chatbot products would likely disseminate unauthorized copies of the Publishers' works without CMI to its customers.

178. OpenAI knew that the Publishers' works would become outputs of OpenAI's GenAI products, conveyed without accompanying CMI and without permission. That harmed Publishers, who rely on CMI information to protect their copyright interests in their works.

179. The stripping of CMI from the Publishers' works benefitted OpenAI—and upon information and belief, OpenAI intended to receive that benefit. For one, had CMI not been removed, OpenAI's products would be trained on associating CMI with the text from the Publishers' work. They would be more likely to deliver outputs that contain CMI, such as copyright and term of use information. Such a characteristic would create hesitation among users who, upon seeing that CMI associated with the outputs, would become concerned about unlawful use. For another, removing CMI in the training dataset would conceal Defendants' infringing

activity because the outputs dissociated from CMI would not tip off the public, and importantly the Publishers, that Defendants had illegally removed CMI in prior steps.

180. Moreover, OpenAI expressly contemplated the possibility that their customers "use or distribution of Output" may "infringe[] a third party's intellectual property right." In its Service Terms, OpenAI indemnifies its API customers for such infringements.[37] The provision exempts certain scenarios where the customer or customer end user took other actions against OpenAI policy. The commitment to paying costs and indemnifying against liability—all while it refuses to compensate those whose works it has pirated—demonstrates that OpenAI understood that even if customers are using ChatGPT according to OpenAI's own terms, they may still be engaging in infringing activity.

181. Thus, OpenAI knew that its removal of the Publishers' CMI in this manner would induce, enable, conceal, or facilitate infringement by end-users of the GenAI products.

E. Defendants' Illegal Conduct Enriched Themselves at the Expense of the Publishers.

182. The Publishers have suffered and will continue to suffer concrete harm as a result of Defendants' unlawful conduct.

183. In creating and supporting local news, the Publishers expend untold hours and many millions of dollars to develop news-gathering, storytelling, community engagement, and information-conveying capabilities. Such activities are not only private, business interests: they also deliver a crucial public good in our democratic society.

184. The Publishers' economic viability to undertake these actions depends on readers who purchase and renew subscriptions to the Publishers' products or otherwise engage with the Publishers' content and advertising. That includes accessing the Publishers' print publications,

[37] OpenAI, *Service terms*(last updated Jun. 12, 2026), https://openai.com/policies/service-terms.

websites—with and without paywalls—mobile applications, newsletters, e-editions, and other services. Some Publishers also license or otherwise agree to provide their content to other entities, such as other media entities, museums, and libraries under contractual agreements. When they do so, these Publishers receive proper renumeration for the use of their works, and such agreements come with clear parameters around the display and use of the works.

185. Whenever prospective readers access the Publishers' works through alternate means—without any compensation to the Publishers—that harms the Publishers' interests in several ways.

186. First, it directly diminishes revenue from subscriptions, licensing agreements, and other direct contracts by siphoning readers.

187. Second, it diminishes advertising revenue by reducing the readership base when readers are less likely to read print periodicals or visit the Publishers' websites and other publication displays.

188. Third, it disincentivizes third parties such as licensers and advertisers from entering into agreements with the Publishers.

189. Fourth, it damages the Publishers' ability to attract and retain journalism talent, creating a cyclical effect on Publishers' business and further eroding the local media industry upon which this nation relies.

190. The above harms to the Publishers come at an acute time in the journalism industry. Local journalism operates in a fragile economic environment. But its impact on our civil society and our democracy cannot be overstated. As a Democracy Fund analysis of numerous studies

demonstrates, local journalism has demonstrably increased civic participation, produced greater cohesiveness in communities, and reduced public corruption.[38]

191. Moreover, at a time when the American public is deeply divided, local news remains a beacon of hope and a uniquely trusted source of information. According to the Pew Research Center, 71% of Americans believe their local news outlets are doing a good job at reporting the news accurately. Sixty-two percent believe local news is effective at dealing fairly with all sides.[39]

192. Defendants' illegal and unauthorized use of the Publishers' works as described above wreaks havoc on the viability of local journalism. By taking the Publisher's works without attribution or compensation, Defendants threaten the Publishers' business model in fundamental ways.

