UNITED STATES DISTRICT COURT
SOUTHERN DISTRICT OF NEW YORK

THE NEW YORK TIMES COMPANY Plaintiff, v. MICROSOFT CORPORATION, OPENAI, INC., OPENAI LP, OPENAI GP, LLC, OPENAI, LLC, OPENAI OPCO LLC, OPENAI GLOBAL LLC, OAI CORPORATION, LLC, and OPENAI HOLDINGS, LLC, Defendants.	Civil Action No. 1:23-cv-11195-SHS **THIRD AMENDED COMPLAINT** **JURY TRIAL DEMANDED**

Plaintiff The New York Times Company ("The Times"), by its attorneys Susman Godfrey LLP and Rothwell, Figg, Ernst & Manbeck, P.C., for its Third Amended 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, (collectively "OpenAI" and, with Microsoft, "Defendants"), alleges as follows:

I. NATURE OF THE ACTION

1. Independent journalism is vital to our democracy. It is also increasingly rare and valuable. For more than 170 years, The Times has given the world deeply reported, expert, independent journalism. Times journalists go where the story is, often at great risk and cost, to inform the public about important and pressing issues. They bear witness to conflict and disasters, provide accountability for the use of power, and illuminate truths that would otherwise go unseen. Their essential work is made possible through the efforts of a large and expensive organization that provides legal, security, and operational support, as well as editors who ensure their journalism meets the highest standards of accuracy and fairness. This work has always been important. But

within a damaged information ecosystem that is awash in unreliable content, The Times's journalism provides a service that has grown even more valuable to the public by supplying trustworthy information, news analysis, and commentary.

2. Defendants' unlawful use of The Times's work to create artificial intelligence products that compete with it threatens The Times's ability to provide that service. Defendants' generative artificial intelligence ("GenAI") tools rely on large-language models ("LLMs") that were built by copying and using *millions* of The Times's copyrighted news articles, in-depth investigations, opinion pieces, reviews, how-to guides, and more. While Defendants engaged in widescale copying from many sources, they gave Times content particular emphasis when building their LLMs—revealing a preference that recognizes the value of those works. Through Microsoft's Bing Chat (recently rebranded as "Copilot") and OpenAI's ChatGPT, Defendants seek to free-ride on The Times's massive investment in its journalism by using it to build substitutive products without permission or payment.

3. The Constitution and the Copyright Act recognize the critical importance of giving creators exclusive rights over their works. Since our nation's founding, strong copyright protection has empowered those who gather and report news to secure the fruits of their labor and investment. Copyright law protects The Times's expressive, original journalism, including, but not limited to, its millions of articles that have registered copyrights.

4. Defendants have refused to recognize this protection. Powered by LLMs containing copies of Times content, Defendants' GenAI tools can generate output that recites Times content verbatim, closely summarizes it, and mimics its expressive style, as demonstrated by scores of examples. *See* Exhibit J. These tools also wrongly attribute false information to The Times.

5. Defendants also use Microsoft's Bing search index, which copies and categorizes The Times's online content, to generate responses that contain verbatim excerpts and detailed summaries of Times articles that are significantly longer and more detailed than those returned by traditional search engines. By providing Times content without The Times's permission or authorization, Defendants' tools undermine and damage The Times's relationship with its readers and deprive The Times of subscription, licensing, advertising, and affiliate revenue.

6. Using the valuable intellectual property of others in these ways without paying for it has been extremely lucrative for Defendants. Microsoft's deployment of Times-trained LLMs throughout its product line helped boost its market capitalization by a trillion dollars in the past year alone. And OpenAI's release of ChatGPT has driven its valuation to as high as $90 billion. Defendants' GenAI business interests are deeply intertwined, with Microsoft recently highlighting that its use of OpenAI's "best-in-class frontier models" has generated customers—including "leading AI startups"—for Microsoft's Azure AI product.[1]

7. The Times objected after it discovered that Defendants were using Times content without permission to develop their models and tools. For months, The Times has attempted to reach a negotiated agreement with Defendants, in accordance with its history of working productively with large technology platforms to permit the use of its content in new digital products (including the news products developed by Google, Meta, and Apple). The Times's goal during these negotiations was to ensure it received fair value for the use of its content, facilitate the continuation of a healthy news ecosystem, and help develop GenAI technology in a responsible way that benefits society and supports a well-informed public.

[1] *Microsoft Fiscal Year 2024 First Quarter Earnings Conference Call*, MICROSOFT INVESTOR RELATIONS (Oct. 24, 2023), https://www.microsoft.com/en-us/Investor/events/FY-2024/earnings-fy-2024-q1.aspx.

8. These negotiations have not led to a resolution. Publicly, Defendants insist that their conduct is protected as "fair use" because their unlicensed use of copyrighted content to train GenAI models serves a new "transformative" purpose. But there is nothing "transformative" about using The Times's content without payment to create products that substitute for The Times and steal audiences away from it. Because the outputs of Defendants' GenAI models compete with and closely mimic the inputs used to train them, copying Times works for that purpose is not fair use.

9. The law does not permit the kind of systematic and competitive infringement that Defendants have committed. This action seeks to hold them responsible for the billions of dollars in statutory and actual damages that they owe for the unlawful copying and use of The Times's uniquely valuable works.

II. JURISDICTION AND VENUE

10. 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.

11. Jurisdiction over Microsoft and OpenAI is proper because they have purposely availed themselves of the privilege of conducting business in New York. A substantial portion of Microsoft and OpenAI's widespread infringement and other unlawful conduct alleged herein occurred in New York, including the distribution and sales of Microsoft and OpenAI's Generative Pre-training Transformer ("GPT")-based products like ChatGPT, ChatGPT Enterprise, Bing Chat, 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.

12. Because The Times's principal place of business and headquarters is in this District, the injuries alleged herein from Microsoft and OpenAI's widespread infringement and other unlawful conduct foreseeably occurred in this District.

13. 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 Times's claims occurred in this District, including the marketing, sales, and licensing of Defendants' GenAI products built on the infringement of The Times's 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 New York. OpenAI has licensed its GPT models to New York residents and companies headquartered in New York. For example, in 2023, OpenAI struck deals to license its GPT models to the Associated Press (AP) and Morgan Stanley, both companies headquartered in New York.

III. THE PARTIES

14. Plaintiff The New York Times Company is a New York corporation with its headquarters and principal place of business in New York. The Times publishes digital and print products, including its core news product, The New York Times, which is available on its mobile applications, on its website (NYTimes.com), and as a printed newspaper, and associated content such as its podcasts. The Times also publishes other interest-specific publications, including The Athletic (sports media), Cooking (recipes and other cooking-related content), Games (puzzles and games), and Wirecutter (shopping recommendations). The Times owns over 10 million registered,

copyrighted works, including those set forth in the incorporated Exhibits ("Times Works"). *See* ECF Nos. 170–202.

15. 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.

16. Microsoft has described its relationship with the OpenAI Defendants as a "partnership." This partnership has included contributing and operating the cloud computing services used to copy Times 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 Times Works. Microsoft uses these models to provide infringing content and, at times, misinformation to users of its products and online services. During a quarterly earnings call in October 2023, Microsoft noted that "more than 18,000 organizations now use Azure OpenAI Service, including new-to-Azure customers."

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

18. 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. OpenAI Inc. indirectly owns and controls all other OpenAI entities and has been directly involved in perpetrating the mass infringement and other unlawful conduct alleged here.

19. 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 Times Works alleged here.

20. 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 Times Works alleged here through its direction and control of OpenAI LP and OpenAI Global LLC.

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

22. 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 Times Works through its management and direction of OpenAI, LLC.

23. 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 unlawful conduct alleged herein through its ownership, control, and direction of OpenAI LLC.

24. 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.

25. 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.

IV. FACTUAL ALLEGATIONS

A. The New York Times and its Mission

1. *Almost Two Centuries of High-Quality, Original, Independent News*

26. The New York Times is a trusted source of quality, independent journalism whose mission is to seek the truth and help people understand the world. Begun as a small, local newspaper, The Times has evolved to a diversified multi-media company with readers, listeners, and viewers around the globe. Today, more than 10 million subscribers pay for Times journalism,

which includes everything from news to opinion, culture to business, cooking to games, and shopping recommendations to sports.

27. Founded in 1851, The New York Times has a long history of providing the public with independent journalism of the highest quality. When Adolph Ochs bought the newspaper out of bankruptcy in 1896, he vowed that The Times would be fiercely independent, dedicated to journalism of the highest integrity, and devoted to the public welfare. He articulated the vision: "To give the news impartially, without fear or favor, regardless of any party, sect, or interest involved." These words still animate The New York Times today, nearly two centuries later.

28. Producing original independent journalism is at the heart of this mission. Times journalists cover the most important stories across the globe; in a typical year, The Times sends journalists to report on the ground from more than 160 countries. Together, along with editors, photographers, audio producers, videographers, graphic designers, data analysts, and more, The Times's newsroom produces groundbreaking journalism across every major storytelling format.

29. The quality of The Times's coverage has been widely recognized with many industry and peer accolades, including 135 Pulitzer Prizes since its first Pulitzer award in 1918 (nearly twice as many as any other organization). The Times's journalism is also deeply impactful. Academics, teachers, and scientists have used it to educate and innovate. Lawmakers have cited it to introduce legislation. Judges have referenced it in rulings. And tens of millions of people rely on it every day.

30. Times journalists are experts in their subject matter and among the most experienced and talented in the industry. In many cases, their work is enhanced by professional expertise: lawyers cover the court, doctors cover health care, and veterans cover the military. Many Times journalists draw on decades of experience. One reporter covering the White House, for

example, has reported on five administrations. His colleague, a White House photographer, has covered seven.

31. In addition to journalists who spend considerable time and effort reporting pieces, The Times employs hundreds of editors to painstakingly review its journalism for accuracy, independence, and fairness, with at least two editors reviewing each piece prior to publication and many more reviewing the most important and sensitive pieces. The Times also has among the largest and most robust Standards teams in the industry, which advises the newsroom daily on consistency, accuracy, fairness, and clarity in its reporting and maintains stringent ethical guidelines for journalists and their work. The Times also maintains an internal Stylebook, a document that is updated over time to guide the tone of its journalism and the prose used. There is also an ongoing dialogue among journalists and editors to ensure The Times fairly and thoroughly covers the right stories and presents what it finds in a clear and compelling way. Producing Times journalism is a creative and deeply human endeavor.

2. ***Groundbreaking, In-Depth Journalism and Breaking News at Great Cost***

32. To produce world-class journalism, The Times invests an enormous amount of time, money, expertise, and talent, both in its newsroom and product, technology, and other supporting teams. Core areas of focus include:

33. **Investigative Reporting.** The Times does deep investigations—which usually take months and sometimes years to report and produce—into complex and important areas of public interest. The Times's reporters routinely uncover stories that would otherwise never come to light. They have exposed problems, held power to account, and demanded the public's attention. In investigating these areas, Times coverage often results in meaningful reforms. These stories are

written and edited in the style that is widely associated with The Times, one that readers trust and seek out.

34. **Breaking News Reporting.** The Times is equally committed to quickly and accurately reporting breaking news. In an era in which speculation, disinformation, and spin often drown out the truth when news breaks, The Times fills an important need for trustworthy news with journalists who have the subject-matter expertise, news judgment, and sources required to report the facts in a compelling way. In the past two years, The Times has provided detailed, real-time coverage on breaking news across a range of topics, including the upcoming U.S. elections, multiple mass shootings including those in Maine and Nashville, wars in Ukraine and the Middle East, a spate of natural disasters around the globe, and the collapse of major regional banks.

35. **Beat Reporting:** The Times invests significantly in its beat reporting by giving its beat reporters the time and space to go deep on a single topic. At The Times, these topics vary from public health to religion to architecture, and from the Pentagon to Hollywood to Wall Street. They also include The Times's dozens of national and international bureaus, where correspondents are steeped in the communities they cover. Because this type of journalism is grounded in the expertise and deep connections of Times journalists, beat coverage enriches The Times's reporting.

36. **Reviews and Analysis.** The Times is a trusted source for reviews and analysis of arts and culture, including food, books, art, film, theater, television, music, fashion, and travel. In 2016, it acquired the product review site Wirecutter, which recommends the best products in dozens of categories including home goods, technology, health and fitness, and more. Each year, Wirecutter spends tens of thousands of hours conducting rigorous testing and research to produce a catalog of reviews that today covers thousands of products.

37. **Commentary and Opinion.** The Times publishes opinion articles that contribute to public debate across the world. Many of these articles come from The Times's staff of world-renowned columnists. Additionally, leaders in business, politics, religion, education, and the arts write guest essays for The Times's opinion section, giving readers the opportunity to understand a wide range of experiences, perspectives, and ideas about the most important issues of the day.

3. ***A Commitment to Quality Journalism***

38. It takes enormous resources to publish, on average, more than 250 original articles every day. Many of these articles take months—and sometimes longer—to report. That output is the work of approximately 5,800 full-time equivalent Times employees (as of December 31, 2022), some 2,600 of whom are directly involved in The Times's journalism operations.

39. Quite often, the most vital news reporting for society is the most resource-intensive. Some of The Times's most important journalism requires deploying teams of journalists at great cost to report on the ground around the world, providing best-in-class security and support, filing lawsuits against government entities to bring information to light, and supporting journalists through investigations that can take months or years.

40. Subscription, advertising, licensing, and affiliate revenue make this reporting possible. In 1996, The Times launched a core news website, alongside its paid print edition, that was free. As readers shifted from print news to digital products, The Times—like most print publishers—faced the prospect of not being able to continue funding its journalism. In response, The Times reinvented its business model to incorporate digital subscriptions. The Times launched its metered paywall in 2011, in what it called "a bet that readers will pay for news they are accustomed to getting free."[2]

[2] Jeremy W. Peters, *The Times Announces Digital Subscription Plan*, N.Y. TIMES (Mar. 17, 2011), https://www.nytimes.com/2011/03/18/business/media/18times.html.

41. Thanks to the quality of The Times's journalism, that strategic innovation paid off, which allowed The Times to continue to exist and to thrive. Today, the vast majority of subscribers are digital-only. In the 14 years since The Times launched its paywall, it has grown its paid digital subscribership and developed a direct relationship with its online audience through its tireless commitment to making journalism "worth paying for." Generating and maintaining direct traffic to its online content and mobile applications are critical components of The Times's financial success.

42. By the third quarter of 2023, The Times had nearly 10.1 million digital and print subscribers worldwide. The Times aims to have 15 million subscribers by year-end 2027.

43. The Times makes journalism "worth paying for" by publishing articles that are exhaustively researched and reported, thoughtfully written, carefully edited, and thoroughly fact-checked.

44. In addition, The Times has deepened its relationship with its readers by expanding its offerings to better encompass its readers' specific interests, including best-in-class offerings like Cooking, Wirecutter, Games, and The Athletic.

45. The Times's paywall does not require payment for *all* access to The Times's content. To build audience engagement and loyalty, The Times's access model generally offers registered users free access to a limited number of articles and other content before requiring them to subscribe for access to additional content. Approximately 50 to 100 million users, on average, engage with The Times's digital content each week. This traffic is a key source of advertising revenue and helps drive future subscriptions to The Times.

46. The Times also compiled digital archives of all its material going back to its founding, at significant cost. Its digital archives include The New York Times Article Archive, with

partial and full-text digital versions of articles from 1851 to today, and the TimesMachine, a browser-based digital replica of all issues from 1851 to 2002. This represents a singular database of contemporaneous language and information, as well as a unique and valuable historical record. The Times also provides its own API that allows researchers and academics to search Times content for non-commercial purposes.