193. By contrast, Defendants have reaped astronomical financial benefits off the backs of the Publishers.

194. In March 2026, OpenAI announced that it has a post money valuation of $852 billion.[40] It touts $2 billion in revenue per month and over 900 million active users weekly.[41] Its clients include millions of individuals and thousands of companies, including nearly all Fortune 500 companies, and many on paying subscriptions.[42]

---

[38] Josh Stearns & Christine Schmidt, *How We Know Journalism is Good for Democracy*, Democracy Fund (Sept. 15, 2022), https://democracyfund.org/idea/how-we-know-journalism-is-good-for-democracy.
[39] *See* John Gramlich, Q&*A: What Pew Research Center's new survey says about local news in the U.S.*, Pew Research Ctr. (Mar. 16, 2019), https://www.pewresearch.org/short-reads/2019/03/26/qa-what-pew-research-centers-new-survey-says-about-local-news-in-the-u-s.
[40] *See* OpenAI, *OpenAI raises $122 billion to accelerate the next phase of AI* (Mar. 31, 2026), https://openai.com/index/accelerating-the-next-phase-ai..
[41] *See id.*
[42] *See* Sameer Khan, *50 ChatGPT Statistics Every Business Leader Should Know in 2026*, AI Business Weekly (Feb. 28, 2026), https://aibusinessweekly.net/p/chatgpt-statistics

195. In June 2026, OpenAI filed registration statements ahead of an IPO with the Securities and Exchange Commission.[43] Some analysts suggest OpenAI's IPO valuation could exceed $1 trillion, a number that exceeds the 2025 gross domestic product of all but 21 countries.[44]

196. Microsoft's investment in OpenAI has proven extremely lucrative. After its initial $1 billion investment in 2019, Microsoft reportedly added another $12 billion to its investment, rendering its stake in the business to be approximately 27%. As noted above, OpenAI is preparing one of the most lucrative IPOs in history, and Microsoft is its biggest shareholder.

197. Moreover, Microsoft's own products have stood to gain from the enterprise. Its Bing search engine's popularity rose after it integrated ChatGPT. And Microsoft also uses ChatGPT in its Copilot product, which is streamlined with its Office 365 premium products subscription. Microsoft's reported revenue in March 2026 was $82.9 billion, nearly a 20% increase year over year.[45]

198. Defendants' successes would not have been possible without the use of copyrighted material such as the Publishers' works.

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

On Behalf of Plaintiffs Arkansas Democrat-Gazette, Inc.; Concord Publishing House, Inc.; H.S. Gere & Sons, Inc.; The New Mexican Inc.; and Newspapers of New Hampshire, Inc. Against All Defendants

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

---

[43] Adam Levy, *OpenAI Just Took the First Step Towards Its IPO. Here's How to Invest in the Stock Now.*, The Motley Fool (Jun. 14, 2026), https://www.fool.com/investing/2026/06/14/openai-just-took-the-first-step-toward-its-ipo-her.
[44] Douglas A. McIntyre, *OpenAI's $1 Trillion IPO*, Yahoo! Finance (Oct. 30, 2025), https://finance.yahoo.com/news/openai-1-trillion-ipo-141519224.html.
[45] Microsoft, *Microsoft Cloud and AI strength fuels third quarter results* (Apr. 29, 2026), https://news.microsoft.com/source/2026/04/29/microsoft-cloud-and-ai-strength-fuels-third-quarter-results.

200. Plaintiffs Arkansas Democrat-Gazette, Inc.; Concord Publishing House, Inc.; H.S. Gere & Sons, Inc.; The New Mexican Inc.; and Newspapers of New Hampshire, Inc. are the owners of the registered copyrights, including those listed in Exhibits A-E ("Registration Publishers"). The Publishers reserve the right to revise or supplement these exhibits if it becomes clear during discovery that additional registered works of any Publishers were also copied by Defendants.

201. Registration Publishers have a business practice of publishing in electronic format on their respective websites articles that also appears in that newspaper's print edition. It has additionally been the business practice of these Publishers to publish in electronic format on those Publishers' respective websites articles from older print editions of those newspapers. The electronic versions of the articles are substantially the same as their print-edition counterparts.

202. As the owners of the registered copyrights in these works, the Registration Publishers hold the exclusive rights to the works under 17 U.S.C. § 106.

203. Electronic versions of the articles published by the Registered Publishers were copied to train Defendants' GPT models and, in many cases, have been distributed by and encoded within Defendants' GPT models.

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

205. By illegally storing, processing, and reproducing the training datasets containing the Registered Publishers' works to train the GPT models on Microsoft's supercomputing platform,

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

206. On information and belief, by storing, processing, and reproducing the GPT models trained on the Registered Publishers' works, which GPT models themselves have memorized, on Microsoft's supercomputing platform, Microsoft and the OpenAI Defendants have jointly directly infringed the Registered Publishers' works exclusive rights in their copyrighted works.