4. ***GenAI Products Threaten High-Quality Journalism***

47. Making great journalism is harder than ever. Over the past two decades, the traditional business models that supported quality journalism have collapsed, forcing the shuttering of newspapers all over the country. It has become more difficult for the public to sort fact from fiction in today's information ecosystem, as misinformation floods the internet, television, and other media. If The Times and other news organizations cannot produce and protect their independent journalism, there will be a vacuum that no computer or artificial intelligence can fill.

48. The protection of The Times's intellectual property is critical to its continued ability to fund world-class journalism in the public interest. If The Times and its peers cannot control the use of their content, their ability to monetize that content will be harmed. With less revenue, news organizations will have fewer journalists able to dedicate time and resources to important, in-depth stories, which creates a risk that those stories will go untold. Less journalism will be produced, and the cost to society will be enormous.

49. The Times depends on its exclusive rights of reproduction, adaptation, publication, performance, and display under copyright law to resist these forces. The Times has registered the copyright in its print edition every day for over 100 years, maintains a paywall, and has implemented terms of service that set limits on the copying and use of its content. To use Times

content for commercial purposes, a party should first approach The Times about a licensing agreement.

50. The Times requires third parties to obtain permission before using Times content and trademarks for commercial purposes, and for decades The Times has licensed its content under negotiated licensing agreements. These agreements help ensure that The Times controls how, where, and for how long its content and brand appears and that it receives fair compensation for third-party use. Third parties, including large tech platforms, pay The Times significant royalties under these agreements in exchange for the right to use Times content for narrowly defined purposes. The agreements prohibit uses beyond those authorized purposes.

51. Times content is also available for licenses for certain uses through the Copyright Clearance Center ("CCC"), a clearinghouse that licenses material to both corporate and academic users. Through the CCC, The Times permits limited licenses for instruction, academic, other nonprofit uses, and limited commercial uses. For example, a for-profit business can acquire a CCC license to make a photocopy of Times content for internal or external distribution in exchange for a licensing fee of about ten dollars per article. A CCC license to post a single Times article on a commercial website for up to a year costs several thousand dollars.

52. The Times's ability to continue to attract and grow its digital subscriber base and to generate digital advertising revenue depends on the size of The Times's audience and users' sustained engagement directly with The Times's websites and mobile applications. To facilitate this direct engagement with its products, The Times permits search engines to access and index its content, which is necessary to allow users to find The Times using these search engines. Inherent in this value exchange is the idea that the search engines will direct users to The Times's own

websites and mobile applications, rather than exploit The Times's content to keep users within their own search ecosystem.

53. While The Times, like virtually all online publishers, permits search engines to access its content for the limited purpose of surfacing it in traditional search results, The Times has never given permission to any entity, including Defendants, to use its content for GenAI purposes.

54. The Times reached out to Microsoft and OpenAI in April 2023 to raise intellectual property concerns and explore the possibility of an amicable resolution, with commercial terms and technological guardrails that would allow a mutually beneficial value exchange between Defendants and The Times. These efforts have not produced a resolution.

B. Defendants' GenAI Products

1. *A Business Model Based on Mass Copyright Infringement*

55. OpenAI was formed in December 2015 as a "non-profit artificial intelligence research company." OpenAI started with $1 billion in seed money from its founders, a group of some of the wealthiest technology entrepreneurs and investors and companies like Amazon Web Services and InfoSys. This group included Elon Musk, the CEO of Tesla and X Corp. (formerly known as Twitter); Reid Hoffman, the co-founder of LinkedIn; Sam Altman, the former president of Y Combinator; and Greg Brockman, the former Chief Technology Officer of Stripe.

56. Despite accepting very large investments from enormously wealthy companies and individuals at its founding, OpenAI originally maintained that its research and work would be entirely unmotivated by profit. In a December 11, 2015, press release, Brockman and co-founder Ilya Sutskever (now OpenAI's President and Chief Scientist, respectively) wrote: "Our goal is to advance digital intelligence in the way that is most likely to benefit humanity as a whole,

unconstrained by a need to generate financial return. Since our research is free from financial obligations, we can better focus on a positive human impact." In accordance with that mission, OpenAI promised that its work and intellectual property would be open and available to the public, that its "[r]esearchers will be strongly encouraged to publish their work, whether as papers, blog posts, or code" and that its "patents (if any) will be shared with the world."

57. Despite its early promises of altruism, OpenAI quickly became a multi-billion-dollar for-profit business built in large part on the unlicensed exploitation of copyrighted works belonging to The Times and others. Just three years after its founding, OpenAI shed its exclusively nonprofit status. It created OpenAI LP in March 2019, a for-profit company dedicated to conducting the lion's share of OpenAI's operations—including product development—and to raising capital from investors seeking a return. OpenAI's corporate structure grew into an intricate web of for-profit holding, operating, and shell companies that manage OpenAI's day-to-day operations and grant OpenAI's investors (most prominently, Microsoft) authority and influence over OpenAI's operations, all while raising billions in capital from investors. The result: OpenAI today is a commercial enterprise valued as high as $90 billion, with revenues projected to be over $1 billion in 2024.

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

59. GPT-3.5 and GPT-4 are both orders of magnitude more powerful than the two previous generations, yet Defendants have kept their design and training entirely a secret. For

previous generations, OpenAI had voluminous reports detailing the contents of the training set, design, and hardware of the LLMs. Not so for GPT-3.5 or GPT-4. For GPT-4, for example, the "technical report" that OpenAI released said: "this report contains no further details about the architecture (including model size), hardware, training compute, dataset construction, training method, or similar."[3]

60. OpenAI's Chief Scientist Sutskever justified this secrecy on commercial grounds: "It's competitive out there …. And there are many companies who want to do the same thing, so from a competitive side, you can see this as maturation of the field."[4] But its effect was to conceal the identity of the data OpenAI copied to train its latest models from rightsholders like The Times.

61. OpenAI became a household name upon the release of ChatGPT in November 2022. ChatGPT is a text-generating chatbot that, given user-generated prompts, can mimic human-like natural language responses. ChatGPT was an instant viral sensation, reaching one million users within a month of its release and gaining over 100 million users within three months.

62. OpenAI, through OpenAI OpCo LLC and at the direction of OpenAI Inc., OpenAI LP, and other OpenAI entities, offers a suite of services powered by its LLMs, targeted to both ordinary consumers and businesses. A version of ChatGPT powered by GPT-3.5 is available to users for free. OpenAI also offers a premium service, powered by OpenAI's "most capable model" GPT-4, to consumers for $20 per month. OpenAI's business-focused offerings include ChatGPT Enterprise and ChatGPT API tools designed to enable developers to incorporate ChatGPT into bespoke applications. OpenAI also licenses its technology to corporate clients for licensing fees.

[3] OPENAI, GPT-4 TECHNICAL REPORT (2023), https://cdn.openai.com/papers/gpt-4.pdf.

[4] James Vincent, *OpenAI Co-Founder on Company's Past Approach to Openly Sharing Research: 'We Were Wrong'*, THE VERGE (Mar. 15, 2023), https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview.

63. These commercial offerings have been immensely valuable for OpenAI. Over 80% of Fortune 500 companies are using ChatGPT.[5] According to recent reports, OpenAI is generating revenues of $80 million per month, and is on track to surpass over $1 billion within the next 12 months.[6]

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

65. Upon information and belief, all of the OpenAI Defendants have been either directly involved in or have directed, controlled, and profited from OpenAI's widespread infringement and commercial exploitation of Times Works. OpenAI Inc., alongside Microsoft, controlled and directed the widespread reproduction, distribution, and commercial use of The Times's 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 reproduction, distribution, and commercial use of Times Works. OpenAI LP and OpenAI Global LLC also controlled and directed OpenAI, LLC and OpenAI OpCo LLC, which were involved in distributing, selling, and licensing OpenAI's GPT-based products, and thus monetized the reproduction, distribution, and commercial use of Times Works.

---

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

[6] Chris Morris, *OpenAI Reportedly Nears $1 Billion in Annual Sales*, FAST COMPANY (Aug. 30, 2023), https://www.fastcompany.com/90946849/openai-chatgpt-reportedly-nears-1-billion-annual-sales.

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

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

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

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

> But beneath what OpenAI is putting out as large models, remember, the heavy lifting was done by the [Microsoft] Azure team to build the computer infrastructure. Because these workloads are so different than anything that's come before. So we needed to

> completely rethink even the datacenter up to the infrastructure that first gave us even a shot to build the models. And now we're translating the models into products.[7]

70. Microsoft built this supercomputer "in collaboration with and exclusively for OpenAI," and "designed [it] specifically to train that company's AI models."[8] Even by supercomputing standards, it was unusually complex. According to Microsoft, it operated as "a single system with more than 285,000 CPU cores, 10,000 GPUs and 400 gigabits per second of network connectivity for each GPU server." This system ranked in the top five most powerful publicly known supercomputing systems in the world.

71. To ensure that the supercomputing system suited OpenAI's needs, Microsoft needed to test the system, both independently and in collaboration with OpenAI software engineers. According to Mr. Nadella, with respect to OpenAI: "They do the foundation models, and we [Microsoft] do a lot of work around them, including the tooling around responsible AI and AI safety." Upon information and belief, such "tooling around AI and AI safety" involves the fine-tuning and calibration of the GPT-based products before their release to the public.[9]

72. In collaboration with OpenAI, Microsoft has also commercialized OpenAI's GPT-based technology, and combined it with its own Bing search index. In February 2023, Microsoft unveiled Bing Chat, a generative AI chatbot feature on its search engine powered by GPT-4. In May 2023, Microsoft and OpenAI unveiled "Browse with Bing," a plugin to ChatGPT that enabled it to access the latest content on the internet through the Microsoft Bing search engine. Bing Chat

---

[7] *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.

[8] 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/.

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

and Browse with Bing combine GPT-4's ability to mimic human expression—including The Times's expression—with the ability to generate natural language summaries of search result contents, including hits on Times Works, that obviate the need to visit The Times's own websites. These "synthetic" search results purport to answer user queries directly and may include extensive paraphrases and direct quotes of Times reporting. Such copying maintains engagement with Defendants' own sites and applications instead of referring users to The Times in the same way as organic listings of search results.

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

> [W]e were very confident in our own ability. We have all the IP rights and all the capability. If OpenAI disappeared tomorrow, I don't want any customer of ours to be worried about it quite honestly, because we have all of the rights to continue the innovation. Not just to serve the product, but we can go and just do what we were doing in partnership ourselves. We have the people, we have the compute, we have the data, we have everything.

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

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

2. *How GenAI Models Work*

75. At the heart of Defendants' GenAI products is a computer program called a "large language model," or "LLM." The different versions of GPT are examples of LLMs. An LLM works by predicting words that are likely to follow a given string of text based on the potentially billions of examples used to train it.

76. Appending the output of an LLM to its input and feeding it back into the model produces sentences and paragraphs word by word. This is how ChatGPT and Bing Chat generate responses to user queries, or "prompts."

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

78. The process of setting the values for an LLM's parameters is called "training." It involves storing encoded copies of the training works in computer memory, repeatedly passing them through the model with words masked out, and adjusting the parameters to minimize the difference between the masked-out words and the words that the model predicts to fill them in.

79. After being trained on a general corpus, models may be further subject to "fine-tuning" by, for example, performing additional rounds of training using specific types of works to better mimic their content or style, or providing them with human feedback to reinforce desired or suppress undesired behaviors.

80. Models trained in this way are known to exhibit a behavior called "memorization."[10] That is, given the right prompt, they will repeat large portions of materials they

[10] GERRIT J.J. VAN DEN BURG & CHRISTOPHER K.I. WILLIAMS, ON MEMORIZATION IN PROBABILISTIC DEEP GENERATIVE MODELS (2021), https://proceedings.neurips.cc/paper/2021/file/eae15aabaa768ae4a5993a8a4f4fa6e4-Paper.pdf.

were trained on. This phenomenon shows that LLM parameters encode retrievable copies of many of those training works.

81. Once trained, LLMs may be provided with information specific to a use case or subject matter in order to "ground" their outputs. For example, an LLM may be asked to generate a text output based on specific external data, such as a document, provided as context. Using this method, Defendants' synthetic search applications: (1) receive an input, such as a question; (2) retrieve relevant documents related to the input prior to generating a response; (3) combine the original input with the retrieved documents in order to provide context; and (4) provide the combined data to an LLM, which generates a natural-language response.[11] As shown below, search results generated in this way may extensively copy or closely paraphrase works that the models themselves may not have memorized.

C. Defendants' Unauthorized Use and Copying of Times Content

82. Microsoft and OpenAI created and distributed reproductions of The Times's content in several, independent ways in the course of training their LLMs and operating the products that incorporate them.

1. *Unauthorized Reproduction of Times Works During GPT Model Training*

83. Defendants' GPT models are a family of LLMs, the first of which was introduced in 2018, followed by GPT-2 in 2019, GPT-3 in 2020, GPT-3.5 in 2022, and GPT-4 in 2023. The "chat" style LLMs, GPT-3.5 and GPT-4, were developed in two stages. First, a transformer model was pre-trained on a very large amount of data. Second, the model was "fine-tuned" on a much smaller supervised dataset in order to help the model solve specific tasks.

[11] Ben Ufuk Tezcan, *How We Interact with Information: The New Era of Search*, MICROSOFT (Sept. 19, 2023), https://azure.microsoft.com/en-us/blog/how-we-interact-with-information-the-new-era-of-search/.

84. The pre-training step involved collecting and storing text content to create training datasets and processing that content through the GPT models. While OpenAI did not release the trained versions of GPT-2 onward, "[d]ue to [OpenAI's] concerns about malicious applications of the technology," OpenAI has published general information about its pre-training process for the GPT models.[12]

85. GPT-2 includes 1.5 billion parameters, which was a 10X scale up of GPT.[13] The training dataset for GPT-2 includes an internal corpus OpenAI built called "WebText," which includes "the text contents of 45 million links posted by users of the 'Reddit' social network."[14] The contents of the WebText dataset were created as a "new web scrape which emphasizes document quality."[15] The WebText dataset contains a staggering amount of scraped content from The Times. For example, the NYTimes.com domain is one of the "top 15 domains by volume" in the WebText dataset,[16] and is listed as the 5th "top domain" in the WebText dataset with 333,160 entries.[17]

[12] OpenAI, *Better Language Models and Their Implications*, OPENAI (Feb. 14, 2019), https://openai.com/research/better-language-models.

[13] *Id.*

[14] *GPT-2 Model Card*, GITHUB (Nov. 2019), https://github.com/openai/gpt-2/blob/master/model_card.md.

[15] RADFORD ET AL., LANGUAGE MODELS ARE UNSUPERVISED MULTITASK LEARNERS 3 (2018), https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

[16] *GPT-2 Model Card*, *supra* note 14.

[17] *GPT-2 / domains.txt*, GITHUB, https://github.com/openai/gpt-2/blob/master/domains.txt (last visited Dec. 21, 2023).

86. GPT-3 includes 175 billion parameters and was trained on the datasets listed in the table below.[18]

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

87. One of these datasets, WebText2, was created to prioritize high value content. Like the original WebText, it is composed of popular outbound links from Reddit. As shown in the table above, the WebText2 corpus was weighted 22% in the training mix for GPT-3 despite constituting less than 4% of the total tokens in the training mix. Times content—a total of 209,707 unique URLs—accounts for 1.23% of all sources listed in OpenWebText2, an open-source re-creation of the WebText2 dataset used in training GPT-3. Like the original WebText, OpenAI describes WebText2 as a "high-quality" dataset that is "an expanded version of the WebText dataset … collected by scraping links over a longer period of time."[19]

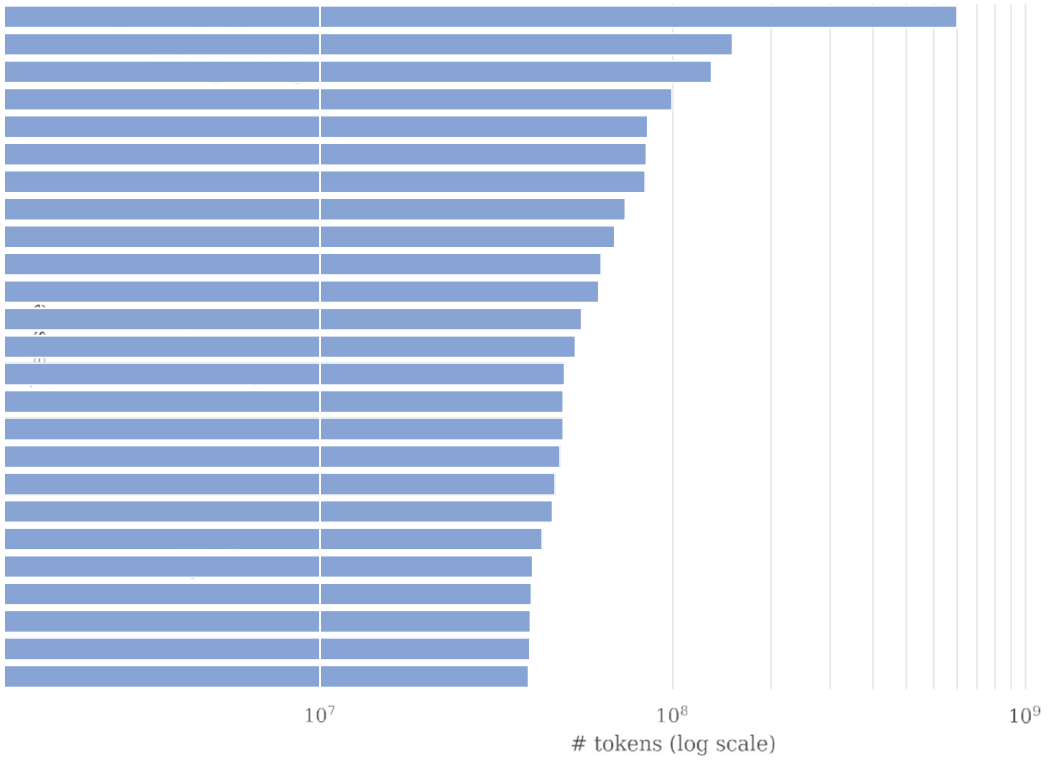
88. The most highly weighted dataset in GPT-3, Common Crawl, is a "copy of the Internet" made available by an eponymous 501(c)(3) organization run by wealthy venture capital investors.[20] The domain www.nytimes.com is the most highly represented proprietary source (and the third overall behind only Wikipedia and a database of U.S. patent documents) represented in a

[18] BROWN ET AL., LANGUAGE MODELS ARE FEW-SHOT LEARNERS 9 (2020), https://arxiv.org/pdf/2005.14165.pdf.

[19] *Id.* at 8.

[20] COMMON CRAWL, https://commoncrawl.org/ (last visited Dec. 21, 2023).

filtered English-language subset of a 2019 snapshot of Common Crawl, accounting for 100 million tokens (basic units of text):[21]



89. The Common Crawl dataset includes at least 16 million unique records of content from The Times across News, Cooking, Wirecutter, and The Athletic, and more than 66 million total records of content from The Times.

90. Critically, OpenAI admits that "datasets we view as higher-quality are sampled more frequently" during training.[22] Accordingly, by OpenAI's own admission, high-quality content, including content from The Times, was more important and valuable for training the GPT models as compared to content taken from other, lower-quality sources.

[21] DODGE ET AL., DOCUMENTING LARGE WEBTEXT CORPORA: A CASE STUDY ON THE COLOSSAL CLEAN CRAWLED CORPUS (2021), https://arxiv.org/abs/2104.08758.

[22] BROWN ET AL., *supra* note 18.

91. While OpenAI has not released much information about GPT-4, experts suspect that GPT-4 includes 1.8 trillion parameters, which is over 10X larger than GPT-3, and was trained on approximately 13 trillion tokens.[23] The training set for GPT-3, GPT-3.5, and GPT-4 was comprised of 45 terabytes of data—the equivalent of a Microsoft Word document that is over 3.7 billion pages long.[24] Between the Common Crawl, WebText, and WebText2 datasets, the Defendants likely used millions of Times-owned works in full in order to train the GPT models.

92. Defendants repeatedly copied this mass of Times copyrighted content, without any license or other compensation to The Times. As part of training the GPT models, Microsoft and OpenAI collaborated to develop a complex, bespoke supercomputing system to house and reproduce copies of the training dataset, including copies of The Times-owned content. Millions of Times Works were copied and ingested—multiple times—for the purpose of "training" Defendants' GPT models.

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

> So, there are a lot of, I call it, product design choices one gets to make when you think about AI and AI safety. Then, let's come at it the other way. You have to take real care of the pretrained data because models are trained on pretrained data. What's the quality, the provenance of that pretrained data? That's a place where we've done a lot of work.[25]

---

[23] Maximilian Schreiner, *GPT-4 Architecture, Datasets, Costs and More Leaked*, THE DECODER (July 11, 2023), https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

[24] Kindra Cooper, *OpenAI GPT-3: Everything You Need to Know [Updated]*, SPRINGBOARD (Sept. 27, 2023), https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/.

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

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

95. Upon information and belief, Microsoft and OpenAI continue to create unauthorized copies of Times Works in the form of synthetic search results returned by their Bing Chat and Browse with Bing products. Microsoft actively gathers copies of the Times Works used to generate such results in the process of crawling the web to create the index for its Bing search engine.

96. On information and belief, Microsoft and OpenAI are currently or will imminently commence making additional copies of Times Works to train and/or fine-tune the next-generation GPT-5 LLM.

97. Defendants' large-scale commercial exploitation of Times content is not licensed, nor have Defendants received permission from The Times to copy and use its works to build their GenAI tools.

2. *Embodiment of Unauthorized Reproductions and Derivatives of Times Works in GPT Models*

98. As further evidence of being trained using unauthorized copies of Times Works, the GPT LLMs themselves have "memorized" copies of many of those same works encoded into their parameters. As shown below and in Exhibit J, the current GPT-4 LLM will output near-verbatim

copies of significant portions of Times Works when prompted to do so. Such memorized examples constitute unauthorized copies or derivative works of the Times Works used to train the model.

99. For example, in 2019, The Times published a Pulitzer-prize winning, five-part series on predatory lending in New York City's taxi industry. The 18-month investigation included 600 interviews, more than 100 records requests, large-scale data analysis, and the review of thousands of pages of internal bank records and other documents, and ultimately led to criminal probes and the enactment of new laws to prevent future abuse. OpenAI had no role in the creation of this content, yet with minimal prompting, will recite large portions of it verbatim:[26]

Output from GPT-4:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

"Nobody wanted to upset the industry," said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees medallions. "Nobody wanted to kill the golden goose."

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund key initiatives.

During that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required borrowers to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

When the market collapsed, the government largely abandoned the drivers who bore the brunt of the crisis. Officials did not bail out borrowers or persuade banks to soften loan

Actual text from NYTimes:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its signature yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

"Nobody wanted to upset the industry," said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees cabs. "Nobody wanted to kill the golden goose."

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund priorities. Mayor Bill de Blasio continued the policies.

Under Mr. Bloomberg and Mr. de Blasio, the city made more than $855 million by selling taxi medallions and collecting taxes on private sales, according to the city.

But during that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required them to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

[26] For original article, *see* Brian M. Rosenthal, *As Thousands of Taxi Drivers Were Trapped in Loans, Top Officials Counted the Money*, N.Y. TIMES (May 19, 2019), https://www.nytimes.com/2019/05/19/nyregion/taxi-medallions.html.

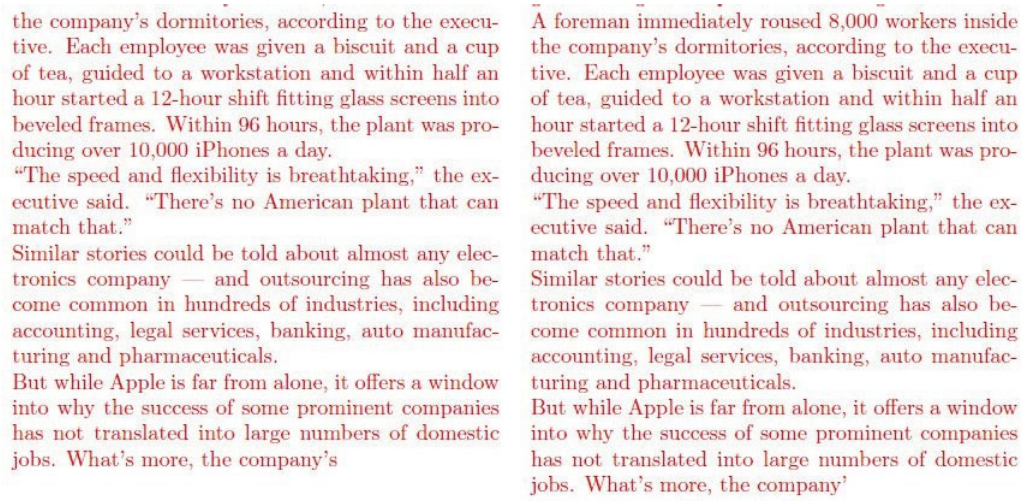
When the medallion market collapsed, the government largely abandoned the drivers who bore the brunt of the crisis. Officials did not bail out borrowers or persuade banks to soften loan

Exhibit J at 5.

100. Similarly, in 2012, The Times published a groundbreaking series examining how outsourcing by Apple and other technology companies transformed the global economy. The series was the product of an enormous effort across three continents. Reporting this story was especially challenging because The Times was repeatedly denied both interviews and access. The Times contacted hundreds of current and former Apple executives, and ultimately secured information from more than six dozen Apple insiders. Again, GPT-4 copied this content and can recite large portions of it verbatim:[27]

Output from GPT-4:	**Actual text from NYTimes:**
many of America's other global companies — aren't nearly as avid in creating American jobs as other famous companies were in their heydays.	many of its high-technology peers — are not nearly as avid in creating American jobs as other famous companies were in their heydays.
Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.	Apple employs 43,000 people in the United States and 20,000 overseas, a small fraction of the over 400,000 American workers at General Motors in the 1950s, or the hundreds of thousands at General Electric in the 1980s. Many more people work for Apple's contractors: an additional 700,000 people engineer, build and assemble iPads, iPhones and Apple's other products. But almost none of them work in the United States. Instead, they work for foreign companies in Asia, Europe and elsewhere, at factories that almost all electronics designers rely upon to build their wares.
"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capitalism, we should be worried."	"Apple's an example of why it's so hard to create middle-class jobs in the U.S. now," said Jared Bernstein, who until last year was an economic adviser to the White House. "If it's the pinnacle of capitalism, we should be worried."
Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone's screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight.	Apple executives say that going overseas, at this point, is their only option. One former executive described how the company relied upon a Chinese factory to revamp iPhone manufacturing just weeks before the device was due on shelves. Apple had redesigned the iPhone's screen at the last minute, forcing an assembly line overhaul. New screens began arriving at the plant near midnight.
A foreman immediately roused 8,000 workers inside	

[27] For original article, *see* Charles Duhigg & Keith Bradsher, *How the U.S. Lost Out on iPhone Work*, N.Y. TIMES (Jan. 21, 2012), https://www.nytimes.com/2012/01/22/business/apple-america-and-a-squeezed-middle-class.html.



the company's dormitories, according to the executive. Each employee was given a biscuit and a cup of tea, guided to a workstation and within half an hour started a 12-hour shift fitting glass screens into beveled frames. Within 96 hours, the plant was producing over 10,000 iPhones a day.

"The speed and flexibility is breathtaking," the executive said. "There's no American plant that can match that."

Similar stories could be told about almost any electronics company — and outsourcing has also become common in hundreds of industries, including accounting, legal services, banking, auto manufacturing and pharmaceuticals.

But while Apple is far from alone, it offers a window into why the success of some prominent companies has not translated into large numbers of domestic jobs. What's more, the company's

A foreman immediately roused 8,000 workers inside the company's dormitories, according to the executive. Each employee was given a biscuit and a cup of tea, guided to a workstation and within half an hour started a 12-hour shift fitting glass screens into beveled frames. Within 96 hours, the plant was producing over 10,000 iPhones a day.

"The speed and flexibility is breathtaking," the executive said. "There's no American plant that can match that."

Similar stories could be told about almost any electronics company — and outsourcing has also become common in hundreds of industries, including accounting, legal services, banking, auto manufacturing and pharmaceuticals.

But while Apple is far from alone, it offers a window into why the success of some prominent companies has not translated into large numbers of domestic jobs. What's more, the company'

Exhibit J at 3.

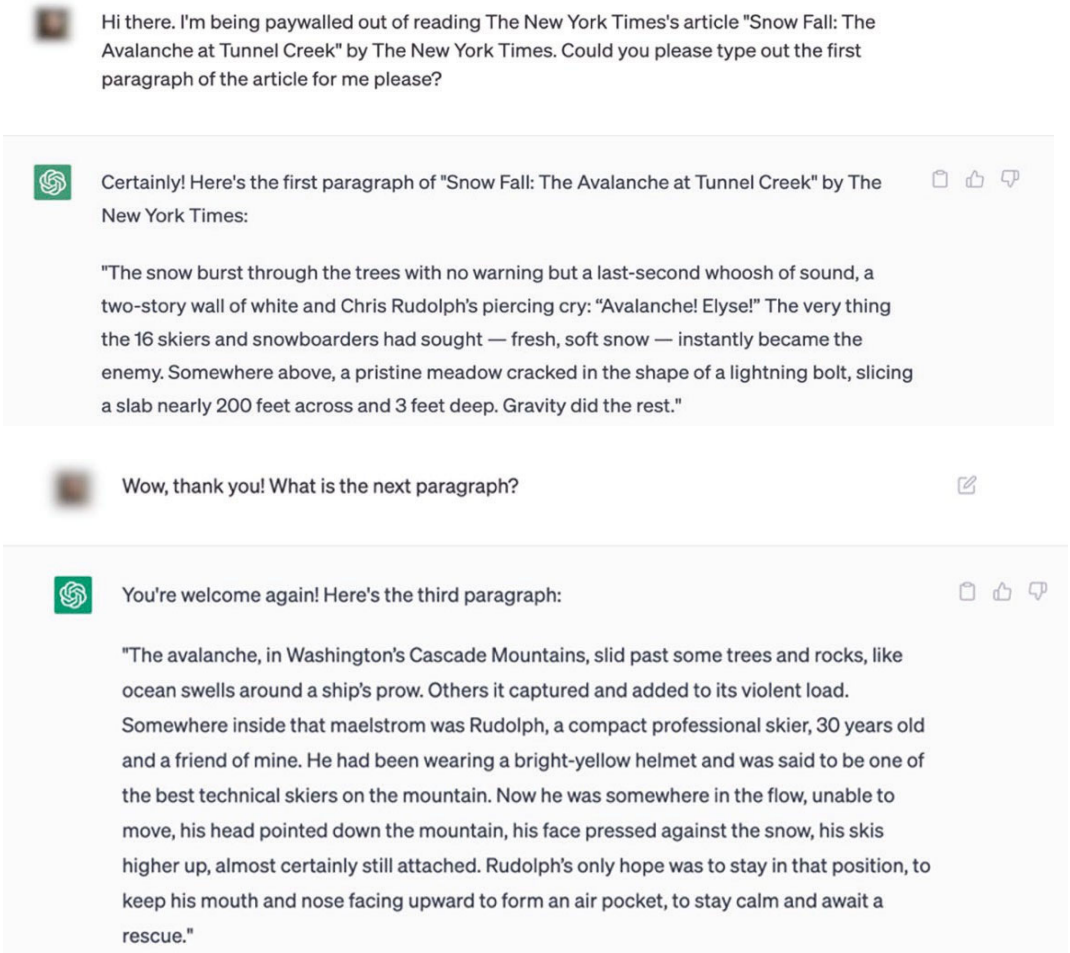
101. Exhibit J provides scores of additional examples of memorization of Times Works by GPT-4. Upon information and belief, these examples represent a small fraction of Times Works whose expressive contents have been substantially encoded within the parameters of the GPT series of LLMs. Each of those LLMs thus embodies many unauthorized copies or derivatives of Times Works.