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

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

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

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

211. As a direct and proximate result of Defendants' infringing conduct alleged herein, the Registered 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 Registered Publishers' works. The Registered Publishers therefore are entitled to permanent injunctive relief restraining and enjoining Defendants' ongoing infringing conduct.

212. The Registered Publishers are further entitled to recover statutory damages, actual damages, restitution of profits, attorney's fees, and other remedies provided by law.

COUNT II: Vicarious Copyright Infringement

On Behalf of Plaintiffs Arkansas Democrat-Gazette, Inc.; Concord Publishing House, Inc.; H.S. Gere & Sons, Inc.; The New Mexican Inc.; and Newspapers of New Hampshire, Inc. Against Microsoft, OpenAI, Inc., OpenAI, GP, OpenAI LP, OAI Corporation, LLC, OpenAI Holdings, LLC, OpenAI Global, LLC, and OpenAI Global PBC

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

214. 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 Registered Publishers' works, the GPT models, and OpenAI's ChatGPT offerings. Microsoft had the ability to control the infringing activity perpetrated by the OpenAI defendants, but failed to do so. Microsoft profited from the infringement perpetrated by the OpenAI Defendants by incorporating the infringing GPT models trained on the Registered Publishers' works into its own product offerings, including Copilot (formerly known as Bing Chat).

215. Defendants OpenAI, Inc.; OpenAI, GP; OAI Corporation, LLC; OpenAI Holdings, LLC; and Microsoft controlled, directed, and benefitted 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 Registered Publishers' works.

216. Defendants OpenAI Global, LLC and OpenAI LP directed, controlled, and benefitted from the infringement perpetrated by Defendants OpenAI OpCo, LLC and OpenAI, LLC, including the reproduction and distribution of the Registered Publishers' works.

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

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

On Behalf of All Plaintiffs Against the OpenAI Defendants

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

219. The Publishers included one or more forms of CMI 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.

220. Without the Publishers' permission, OpenAI copied the Publishers' works and used them as training data for their GenAI models.

221. On information and belief, OpenAI removed the Publishers' CMI in building the training datasets containing copies of the Publishers' works, including removing the Publishers' CMI from the Publishers' works scraped directly from the Publishers' websites and removing the Publishers' CMI from the Publishers' works reproduced from third-party datasets.

222. On information and belief, OpenAI removed the Publishers' CMI through the generation of synthetic search results, including removing Publishers' CMI when scraping the Publishers' works from the Publishers' websites and generating copies or derivatives of the Publishers' works as the output of ChatGPT offerings.

223. On information and belief, OpenAI removed the Publishers' CMI in generating outputs from the GPT models containing copies or derivatives of the Publishers' works.

224. By design, OpenAI's GPT-based products do not preserve anyCMI, 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, OpenAI intentionally removed CMI from the Publishers' works in violation of 17 U.S.C. § 1202(b)(1).

225. OpenAI's removal or alteration of the Publishers' CMI was done knowingly and with the intent to induce, enable, facilitate, or conceal OpenAI's or end-users' infringement of the Publishers' copyrights.

226. OpenAI knew or had reasonable grounds to know that their removal of CMI 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.

227. The Publishers have been injured by OpenAI's removal of CMI. The Publishers are entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law, including full costs and attorney's fees.

PRAYER FOR RELIEF

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

i. Awarding the Publishers statutory damages, including willful infringement damages, compensatory damages, restitution, disgorgement, and any other relief that may be permitted by law or equity;

ii. Permanently enjoining Defendants, their agents and employees, and all person acting in concert or participation with Defendants, from the unlawful, unfair, and infringing conduct alleged herein;

iii. An injunction under 17 U.S.C. § 503(b) requiring Defendants to remove all copies of Registered Works from all GPT or other LLM models and training sets;

iv. An award of costs, expenses, and attorney's fees as permitted by law; and

v. 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: June 24, 2026

By: /s/ *Matthew J. Platkin*

Matthew J. Platkin (#MP0621)
Angela Cai (#AC2014)
Ravi Ramanathan (#RR4590)
Aaron E. Haier (#5787999)
Conor Bradley*
PLATKIN LLP
413 Washington Ave.
Unit 174
Belleville, NJ 07109
Phone: (973) 561-1951
Email: mplatkin@platkinllp.com

Attorneys for Plaintiffs

**pro hac vice application forthcoming*