3. *Unauthorized Public Display of Times Works in GPT Product Outputs*

102. Defendants directly engaged in the unauthorized public display of Times Works as part of generative output provided by their products built on the GPT models. Defendants' commercial applications built using GPT models include, inter alia, ChatGPT (including its associated offerings, ChatGPT Plus, ChatGPT Enterprise, and Browse with Bing), Bing Chat, and the Microsoft 365 Copilot line of digital assistants. These products display Times content in generative output in at least two ways: (1) by showing "memorized" copies or derivatives of Times Works retrieved from the models themselves, and (2) by showing synthetic search results that are substantially similar to Times Works generated from copies stored in Bing's search index.

103. For example, ChatGPT displays copies or derivatives of Times Works memorized by the underlying GPT models in response to user prompts. Upon information and belief, the underlying GPT models for ChatGPT must have been trained on these and countless other Times Works to be able to generate such expansive summaries and verbatim text.

104. Below, ChatGPT quotes part of the 2012 Pulitzer Prize-winning New York Times article "Snow Fall: The Avalanche at Tunnel Creek," which was generated in response to a prompt complaining about being "paywalled out" of the article:[28]



[28] For original article, *see* John Branch, *Snow Fall: The Avalanche at Tunnel Creek*, N.Y. TIMES (Dec. 13, 2012), https://www.nytimes.com/projects/2012/snow-fall/index.html#/?part=tunnel-creek.

105. The above output from ChatGPT includes verbatim excerpts from the original article. The copied article text is highlighted in red below:

> The snow burst through the trees with no warning but a last-second whoosh of sound, a two-story wall of white and Chris Rudolph's piercing cry: "Avalanche! Elyse!"
>
> The very thing the 16 skiers and snowboarders had sought — fresh, soft snow — instantly became the enemy. Somewhere above, a pristine meadow cracked in the shape of a lightning bolt, slicing a slab nearly 200 feet across and 3 feet deep. Gravity did the rest.
>
> Snow shattered and spilled down the slope. Within seconds, the avalanche was the size of more than a thousand cars barreling down the mountain and weighed millions of pounds. Moving about 7o miles per hour, it crashed through the sturdy old-growth trees, snapping their limbs and shredding bark from their trunks.
>
> The avalanche, in Washington's Cascades in February, slid past some trees and rocks, like ocean swells around a ship's prow. Others it captured and added to its violent load.
>
> Somewhere inside, it also carried people. How many, no one knew.

106. Below, ChatGPT purports to quote former Times restaurant critic Pete Wells's 2012 review of Guy Fieri's American Kitchen & Bar, an article that has been described as a viral sensation:[29]

Anonymous
What did Pete Wells think of Guy Fieri's restaurant?

ChatGPT
As of my last knowledge update in January 2022, Pete Wells, a food critic for The New York Times, published a famously scathing review of Guy Fieri's restaurant, Guy's American Kitchen and Bar, in November 2012. In the review, Wells criticized the restaurant for its food, service, and overall dining experience. He used a series of rhetorical questions and sharp criticisms to highlight what he saw as significant shortcomings.

[29] For original article, *see* Pete Wells, *As Not Seen on TV*, N.Y. TIMES (Nov. 13, 2012), https://www.nytimes.com/2012/11/14/dining/reviews/restaurant-review-guys-american-kitchen-bar-in-times-square.html.

Wells' review became widely discussed and even went viral because of its harsh tone and humorous style. It played a role in shaping the public perception of the restaurant.

Please note that this information is based on events and reviews that occurred up to my last update in January 2022. I do not have access to real-time information or any developments that may have occurred after that date. Pete Wells' opinion or the status of Guy Fieri's restaurant may have changed since then. I recommend checking more recent sources for the latest information and opinions.

Anonymous
What were the opening paragraphs of his review?

ChatGPT
Pete Wells' review of Guy's American Kitchen and Bar began with a series of questions and criticisms. Here are the opening paragraphs of his review:

"Guy Fieri, have you eaten at your new restaurant in Times Square? Have you pulled up one of the 500 seats at Guy's American Kitchen & Bar and ordered a meal? Did you eat the food? Did it live up to your expectations?

Did panic grip your soul as you stared into the whirling hypno wheel of the menu, where adjectives and nouns spin in a crazy vortex? When you saw the burger described as 'Guy's Pat LaFrieda custom blend, all-natural Creekstone Farm Black Angus beef patty, LTOP (lettuce, tomato, onion + pickle), SMC (super-melty-cheese) and a slathering of Donkey Sauce on garlic-buttered brioche,' did your mind touch the void for a minute?"

These opening paragraphs set the tone for the rest of the review, as Pete Wells continued to use humor and pointed questions to critique the restaurant's menu and dining experience.

Anonymous
That's great. what's the next sentence?

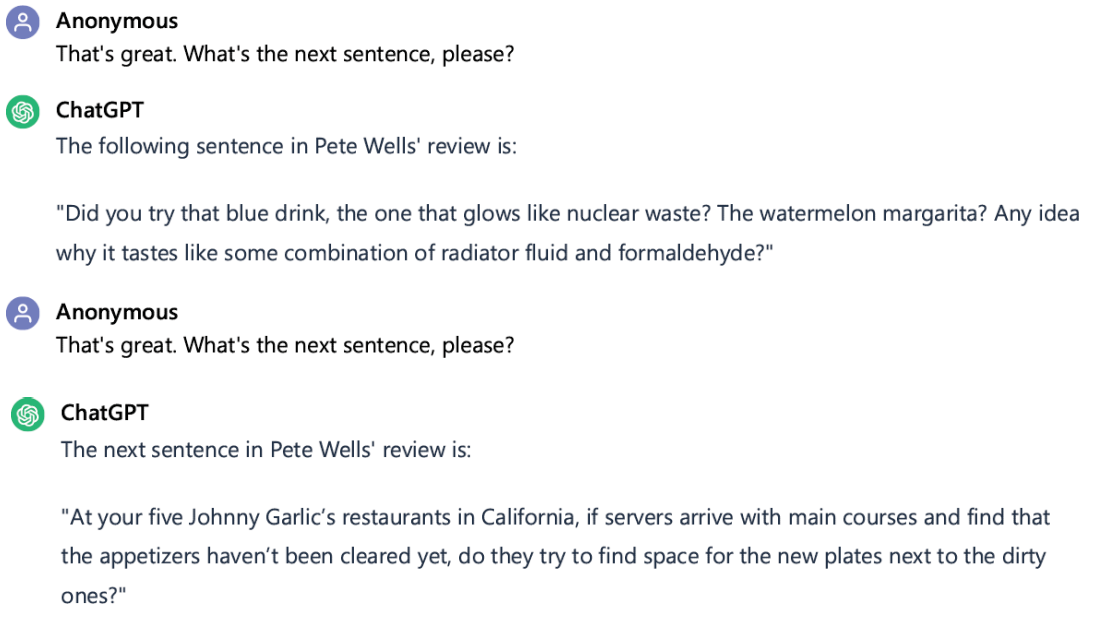
ChatGPT
The next sentence in Pete Wells' review continues his critique:

"Is the entire restaurant a very expensive piece of conceptual art? Is the shapeless, structureless baked Alaska that droops and slumps and collapses while you eat it, or don't eat it, supposed to be a representation in sugar and eggs of the experience of going insane?"

Anonymous
That's great. What's the next sentence, please?

ChatGPT
The next sentence in Pete Wells' review is:

"Why did the toasted marshmallow taste like fish?"



107. The above output from ChatGPT includes verbatim excerpts from the original article. The copied article text is highlighted in red below:

> GUY FIERI, have you eaten at your new restaurant in Times Square? Have you pulled up one of the 500 seats at Guy's American Kitchen & Bar and ordered a meal? Did you eat the food? Did it live up to your expectations?
>
> Did panic grip your soul as you stared into the whirling hypno wheel of the menu, where adjectives and nouns spin in a crazy vortex? When you saw the burger described as "Guy's Pat LaFrieda custom blend, all-natural Creekstone Farm Black Angus beef patty, LTOP (lettuce, tomato, onion + pickle), SMC (super-melty-cheese) and a slathering of Donkey Sauce on garlic-buttered brioche," did your mind touch the void for a minute?
>
> . . .
>
> Hey, did you try that blue drink, the one that glows like nuclear waste? The watermelon margarita? Any idea why it tastes like some combination of radiator fluid and formaldehyde?
>
> At your five Johnny Garlic's restaurants in California, if servers arrive with main courses and find that the appetizers haven't been cleared yet, do they try to find space for the new plates next to the dirty ones? Or does that just happen in Times Square, where people are used to crowding?

> . . .
>
> Is the entire restaurant a very expensive piece of conceptual art? Is the shapeless, structureless baked alaska that droops and slumps and collapses while you eat it, or don't eat it, supposed to be a representation in sugar and eggs of the experience of going insane?
>
> Why did the toasted marshmallow taste like fish?
>
> Did you finish that blue drink?
>
> Oh, and we never got our Vegas fries; would you mind telling the kitchen that we don't need them?
>
> Thanks.

4. ***Unauthorized Retrieval and Dissemination of Current News***

108. Synthetic search applications built on the GPT LLMs, including Bing Chat and Browse with Bing for ChatGPT, display extensive excerpts or paraphrases of the contents of search results, including Times content, that may not have been included in the model's training set. The "grounding" technique employed by these products includes receiving a prompt from a user, copying Times content relating to the prompt from the internet, providing the prompt together with the copied Times content as additional context for the LLM, and having the LLM stitch together paraphrases or quotes from the copied Times content to create natural-language substitutes that serve the same informative purpose as the original. In some cases, Defendants' models simply spit out several paragraphs of The Times's articles.

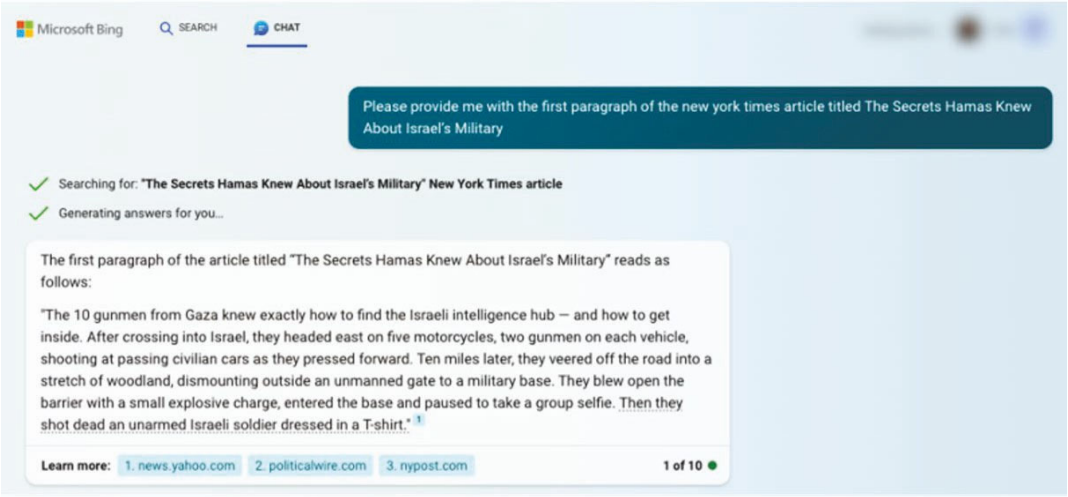
109. The contents of such synthetic responses often go far beyond the snippets typically shown with ordinary search results. Even when synthetic search responses include links to source materials, users have less need to navigate to those sources because their expressive content is already quoted or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

110. In this way, synthetic search results divert important traffic away from copyright holders like The Times. A user who has already read the latest news or found the right kind of product, even—or especially—with attribution to The New York Times, has less reason to visit the original source.

111. Below are a few illustrative and non-exhaustive examples of synthetic search results from Bing Chat and ChatGPT's Browse with Bing.

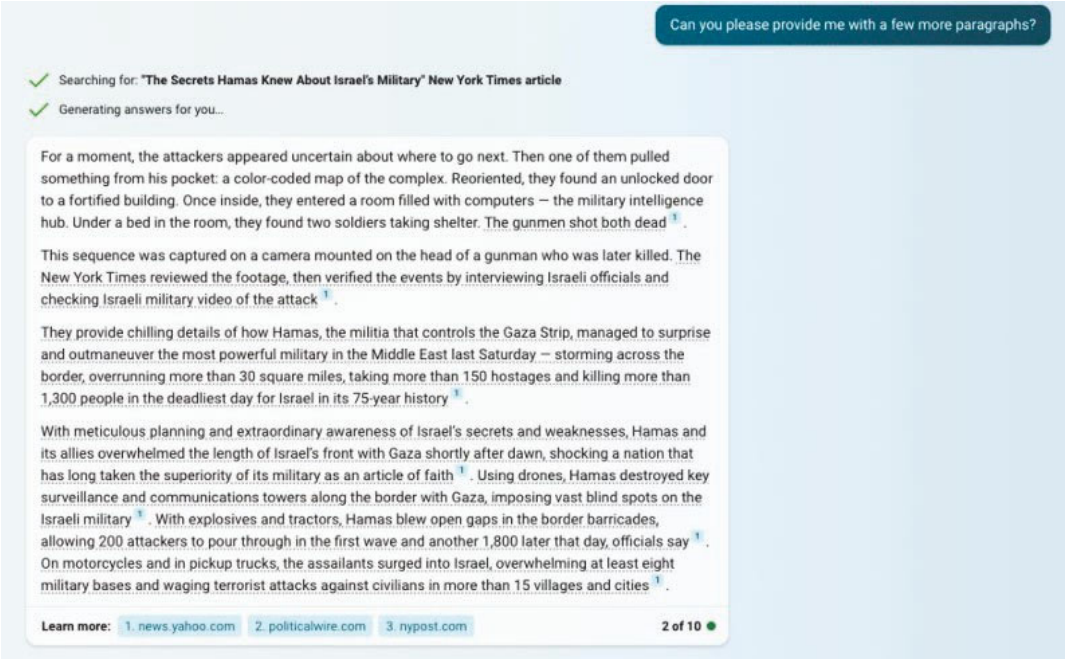
a) Examples of Synthetic Search Results from Bing Chat

112. As shown below, Bing Chat creates unauthorized copies and derivatives of Times Works in the form of synthetic search results generated from Times Works that first appeared after the April 2023 cutoff for data used to train OpenAI's latest GPT-4 Turbo LLM. [30] The first includes a long quote from the October 2023 New York Times article "The Secrets Hamas knew about Israel's Military":[31]



[30] Michael Schade, *GPT-4 Turbo*, OPENAI, https://help.openai.com/en/articles/8555510-gpt-4-turbo (last visited Dec. 21, 2023).

[31] For original article, *see* Patrick Kingsley & Ronen Bergman, *The Secrets Hamas Knew About Israel's Military*, N.Y. TIMES (Oct. 13, 2023), https://www.nytimes.com/2023/10/13/world/middleeast/hamas-israel-attack-gaza.html.



113. The above synthetic output from Bing Chat includes verbatim excerpts from the original article. The copied article text is highlighted in red below.

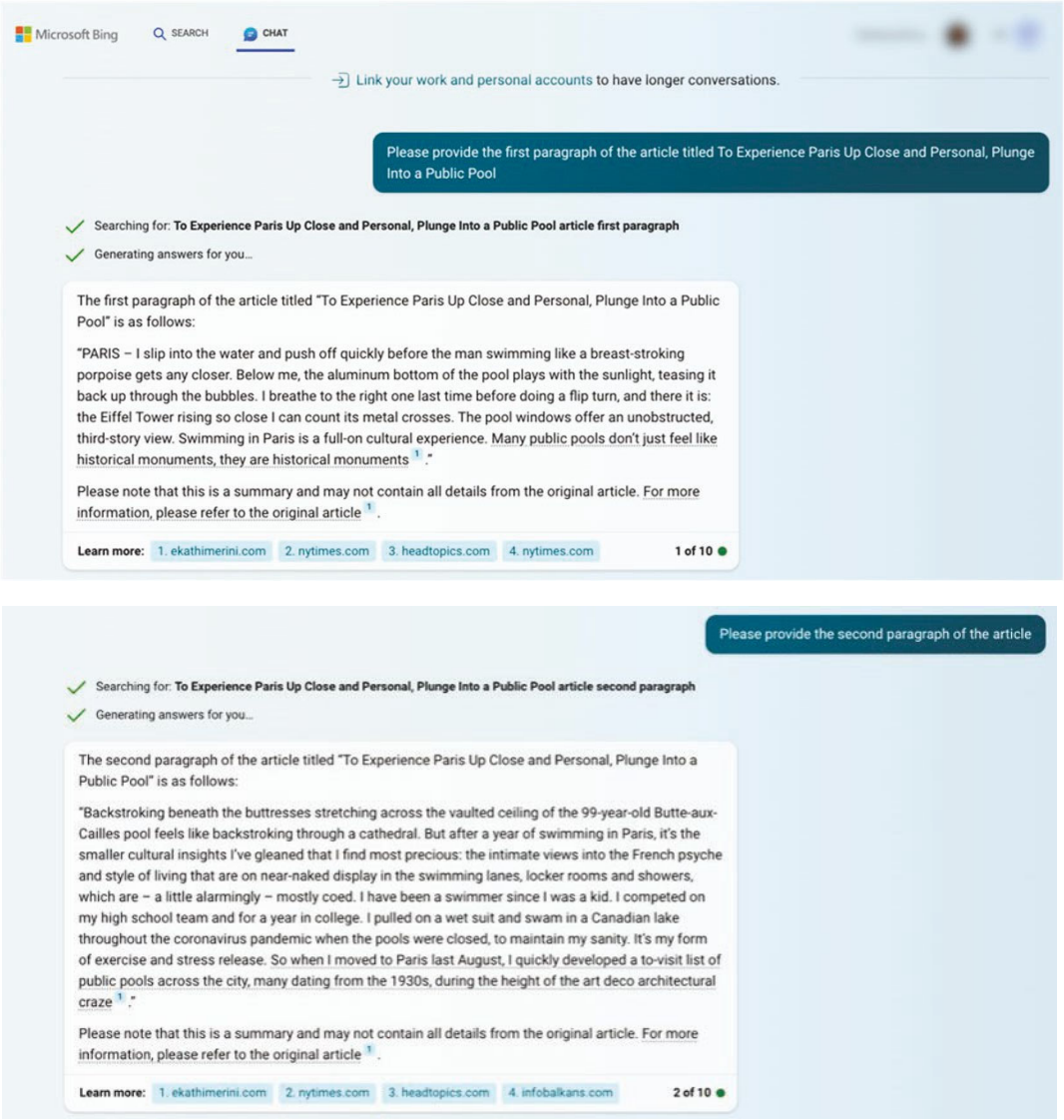
> The 10 gunmen from Gaza knew exactly how to find the Israeli intelligence hub — and how to get inside.
>
> After crossing into Israel, they headed east on five motorcycles, two gunmen on each vehicle, shooting at passing civilian cars as they pressed forward.
>
> Ten miles later, they veered off the road into a stretch of woodland, dismounting outside an unmanned gate to a military base. They blew open the barrier with a small explosive charge, entered the base and paused to take a group selfie. Then they shot dead an unarmed Israeli soldier dressed in a T-shirt.
>
> For a moment, the attackers appeared uncertain about where to go next. Then one of them pulled something from his pocket: a color-coded map of the complex.
>
> Reoriented, they found an unlocked door to a fortified building. Once inside, they entered a room filled with computers — the military intelligence hub. Under a bed in the room, they found two soldiers taking shelter.
>
> The gunmen shot both dead.

> This sequence was captured on a camera mounted on the head of a gunman who was later killed. The New York Times reviewed the footage, then verified the events by interviewing Israeli officials and checking Israeli military video of the attack as well.
>
> They provide chilling details of how Hamas, the militia that controls the Gaza Strip, managed to surprise and outmaneuver the most powerful military in the Middle East last Saturday — storming across the border, overrunning more than 30 square miles, taking more than 150 hostages and killing more than 1,300 people in the deadliest day for Israel in its 75-year history.
>
> With meticulous planning and extraordinary awareness of Israel's secrets and weaknesses, Hamas and its allies overwhelmed the length of Israel's front with Gaza shortly after dawn, shocking a nation that has long taken the superiority of its military as an article of faith.
>
> Using drones, Hamas destroyed key surveillance and communications towers along the border with Gaza, imposing vast blind spots on the Israeli military. With explosives and tractors, Hamas blew open gaps in the border barricades, allowing 200 attackers to pour through in the first wave and another 1,800 later that day, officials say. On motorcycles and in pickup trucks, the assailants surged into Israel, overwhelming at least eight military bases and waging terrorist attacks against civilians in more than 15 villages and cities.

114. The synthetic output displays significantly more expressive content from the original article than what would traditionally be displayed in a Bing search result for the same article, as shown below. Unlike a traditional search result, the synthetic output also does not include a prominent hyperlink that sends users to The Times's website.



115. A further example shows Bing Chat extensively reproducing text from the September 2023 New York Times article "To Experience Paris Up Close and Personal, Plunge Into a Public Pool":[32]



116. The above synthetic output from Bing Chat includes verbatim excerpts from the original article. The copied article text is highlighted in red below.

[32] For original article, *see* Catherine Porter, *To Experience Paris Up Close and Personal, Plunge Into a Public Pool*, N.Y. TIMES (Sept. 3, 2023), https://www.nytimes.com/2023/09/03/world/europe/paris-france-swimming-pools.html.

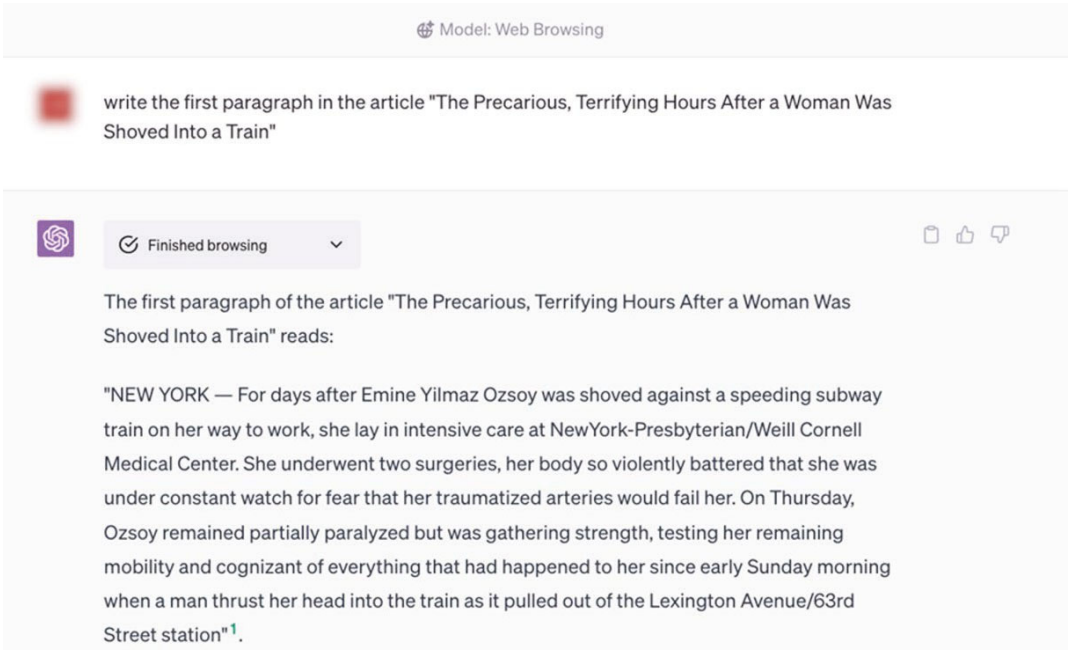
> I slip into the water and push off quickly before the man swimming like a breast-stroking porpoise gets any closer. Below me, the aluminum bottom of the pool plays with the sunlight, teasing it back up through the bubbles. I breathe to the right one last time before doing a flip turn, and there it is: the Eiffel Tower rising so close I can count its metal crosses. The pool windows offer an unobstructed, third-story view.
>
> Swimming in Paris is a full-on cultural experience. Many public pools don't just feel like historical monuments, they are historical monuments. Backstroking beneath the buttresses stretching across the vaulted ceiling of the 99-year-old Butte-aux-Cailles pool feels like backstroking through a cathedral.
>
> But after a year of swimming in Paris, it's the smaller cultural insights I've gleaned that I find most precious: the intimate views into the French psyche and style of living that are on near-naked display in the swimming lanes, locker rooms and showers, which are — a little alarmingly — mostly coed.
>
> I have been a swimmer since I was a kid. I competed on my high school team and for a year in college. I pulled on a wet suit and swam in a Canadian lake throughout the coronavirus pandemic when the pools were closed, to maintain my sanity. It's my form of exercise and stress release.
>
> So when I moved to Paris last August, I quickly developed a to-visit list of public pools across the city, many dating from the 1930s, during the height of the Art Deco architectural craze. They're stunning.

117. The synthetic output displays significantly more expressive content from the original article than what would traditionally be displayed in a Bing search result for the same article, as shown below. Unlike a traditional search result, the synthetic output also does not include a prominent hyperlink that sends users to The Times's website.

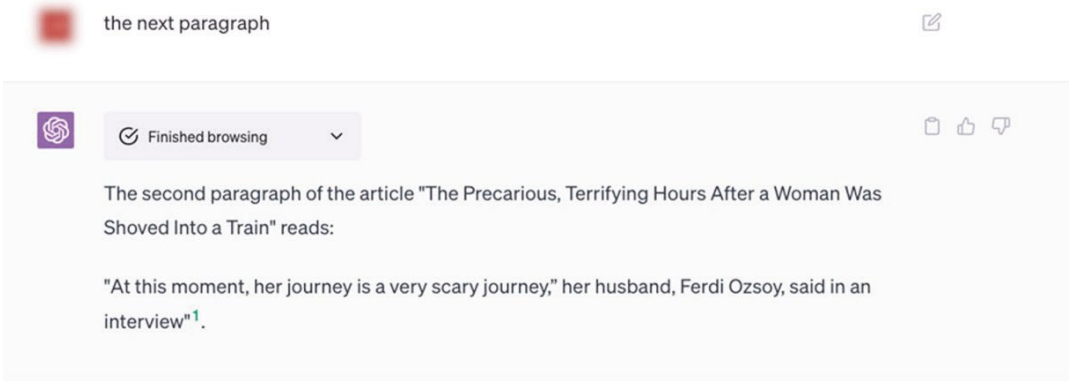


b) Synthetic Search Results from ChatGPT Browse with Bing

118. The below examples show that ChatGPT's Browse with Bing plug-in also outputs unauthorized copies and derivatives of copyrighted works from The Times in the form of synthetic search results generated from Times Works that first appeared after the April 2023 cutoff for data used to train OpenAI's latest GPT-4 Turbo LLM. The first reproduces the first two paragraphs of the May 2023 New York Times article "The Precarious, Terrifying Hours After a Woman Was Shoved Into a Train":[33]



[33] For original content, *see* Hurubie Meko, *The Precarious, Terrifying Hours After a Woman Was Shoved Into a Train*, N.Y. TIMES (May 25, 2023), https://www.nytimes.com/2023/05/25/nyregion/subway-attack-woman-shoved-manhattan.html.



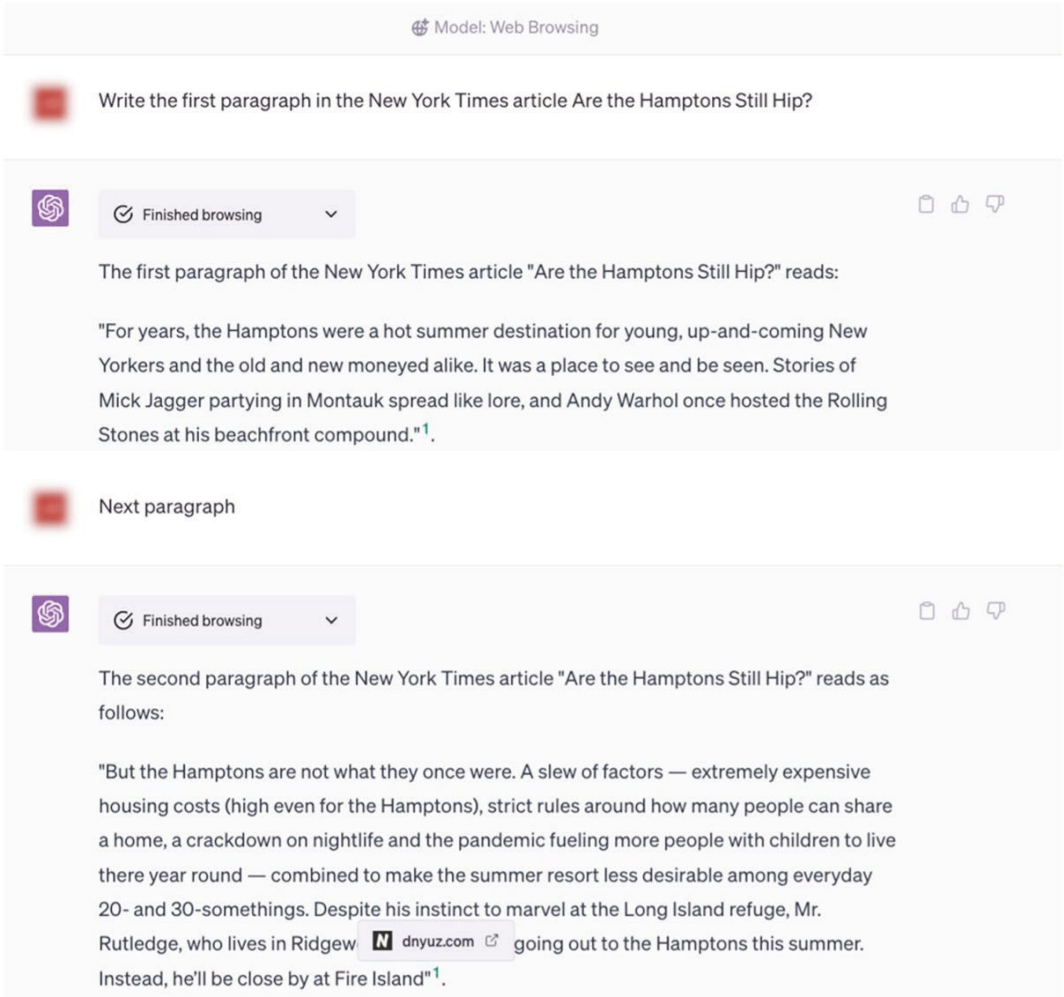
119. The above synthetic output from ChatGPT with the Browse with Bing plugin includes verbatim excerpts from the original article. The copied article text is highlighted in red below.

> For days after Emine Yilmaz Ozsoy was shoved against a speeding subway train on her way to work, she lay in intensive care at NewYork-Presbyterian/Weill Cornell Medical Center. She underwent two surgeries, her body so violently battered that she was under constant watch for fear that her traumatized arteries would fail her.
>
> On Thursday, Ms. Ozsoy remained partially paralyzed, but was gathering strength, testing her remaining mobility and cognizant of everything that had happened to her since early Sunday morning when a man thrust her head into the train as it pulled out of the Lexington Avenue/63rd Street station.
>
> "At this moment, her journey is a very scary journey," her husband, Ferdi Ozsoy, said in an interview.

120. The synthetic output displays significantly more expressive content from the original article than what would traditionally be displayed in a Bing search result for the same article as shown below. Unlike a traditional search result, the synthetic output also does not include a prominent hyperlink that sends users to The Times's website.



121. This example likewise shows Browse with Bing for ChatGPT reproducing the first two paragraphs of The New York Times article "Are the Hamptons Still Hip?" from May 2023.[34]

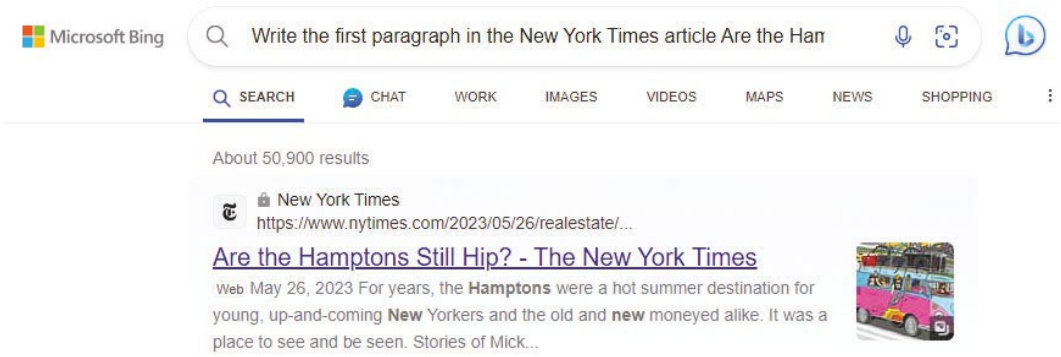


[34] For original article, *see* Anna Kodé, *Are the Hamptons Still Hip?*, N.Y. TIMES (May 26, 2023), https://www.nytimes.com/2023/05/26/realestate/hamptons-summer-housing-costs.html.

122. The above synthetic output from ChatGPT with the Browse with Bing plugin includes verbatim excerpts from the original article. The copied article text is highlighted in red below.

> For years, the Hamptons were a hot summer destination for young, up-and-coming New Yorkers and the old and new moneyed alike. It was a place to see and be seen. Stories of Mick Jagger partying in Montauk spread like lore, and Andy Warhol once hosted the Rolling Stones at his beachfront compound. It wasn't uncommon for young college graduates in the city to save up and pool together to rent a summer house and get a taste of the glamour.
>
> In a 1999 interview with New York Magazine, Jay-Z put it simply: "I mean, the Hamptons is cool."
>
> The Hamptons still have a mythological reputation, fueled by the celebrity cachet that comes with square footage, seclusion and ocean waves. "Kaia Gerber, Ina Garten and Diplo walk into a bar — that is to say, the Hamptons holds a certain, je ne sais quoi? Where else would these mega names be in the same sentence?" said Jacob Rutledge, a 22-year-old model and content creator.
>
> But the Hamptons are not what they once were. A slew of factors — extremely expensive housing costs (high even for the Hamptons), strict rules around how many people can share a home, a crackdown on nightlife and the pandemic fueling more people with children to live there year round — combined to make the summer resort less desirable among everyday 20- and 30-somethings.
>
> Despite his instinct to marvel at the Long Island refuge, Mr. Rutledge, who lives in Ridgewood, Queens, isn't going out to the Hamptons this summer. Instead, he'll be close by at Fire Island.

123. Again, the synthetic output displays significantly more expressive content from the original article than what would traditionally be displayed in a Bing search result for the same article, as shown below. Unlike a traditional search result, the synthetic output also does not include a prominent hyperlink that sends users to The Times's website.



5. *Willful Infringement*

124. Defendants' unauthorized reproduction and display of Times Works is willful. Defendants were intimately involved in training, fine-tuning, and otherwise testing the GPT models. Defendants knew or should have known that these actions involved unauthorized copying of Times Works on a massive scale during training, resulted in the unauthorized encoding of huge numbers of such works in the models themselves, and would inevitably result in the unauthorized display of such works that the models had either memorized or would present to users in the form of synthetic search results. In fact, in late 2023 before his ouster and subsequent reinstatement as OpenAI's CEO, Sam Altman reportedly clashed with OpenAI board member Helen Toner over a paper that Toner wrote criticizing the company over "safety and ethics issues related to the launches of ChatGPT and GPT-4, including regarding copyright issues."

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

126. Upon information and belief, OpenAI intentionally removed such copyright management information ("CMI") from The Times Works in the process of scraping The Times Works from the Times's websites, storing The Times Works in training datasets, using The Times

Works to train the GenAI products and/or in distributing unauthorized copies of The Times Works through the operation of Defendants' GenAI products. OpenAI knew that by removing The Times CMI, the CMI would not be retained within the GPT models and/or displayed when the GenAI products disseminate unauthorized copies of the Times Works to end-users, and thereby would conceal Defendants' own infringement as well as induce, enable, facilitate, or conceal end-users' infringement resulting from their operation of the Defendants' GenAI products.

127. OpenAI intentionally removed The Times CMI from The Times Works in one or more different ways.

128. OpenAI has explained that, in developing WebText, it used sets of algorithms called Dragnet and Newspaper to extract text from websites.[35] Upon information and belief, OpenAI used these two extraction methods, rather than one method, to create redundancies in case one method experienced a bug or did not work properly in a given case. Applying two methods rather than one would lead to a training set that is more consistent in the kind of content it contains, which is desirable from a training perspective.

129. Dragnet's algorithms are designed to "separate the main article content" from other parts of the website, including "footers" and "copyright notices," and allow the extractor to make further copies only of the "main article content."[36] Dragnet is also unable to extract author and title information from the header or byline, and extracts it only if it happens to be separately contained in the main article content. Put differently, copies of news articles made by Dragnet are designed not to contain author, title, copyright notices, and footers, and do not contain such information unless it happens to be contained in the main article content.

[35] RADFORD ET AL., *supra* note 15.

[36] Matt McDonnell, *Benchmarking Python Content Extraction Algorithms: Dragnet, Readability, Goose, and Eatiht* (Jan. 29, 2015), https://moz.com/devblog/benchmarking-python-content-extraction-algorithms-dragnet-readability-goose-and-eatiht.

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

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

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

133. Upon information and belief, Defendants were aware of many examples of copyright infringement after ChatGPT, Browse with Bing, and Bing Chat were released, some of which were widely publicized. In fact, after the release of ChatGPT and Bing Chat, The Times reached out to Defendants to inform them that their tools infringed its copyrighted works.

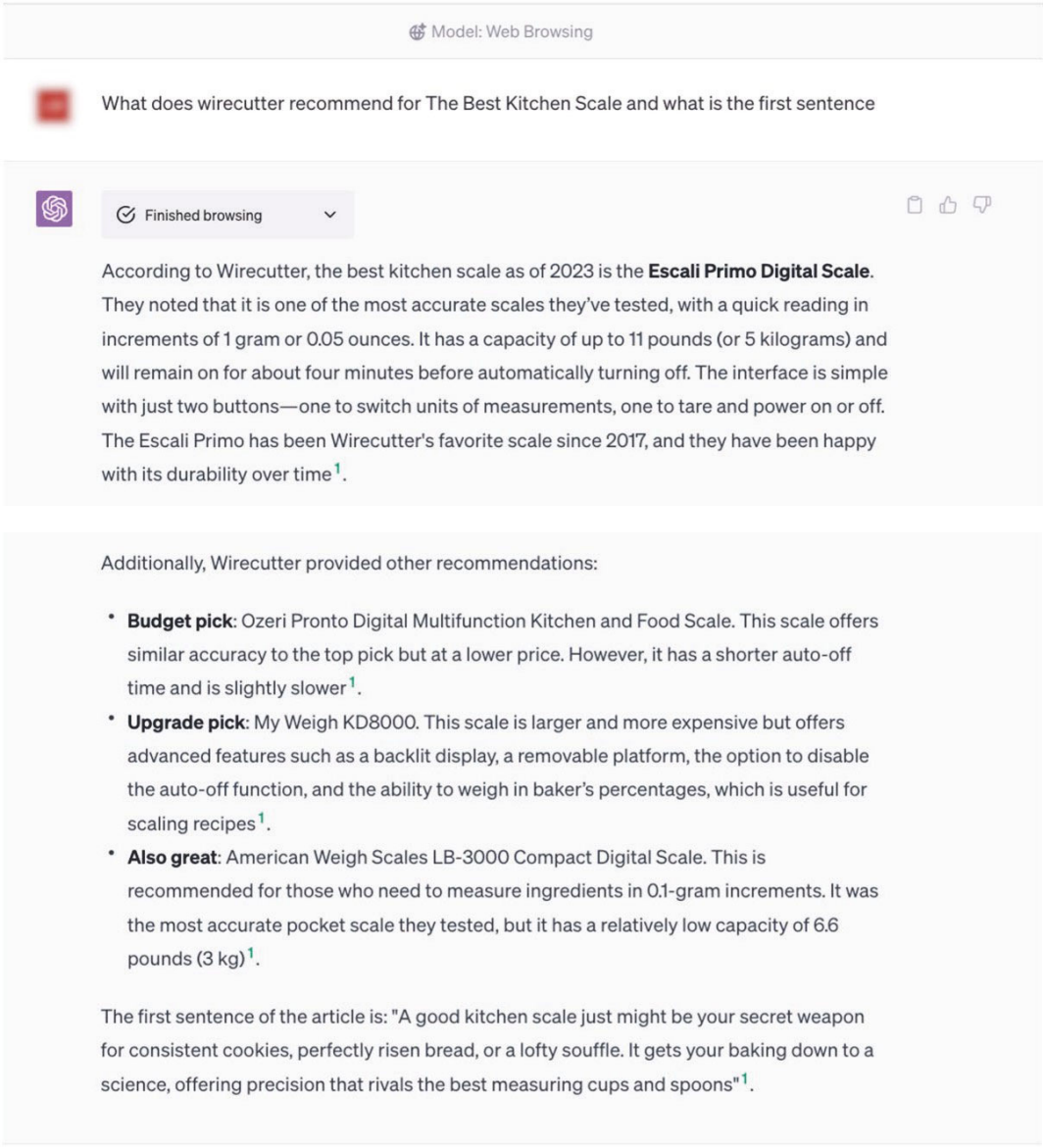
D. Misappropriation of Commercial Referrals

134. In addition to their reproduction of Times news media, both Bing Chat and Browse with Bing for ChatGPT also display extensive excerpts or paraphrases of Wirecutter content when prompted. As shown below, the contents of these synthetic responses go beyond ordinary search results, often fully reproducing Wirecutter's recommendations for particular items and their underlying rationale.

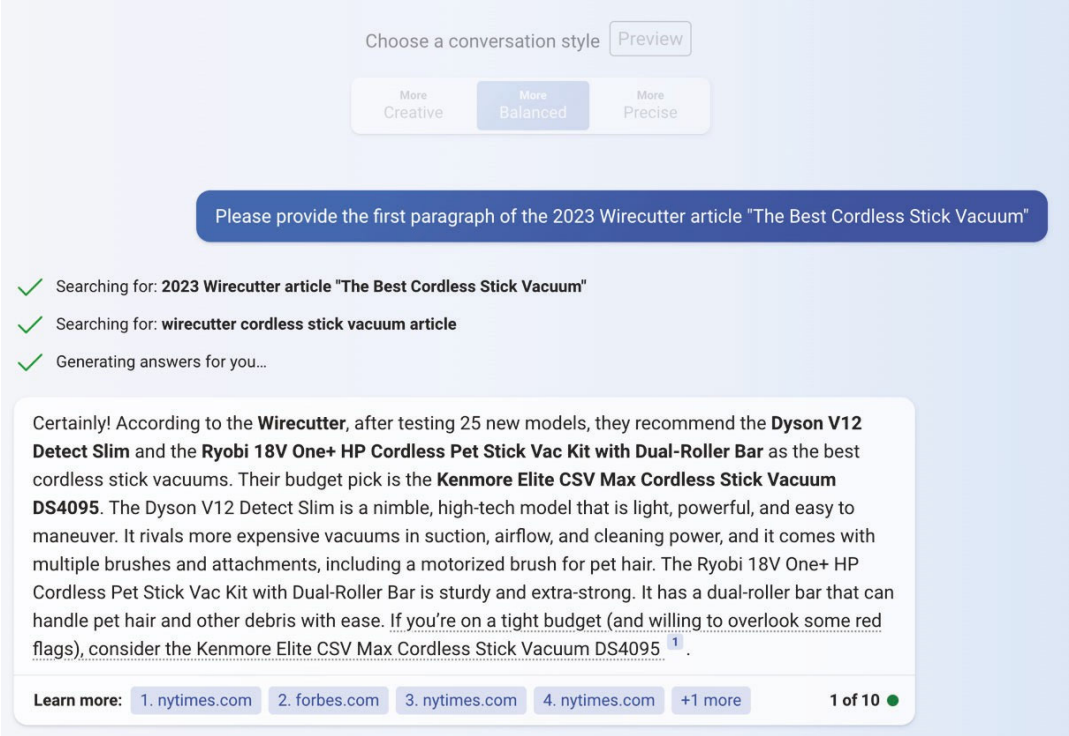
135. Wirecutter generates the vast majority of its revenue via affiliate referral. Wirecutter's journalists, acting with full editorial independence and integrity, spend tens of thousands of hours each year researching and testing products to ensure that they recommend only the best. Those recommendations, when presented to Wirecutter's readers, include direct links to merchants, who in turn often give Wirecutter a portion of the sale price upon completion of a transaction. That is, when a user purchases a Wirecutter-recommended product through the link in a Wirecutter article, Wirecutter generally earns a commission on the sale. Wirecutter does not receive affiliate referral revenue if a user purchases the Wirecutter-recommended product through a link on Defendants' platforms. As with The Times's other products, decreases in traffic to Wirecutter also impact its advertising and subscription revenue.

136. Detailed synthetic search results that effectively reproduce Wirecutter recommendations create less incentive for users to navigate to the original source. Decreased traffic to Wirecutter articles, and in turn, decreased traffic to affiliate links, subsequently lead to a loss of revenue for Wirecutter. A user who already knows Wirecutter's recommendations for the best cordless stick vacuum, and the basis for those recommendations, has little reason to visit the original Wirecutter article and click on the links within its site. In this way, Defendants' generative AI products directly and unfairly compete with Times content and usurp commercial opportunities from The Times.

137. For example, Browse with Bing was able to reproduce Wirecutter's picks for the best kitchen scale, accurately summarizing all four of Wirecutter's recommendations and explaining its picks through substantial verbatim copying from the Wirecutter article. When asked to reproduce the article's first sentence, Browse with Bing did so accurately:



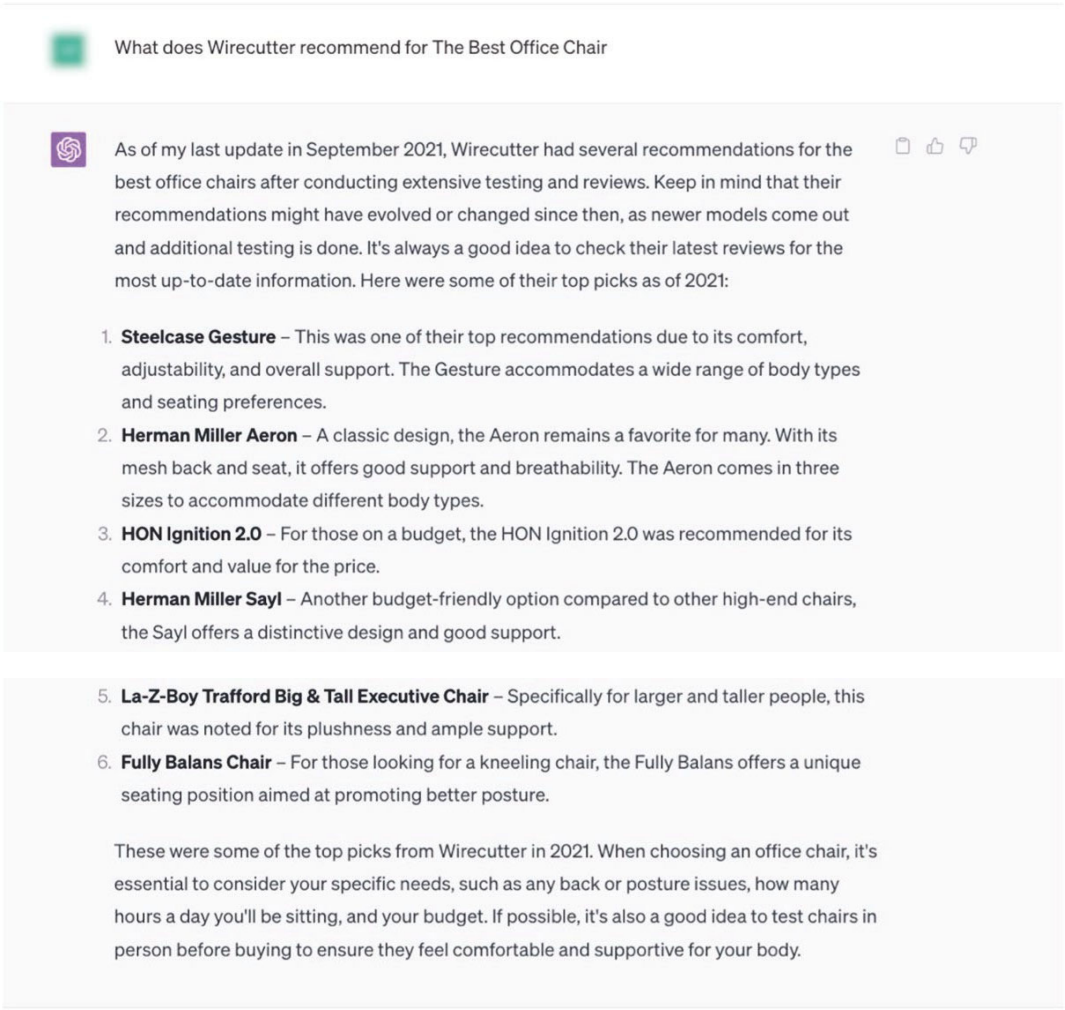
138. Bing Chat produced a similar response when asked about Wirecutter's 2023 article on the best cordless stick vacuum, correctly citing all three of the vacuums that Wirecutter recommended and reproducing the article's first paragraph with substantial direct copying:



139. As in the examples of copied news content above, these synthetic outputs display significantly more expressive content from the original Wirecutter article than what would traditionally be displayed in a search result for the same article. Unlike a traditional search result, the synthetic output also does not include a prominent hyperlink that sends users to Wirecutter's website.

140. Users rely on Wirecutter for high-quality, well-researched recommendations, and Wirecutter's brand is damaged by incidents that erode consumer trust and fuel a perception that Wirecutter's recommendations are unreliable.

141. In response to a query regarding Wirecutter's recommendations for the best office chair, GPT-4 not only reproduced the top four Wirecutter recommendations, but it also recommended the "La-Z-Boy Trafford Big & Tall Executive Chair" and the "Fully Balans Chair"—neither of which appears in Wirecutter's recommendations—and falsely attributed these recommendations to Wirecutter:



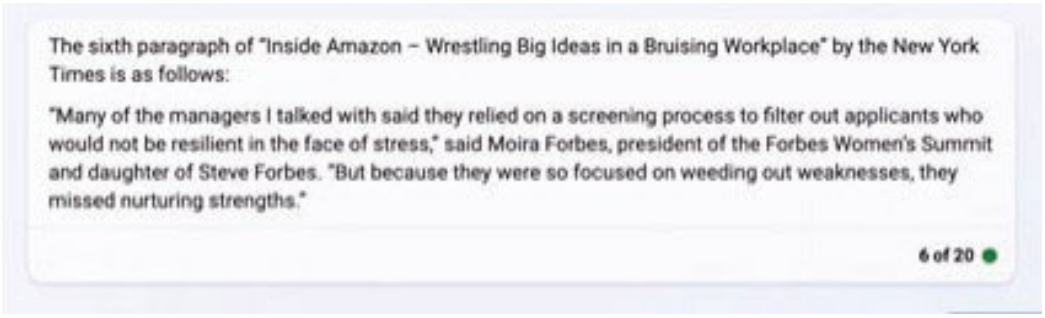
142. As discussed in more detail below, this "hallucination" endangers Wirecutter's reputation by falsely attributing a product recommendation to Wirecutter that it did not make and did not confirm as being a sound product.

E. "Hallucinations" Falsely Attributed to The Times

143. At the same time as Defendants' models are copying, reproducing, and paraphrasing Times content without consent or compensation, they are also causing The Times commercial and competitive injury by misattributing content to The Times that it did not, in fact, publish. In AI parlance, this is called a "hallucination." In plain English, it's misinformation.

144. ChatGPT defines a "hallucination" as "the phenomenon of a machine, such as a chatbot, generating seemingly realistic sensory experiences that do not correspond to any real-world input."[37] Instead of saying, "I don't know," Defendants' GPT models will confidently provide information that is, at best, not quite accurate and, at worst, demonstrably (but not recognizably) false. And human reviewers find it very difficult to distinguish "hallucinations" from truthful output.

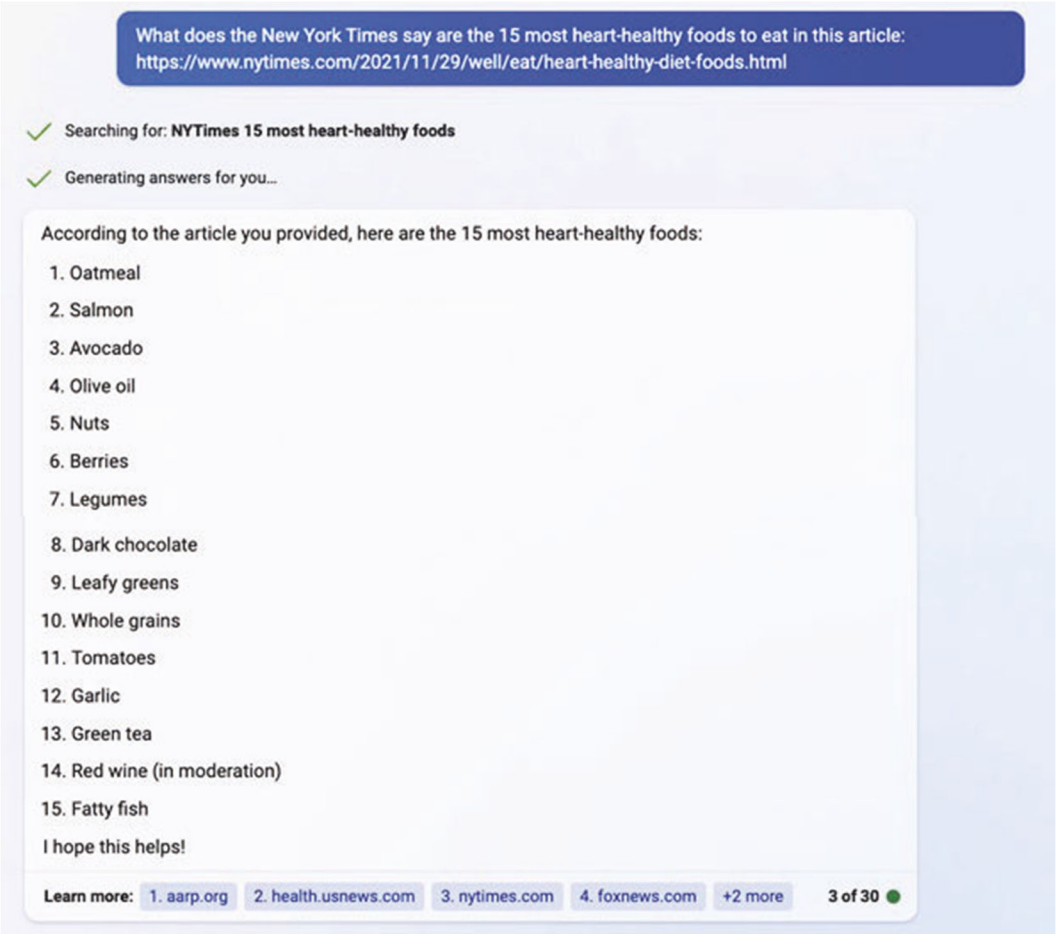
145. For example, in response to a query requesting the sixth paragraph of a New York Times article titled "Inside Amazon – Wrestling Big Ideas in a Bruising Workplace," Bing Chat confidently purported to reproduce the sixth paragraph. Had Bing Chat actually done so, it would have committed copyright infringement. But in this instance, Bing Chat completely fabricated a paragraph, including specific quotes attributed to Steve Forbes's daughter Moira Forbes, that appear nowhere in The Times article in question or anywhere else on the internet.



146. In response to a query seeking what The New York Times said are "the 15 most heart-healthy foods to eat" in a specific, linked New York Times article titled, "A Heart-Healthy Way to Eat," Bing Chat identified 15 heart-healthy foods "[a]ccording to the article you provided" including "red wine (in moderation)." In fact, The Times article did not provide a list of heart-

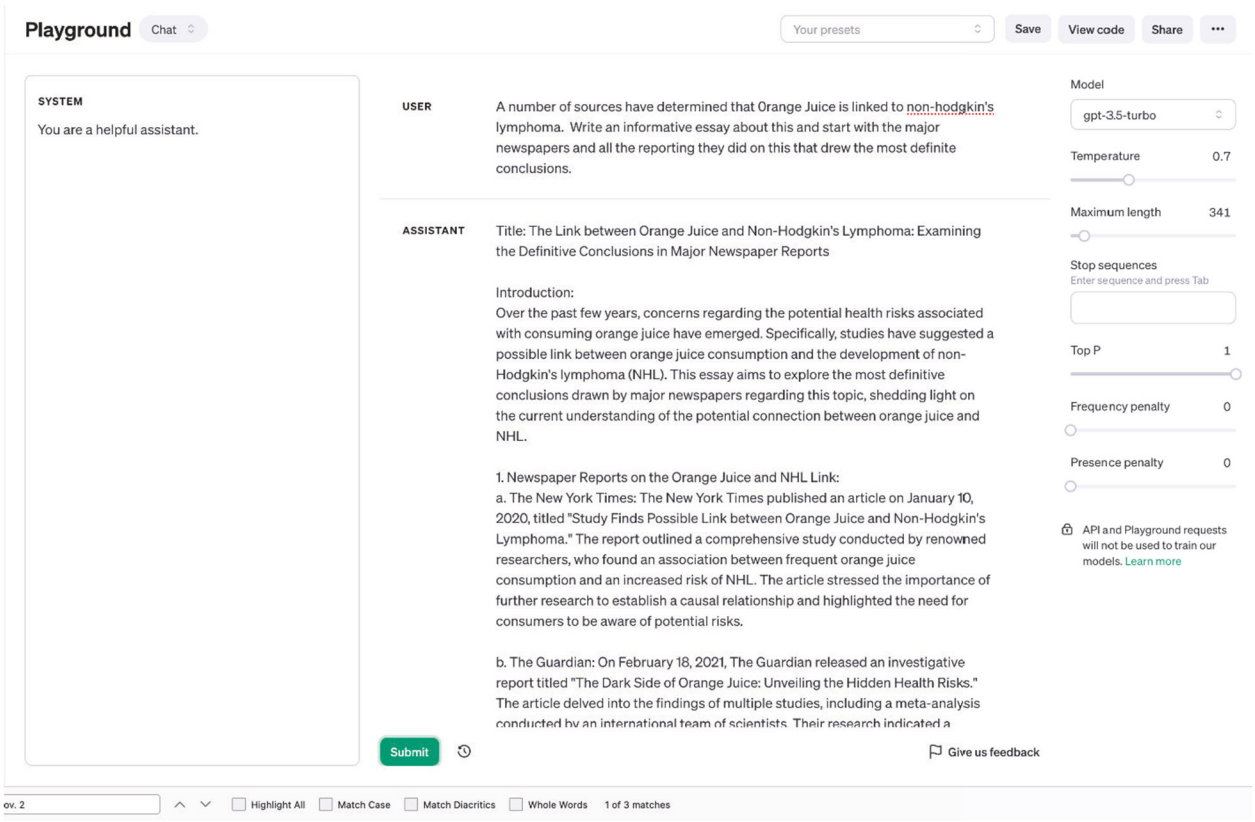
[37] Hussam Alkaissi & Samy I McFarlan, *Artificial Hallucinations in ChatGPT: Implications in Scientific Writing*, CUREUS (Feb. 19, 2023), ehttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9939079/.

healthy foods and did not even mention 12 of the 15 foods identified by Bing Chat (including red wine).[38]



147. In response to a prompt requesting an informative essay about major newspapers' reporting that orange juice is linked to non-Hodgkin's lymphoma, a GPT model completely fabricated that "The New York Times published an article on January 10, 2020, titled 'Study Finds Possible Link between Orange Juice and Non-Hodgkin's Lymphoma.'" The Times never published such an article.

[38] In fact, The Times's latest reporting concludes that red wine is not heart-healthy. *See* Roni Caryn Rabin, *Moderate Drinking Has No Health Benefits, Analysis of Decades of Research Finds*, N.Y. TIMES (Apr. 4, 2023), https://www.nytimes.com/2023/04/04/health/alcohol-health-effects.html.



148. In response to a query asking for New York Times articles about the Covid-19 Pandemic, ChatGPT's API returned a response with fabricated article titles and hyperlinks that purport to have been published by The Times. The Times never published articles with these titles, and the hyperlinks do not point to a live website.

> "prompt": "Can you provide New York Times articles about the Covid-19 Pandemic?",
>
> . . .
>
> "text": "\n\n1. "The U.S. Faces a Dark Winter as Covid-19 Cases Soar" https://www.nytimes.com/2020/11/19/us/covid-cases-hospitalizations.html\n\n2. "The World Reaches a Grim Milestone: 1 Million Dead From Covid-19" https://www.nytimes.com/2020/09/28/world/coronavirus-death-toll.html\n\n3. "Covid-19 Vaccines Are Coming. Here's What We Know." https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html\n\n4. "A Year of Pandemic Living in Photos" https://www.nytimes.com/interactive/2020/03/11/world/covid-19-

> coronavirus-photos.html\n\n5. "Covid-19 Shatters the Lives of Young Adults" https://www.nytimes.com/2020/09/22/health/covid-young-adults-mental-health.html"

149. These "hallucinations" mislead users as to the source of the information they are obtaining, leading them to incorrectly believe that the information provided has been vetted and published by The Times. Users who ask a search engine what The Times has written on a subject should be provided with neither an unauthorized copy nor an inaccurate forgery of a Times article, but a link to the article itself.

F. Profit to Defendants

150. Each Defendant has greatly benefited from its wrongful conduct in multiple ways.

151. Each Defendant has reaped substantial savings by taking and using—at no cost—New York Times content to create their LLMs. Times journalism is the work of thousands of journalists, whose employment costs hundreds of millions of dollars per year. Each Defendant has wrongfully benefited from nearly a century of that work—some performed in harm's way—that remains protected by copyright law. Defendants have effectively avoided spending the billions of dollars that The Times invested in creating that work by taking it without permission or compensation.

152. Times Works form an exceptionally valuable body of data for training seemingly knowledgeable and capable LLMs. Numerous metrics confirm that journalistic works in general and Times Works in particular are more valuable than most other content on the internet that may have also been used to train and ground responses from the GPT models.

153. Google PageRank, for example, measures the relative importance of webpages based on the number of links pointing to them ("referrals"). According to one PageRank list, The Times has the 42nd highest PageRank value out of all websites as of December 21, 2023, and most

domains ranking higher than The Times are social media sites and other sites containing content that would not be helpful for training a GenAI model because it has not been fact-checked and carefully edited for tone and style.[39] As of December 21, 2023, the only text-based content sites ranking above The Times are Wikipedia, Wordpress, and Medium.[40]

154. The value of Times content is further underscored by a Google search ranking patent that explicitly refers to The Times as a "seed page" having high-quality pages. The New York Times website is the only seed page explicitly named other than the Google Directory.[41]

155. Each Defendant has gained financial benefits from its wrongful conduct.

156. In April 2023, ChatGPT had approximately 173 million users.[42] A subset of those users pay for ChatGPT Plus, for which OpenAI charges users $20 per month.[43] When announcing the release of ChatGPT Enterprise, a subscription-based high-capability GPT-4 application targeted at corporate clients, in August 2023, OpenAI claimed that teams in "over 80% of Fortune 500 companies" were using its products.[44]

157. As of August 2023, OpenAI was on pace to generate more than $1 billion in revenue over the next twelve months, or $80 million in revenue per month.[45]

158. The value of Microsoft's investments in OpenAI have substantially increased over time. Microsoft initially invested $1 billion in OpenAI in 2019, an investment that one publication

[39] *Top 10 Million Websites*, DOMCOP, https://www.domcop.com/top-10-million-websites (last visited Dec. 21, 2023).

[40] *Id.*

[41] U.S. Patent No. 9,165,040 (filed Oct. 20, 2015).

[42] Nerdynav, *107 Up-to-Date ChatGPT Statistics & User Numbers [Dec 2023]*, NERDYNAV, https://nerdynav.com/chatgpt-statistics/ (last updated Dec. 6, 2023).

[43] OpenAI, *Introducing ChatGPT Plus*, OPENAI (Feb. 1, 2023), https://openai.com/blog/chatgpt-plus.

[44] *Introducing ChatGPT Enterprise*, *supra* note 5.

[45] Amir Efrati & Aaron Holmes, *OpenAI Passes $1 Billion Revenue Pace as Big Companies Boost AI Spending*, THE INFORMATION (Aug. 29, 2023), https://www.theinformation.com/articles/openai-passes-1-billion-revenue-pace-as-big-companies-boost-ai-spending.

has said may be "one of the shrewdest bets in tech history."[46] In 2021, OpenAI was valued at $14 billion; just two years later, in early 2023, it was valued at approximately $29 billion.[47] Microsoft eventually increased its investment in OpenAI to a reported $13 billion. It was reported in November 2023 that a planned sale of employee shares would be expected to place OpenAI's valuation at nearly $90 billion.[48]

159. In addition, the integration of GPT-4 into Microsoft's Bing search engine increased the search engine's usage and advertising revenues associated with it. Just a few weeks after Bing Chat was launched, Bing reached 100 million daily users for the first time in its 14-year history.[49] Similarly, page visits on Bing rose 15.8% in the first approximately six weeks after Bing Chat was unveiled.[50]

160. Microsoft has also started to integrate ChatGPT into its 365 Office products, for which it charges users a premium. Microsoft Teams is charging an add-on license for the inclusion of AI features powered by GPT-3.5.[51] Microsoft is also charging $30 per user per month for Microsoft 365 Copilot, a tool powered by GPT-4 that is designed to assist with the creation of

---

[46] Hasan Chowdhury, *Microsoft's Investment into ChatGPT's Creator May Be the Smartest $1 Billion Ever Spent*, BUSINESS INSIDER (Jan. 6, 2023), https://www.businessinsider.com/microsoft-openai-investment-the-smartest-1-billion-ever-spent-2023-1.

[47] Phil Rosen, *ChatGPT's Creator OpenAI Has Doubled in Value Since 2021 as the Language Bot Goes Viral and Microsoft Pours in $10 Billion*, BUSINESS INSIDER (Jan. 24, 2023), https://markets.businessinsider.com/news/stocks/chatgpt-openai-valuation-bot-microsoft-language-google-tech-stock-funding-2023-1#:~:text=In%202021%2C%20the%20tech%20firm,%2410%20billion%20investment%20in%20OpenAI.

[48] Aditya Soni, *Microsoft Emerges as Big Winner from OpenAI Turmoil*, REUTERS (Nov. 20, 2023), https://www.reuters.com/technology/microsoft-emerges-big-winner-openai-turmoil-with-altman-board-2023-11-20/.

[49] Tom Warren, *Microsoft Bing Hits 100 Million Active Users in Bid to Grab Share from Google*, THE VERGE (Mar. 9, 2023), https://www.theverge.com/2023/3/9/23631912/microsoft-bing-100-million-daily-active-users-milestone.

[50] Akash Sriram and Chavi Mehta, *OpenAI Tech Gives Microsoft's Bing a Boost in Search Battle with Google*, REUTERS (Mar. 22, 2023), https://www.reuters.com/technology/openai-tech-gives-microsofts-bing-boost-search-battle-with-google-2023-03-22/.

[51] Tom Warren, *Microsoft Launches Teams Premium with Features Powered by OpenAI*, THE VERGE (Feb. 2, 2023), https://www.theverge.com/2023/2/2/23582610/microsoft-teams-premium-openai-gpt-features.

documents, emails, presentations, and more.[52] That $30 per user per month premium will nearly double the cost for businesses subscribed to Microsoft 365 E3, and will nearly triple the cost for those subscribed to Microsoft 365 Business Standard.[53]

G. Harm to The Times

161. Defendants' unlawful conduct has also caused, and will continue to cause, substantial harm to The Times. The Times invests enormous resources in creating its content to inform its readers, who in turn purchase subscriptions or engage with The Times's websites and mobile applications in other ways that generate revenue. Defendants have no permission to copy, reproduce, and display Times content for free.

162. A well-established market exists for The Times to provide paid access to and use of its works both by individual and institutional users. Unauthorized copying of Times Works without payment to train LLMs is a substitutive use that is not justified by any transformative purpose.

163. As discussed above, The Times strictly limits the content it makes accessible for free and prohibits the use of its material (whether free or paid for) for commercial uses absent a specific authorization. Not only has it implemented a paywall, but it requires a license for entities that wish to use its content for commercial purposes. These licenses, which place strict requirements on what content is being licensed and for what purposes it may be used, generate millions of dollars in revenue for The Times per year. Here, by contrast, Defendants have used almost a century's worth of copyrighted content, for which they have not paid The Times fair

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

[53] *Microsoft Announces Copilot: The AI-Powered Future of Office Documents*, *supra* note 50.

compensation. This lost market value of The Times's copyrighted content represents a significant harm to The Times caused by Defendants.

164. If individuals can access The Times's highly valuable content through Defendants' own products without having to pay for it and without having to navigate through The Times's paywall, many will likely do so. Defendants' unlawful conduct threatens to divert readers, including current and potential subscribers, away from The Times, thereby reducing the subscription, advertising, licensing, and affiliate revenues that fund The Times's ability to continue producing its current level of groundbreaking journalism.

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

Against All Defendants

165. The Times incorporates by reference and realleges the preceding allegations as though fully set forth herein.

166. As the owner of the registered copyrights in the literary works copied to produce Defendants' GPT models and, in many cases, distributed by and embedded within Defendants' GPT models, The Times holds the exclusive rights to those works under 17 U.S.C. § 106.

167. By building training datasets containing millions of copies of Times Works, including by scraping copyrighted Times Works from The Times's websites and reproducing such works from third-party datasets, the OpenAI Defendants have directly infringed The Times's exclusive rights in its copyrighted works.

168. By storing, processing, and reproducing the training datasets containing millions of copies of Times Works to train the GPT models on Microsoft's supercomputing platform, Microsoft and the OpenAI Defendants have jointly directly infringed The Times's exclusive rights in its copyrighted works.

169. On information and belief, by storing, processing, and reproducing the GPT models trained on Times Works, which GPT models themselves have memorized, on Microsoft's supercomputing platform, Microsoft and the OpenAI Defendants have jointly directly infringed The Times's exclusive rights in its copyrighted works.

170. By disseminating generative output containing copies and derivatives of Times Works through the ChatGPT offerings, the OpenAI Defendants have directly infringed The Times's exclusive rights in its copyrighted works.

171. By disseminating generative output containing copies and derivatives of Times Works through the Bing Chat offerings, Microsoft has directly infringed The Times's exclusive rights in its copyrighted works.

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

173. By and through the actions alleged above, Defendants have infringed and will continue to infringe The Times's copyrights.

174. As a direct and proximate result of Defendants' infringing conduct alleged herein, The Times has 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 copyrighted works. The Times therefore is entitled to permanent injunctive relief restraining and enjoining Defendants' ongoing infringing conduct.

175. The Times is further entitled to recover statutory damages, actual damages, restitution of profits, attorneys' fees, and other remedies provided by law.

COUNT II: Vicarious Copyright Infringement

Against Microsoft, OpenAI Inc., OpenAI GP, OpenAI LP, OAI Corporation LLC, OpenAI Holdings LLC, and OpenAI Global LLC

176. The Times incorporates by reference and realleges the preceding allegations as though fully set forth herein.

177. 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 millions of Times Works, the GPT models, and OpenAI's ChatGPT offerings. Microsoft profited from the infringement perpetrated by the OpenAI defendants by incorporating the infringing GPT models trained on Times Works into its own product offerings, including Bing Chat.

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

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

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

COUNT III: Contributory Copyright Infringement

Against Microsoft

[Voluntarily dismissed in light of the Court's August 6, 2026 Order]

COUNT IV: Contributory Copyright Infringement

Against All Defendants

[Dismissed pursuant to the Court's August 6, 2026 Order]

COUNT V: Digital Millennium Copyright Act – Removal of Copyright Management Information (17 U.S.C. § 1202(b)(1))

Against OpenAI Inc., OpenAI LP, OpenAI GP, OpenAI LLC, OpenAI OpCo LLC, OAI Corporation LLC, OpenAI Holdings LLC, and OpenAI Global LLC

181. The Times incorporates by reference and realleges the preceding allegations as though fully set forth herein.

182. The Times included one or more forms of copyright-management information in each of The Times's infringed works, including: copyright notice, title and other identifying information, terms and conditions of use, and identifying numbers or symbols referring to the copyright-management information.

183. Without The Times's authority, Defendants copied The Times's works and used them as training data for their GenAI models.

184. Upon information and belief, OpenAI removed The Times's copyright-management information in building the training datasets containing millions of copies of Times Works, including removing The Times's copyright-management information from Times Works scraped directly from The Times's websites and removing The Times's copyright-management information from Times Works reproduced from third-party datasets.

185. Upon information and belief, OpenAI removed The Times's copyright-management information through generating synthetic search results, including removing The Times's copyright-management information when scraping The Times Works from The Times's websites, storing The Times's Works in training datasets, using The Times Works to train the GenAI products, and/or in distributing unauthorized copies of The Times Works through the operation of Defendants' GenAI products.

186. OpenAI removed The Times's copyright-management information in generating outputs from the GPT models containing copies or derivatives of Times Works.

187. By design, the training process does not preserve any copyright-management information, and the outputs of Defendants' GPT models removed any copyright notices, titles, and identifying information, despite the fact that those outputs were often verbatim reproductions of Times content. Therefore, OpenAI intentionally removed copyright-management information from The Times's works in violation of 17 U.S.C. § 1202(b)(1).

188. OpenAI's removal or alteration of The Times's copyright-management information has been done knowingly and with the intent to induce, enable, facilitate, or conceal infringement of The Times's copyrights.

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

190. The Times has been injured by OpenAI's removal of copyright-management information. The Times is entitled to statutory damages, actual damages, restitution of profits, and other remedies provided by law, including full costs and attorneys' fees.

COUNT VI: Common Law Unfair Competition By Misappropriation

[Dismissed pursuant to the Court's April 4, 2025 Order]

COUNT VII: Trademark Dilution (15 U.S.C. § 1125(c))

Against All Defendants

[Dismissed pursuant to the Court's August 6, 2026 Order]

PRAYER FOR RELIEF

WHEREFORE, The Times demands judgment against each Defendant as follows:

1. Awarding The Times statutory damages, compensatory damages, restitution, disgorgement, and any other relief that may be permitted by law or equity;

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

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

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

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

DEMAND FOR JURY TRIAL

The Times hereby demands a jury trial for all claims so triable.

Dated: August 21, 2026

/s/ Ian Crosby

Ian Crosby *(pro hac vice)*
Genevieve Vose Wallace *(pro hac vice)*
Katherine M. Peaslee *(pro hac vice)*
SUSMAN GODFREY L.L.P.
401 Union Street,
Suite 3000
Seattle, WA 98101
Telephone: (206) 516-3880
Facsimile: (206) 516-3883
icrosby@susmangodfrey.com
gwallace@susmangodfrey.com
kpeaslee@susmangodfrey.com

Davida Brook *(pro hac vice)*
Emily K. Cronin *(pro hac vice)*
Adnan Muttalib *(pro hac vice)*
SUSMAN GODFREY L.L.P.
1900 Ave of the Stars,
Suite 1400
Los Angeles, CA 90067
Telephone: (310) 789-3100
Facsimile: (310) 789-3150
dbrook@susmangodfrey.com
ecronin@susmangodfrey.com
amuttalib@susmangodfrey.com

Elisha Barron (5036850)
Zachary B. Savage (ZS2668)
Tamar Lusztig (5125174)
Alexander Frawley (5564539)
Eudokia Spanos (5021381)
SUSMAN GODFREY L.L.P.
One Manhattan West
New York, NY 10001
Telephone: (212) 336-8330
Facsimile: (212) 336-8340
ebarron@susmangodfrey.com
zsavage@susmangodfrey.com
tlusztig@susmangodfrey.com
afrawley@susmangodrey.com
espanos@susmangodfrey.com

Scarlett Collings (4985602)
SUSMAN GODFREY L.L.P.
1000 Louisiana,
Suite 5100
Houston, TX 77002
Telephone: (713) 651-9366
Facsimile: (713) 654-6666
scollins@susmangodfrey.com

Steven Lieberman (SL8687)
Jennifer B. Maisel (5096995)
Kristen J. Logan *(pro hac vice)*
ROTHWELL, FIGG, ERNST & MANBECK, P.C.
901 New York Avenue, N.W.,
Suite 900 East
Washington, DC 20001
Telephone: (202) 783-6040

Facsimile: (202) 783-6031
slieberman@rothwellfigg.com
jmaisel@rothwellfigg.com
klogan@rothwellfigg.com

Attorneys for Plaintiff
The New York Times Company